WILD CARD WEIGHTING

JACK COOK AND JACKSON POND

Abstract. The MLB postseason is known to be unpredictable. Over the last decade, wild card teams sometimes outperform division winners. This paper aims to investigate and quantify the underlying reasons behind this phenomenon. Using a dataset of MLB games from 2014–2025, we train both Random Forest and XGBoost models to identify the strongest features influencing postseason success and evaluate whether wild card teams possess structural advantages that are not captured by regular season standings.

1. Research question and overview of the data

In Major League Baseball, the wild card teams are those who are have comparitively weaker records than the division winners, but are still selected to advance to the postseason. Over the last decade, wild card teams have performed surprisingly well in the postseason, sometimes beating division winners. This raises important questions for both baseball fans and analysts:

> *How should the fact that a team is a wild card factor in to predicting their success?*

This question is interesting because it challenges the assumption that regular season success is a reliable predictor of postseason performance. It also relates to the design of playoff formats and whether certain types of teams are structurally advantaged in short playoff series.

2. Background and Prior Work

Several previous studies have examined what factors into MLB postseason performance. FanGraphs applied a clustering model to determine which bullpen matchups produced the highest probability of postseason success [Fan25b]. Furthermore, FanGraphs also performed a study that looked at the correlation of various statistics for predicting postseason success, only to find that playoff success is random and unpredictable by regular season performance [Fan25a]. We hope to expand on these studies by using random forest and XGBoost models to identify the most important features influencing postseason success.

Date: 4 October 2024.

3. Data Set

We use MLB game data from 2014-2025. The data includes variables such as pitcher skill level and team average outs above average (OAA), team average weighted on base average (WOBA), bullpen fielding independent pitching (FIP), bullpen freshness, and whether the team is a wild card. Our final dataset incorporates data from 1) The pitcher logs for each game for each pitcher during the regular seasons and post seasons from 2014-2025, and 2) team level statistics (such as outs above average, weighted on-base average, whether the team is a wild card) for each team for each year from 2014 - 2025.

A shortcoming of this dataset is that the quantity of postseason games (317) is dwarfed by the quantity of regular season games (20,000). Because the state of whether a team is a wild card or not is only pertinent in the postseason, this dataset (which reflects the real life proportion) is wildly unbalanced.

Based on our prior experience with baseball, we form the following hypothesis:

(1) Wild card teams may enter the later rounds of the postseason with the advantages of momentum (having played other wild card teams while division winners rested) and the high stakes of a Cinderella story, neither of which would be otherwise captured in the dataset.

4. Data Cleaning / Feature Engineering

Our dataset required extensive data cleaning and feature engineering.

We decided to create a match-up model to with the following features for both teams: starting pitcher's FIP, starting pitcher's freshness, an aggregated value for the FIP and freshness of the pitchers in the bullpen (bank of relief pitchers), average OAA (as an indication of the efficacy of the team's defense), average WOBA (as an indication of the efficacy of the team's offense), whether the team is a wild card in the playoffs, and whether the team is at home.

The pitchers' logs (sourced from Baseball-Reference[BR25]) included the following pertinent features: Date, Team, a unique Pitcher ID, FIP, Innings played, and the result of the game. In order to assemble it into a useable format for our problem, we made a few tweaks and added some new columns.

- After converting the Date column to pd.DateTime, we sorted the dataset by Pitcher ID and Date and created a Freshness column by subtracting the previous game's date from each game. A pitcher's first game was assigned the dataset's maximum freshness.
- We added a Starter feature that one-hot-encodes whether the pitcher started pitching that game.
- We created a Game ID from the date and two teams playing. Teams occasionally play two games on the same day. Because this is both problematic and rare, we decided to drop these games.

The dataset of team level statistics by year (sourced from Baseball Savant[bas25]) included the following pertinent features: Team, Year, OOA (averaged over the season), WOBA (averaged over the season), and Is-Wildcard. This data was quite clean, except that it was missing all WOBA values for 2024. We decided the lowest impact filler would be the average of all WOBA values. After readying the data, we merged it with the pitcher logs on Year and Team.

Equipped with a newly-merged datset (with features Date, Team, Pitcher ID, FIP, Innings played, Game Result, Freshness, Starter, Game ID, Team OOA, Team WOBA, Is-Wildcard for each pitcher in each game) we were ready to assemble a final games dataset. To do so, we grouped the newly-merged dataset by Game ID and then by team. For each team in each game, we assembled a dictionary containing the starter's FIP and freshness, the meaned FIP and freshness of the relief pitchers, whether the team was at home, team OOA, team FIP, whether it was a post-season game, and if so, whether the team was a wild card during the game. After shuffling the order of the teams, we concatenated the two team's stats into a single row, with a results feature signifying whether the first team in the row won or lost. There were no ties in our dataset. There was a single row with missing data, which we dropped.

In the final dataset, there is a row for each game containing stats that were true *by the end of the game*. For instance, the FIP value in the game's row is the FIP of the pitcher calculated at the end of the game. If this were a prediction model, this would be considered a mild form of data leakage. Since our model is not a prediction model, but rather a model trained in order to evaluage feature importance, we may treat these post-game calulated values as true values. The same concept applies to why we evaluated the freshness and skill (FIP) of the pitchers that actually played in the game rather than the pitchers in the bullpen that were available.

With our data in a tabular dataset with a binary label, we were ready to train!

5. Data Visualization and Basic Analysis

Our final dataset is large, with 19,781 rows, one per game, and 20 features, namely **Game ID**, **1 Starter FIP**, **1 Starter Freshness**, **1 Relief FIP**, **1 Relief Freshness**, **1 WOBA**, **1 OOA**, **1 Home** (Is team 1 at home?), **1 Team**, **1 Is Wildcard**, **2 Starter FIP**, **2 Starter Freshness**, **2 Relief FIP**, **2 Relief Freshness**, **2 WOBA**, **2 OOA**, **2 Team**, **2 Is Wildcard**, **Is Playoff Game**, and **Result** (Did Team 1 win?). In this section, we will visually inspect the data and evaluate our initial assumptions.

We will first inspect the distributions of key numerical features Starter FIP, Starter Freshness, Relief FIP, Relief Freshness, WOBA, and OOA. We will only look at team 1 since team 1 is a random half of the data, and the corresponding values and distributions for team 2 match.

As seen in Figure 1, skill metrics (FIP, WOBA, OOA) are normally distributed, which is as expected. The distribution of freshness, both for starting pitchers and relief pitchers, is heavily skewed towards 0 days of rest.

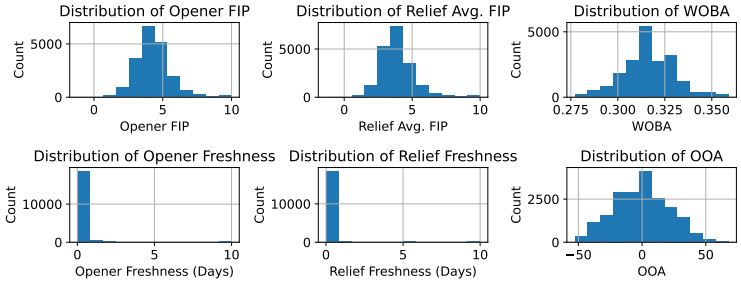

FIGURE 1. Distributions of the features of our final dataset. Skill metrics (FIP, WOBA, OOA) are normally distributed, while freshness is skewed towards 0 days rest.

Figure 2 shows the distributions of each of the six key features grouped by result. As shown in the figure, the distributions for each feature hardly vary by result, implying that Baseball data is quite noisy, with lots of randomness incorporated into the final result.

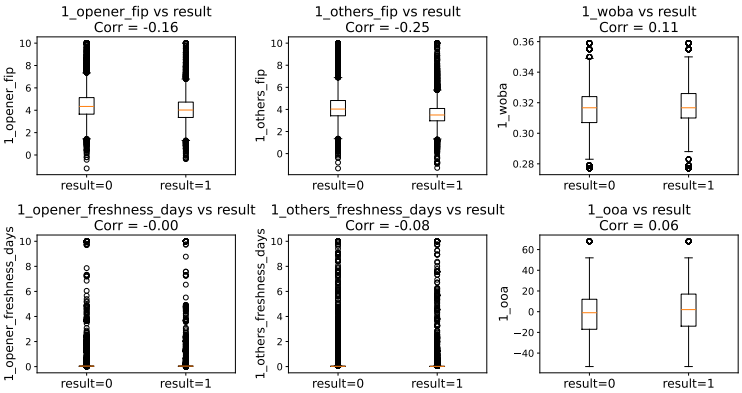

FIGURE 2. Distribution of six features grouped by result (1 for win, 0 for loss). The result-grouped distributions are nearly identical for each feature.

Considered as a whole, the data is clean and comprehensive, including nearly every game from 2014-2025. The feature values are either numerical ratio data or boolean expressed as 0 or 1, and the distributions of each feature are as expected. The format of the dataset is ideal for training data.

6. Learning Algorithms and In-depth Analysis

To answer our research question, we prioritized feature importance over accuracy. We selected Random Forest and XGBoost classifiers specifically

because these tree-based models are more interpretable, providing feature importance scores.

We built a comparative framework by training separate models for the regular season (serving as a baseline) and the postseason. Although both tree-based algorithms were implemented, we focus only on the XGBoost model due to its better performance metrics. Following an 80/20 train-test split, the XGBoost model achieved an accuracy and F1 score of 0.701 for the regular season and 0.733 for the postseason.

Our results show a noticeable shift in which pitching metrics matter most in postseason success. While bullpen FIP was the primary predictor of success in the regular season, starting pitcher FIP became the most important feature in the postseason. Regarding our initial research question, the 'Is Wildcard' feature proved negligible with an importance score of just ≈ 0.037. This indicates that wild card teams possess no structural advantage in the postseason.

See Figure 3.

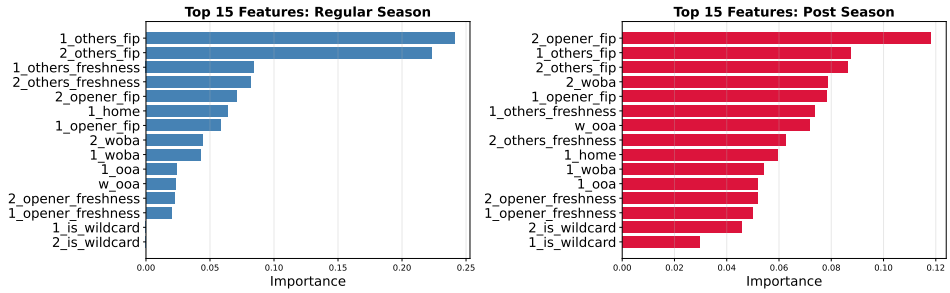

FIGURE 3. Comparison of the top 15 most important features for the regular season (left) versus the postseason (right).

The results show that postseason success depends on talented starting pitching. We also observed an increase in the importance of defensive and offensive metrics (such as OAA and WOBA), reflecting the heightened competitiveness and smaller margins of error in postseason play.

7. Ethical Implications and Conclusions

The ethical implications of this project are relatively minimal. The data is publicly available and we did not collect any new data. The models we trained are used to identify feature importance and not used in a way that is harmful or misleading. Therefore, there is no risk of misuse or misunderstanding.

In conclusion, our project revealed that postseason pitching is the most important factor in determining success. Our original hypothesis that wild card teams would perform better than division winners was not supported by our findings.

References

[bas25] Baseball savant. https://baseballsavant.mlb.com, 2025. Accessed November 25, 2025.

[BR25] Baseball-Reference. Player game logs. https://www.baseball-reference.com, 2025. Accessed: 2025-11-24.

[Fan25a] FanGraphs. Predicting the playoffs. https://community.fangraphs.com/predicting-the-playoffs/, 2025. Accessed November 25, 2025.

[Fan25b] FanGraphs. Using clustering to generate bullpen matchups. https://community.fangraphs.com/using-clustering-to-generate-bullpen-matchups/, 2025. Accessed November 25, 2025.