WILDCARD WEIGHTING

JACK COOK AND JACKSON POND

Abstract. The MLB postseason is known to be unpredictable. Over the last decade, Wild Card teams have outperformed division winners. This paper aims to investigate and quantify the underlying factors driving this phenomenon. Using a dataset of MLB games from 2014–2025, we train both Random Forest and XGBoost models to identify the variables that most strongly influence postseason success and to 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 selected to advance to the postseason. Over the last decade, wild card teams have performed surprisingly well in the postseason, often 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 MLB postseason performance. Baseball Prospectus and FanGraphs have written extensively about the role of randomness, bullpen strength, and matchups in postseason series [Bas25b, Fan25] applying regression models to predict postseason outcomes. Academic work (e.g., Berri & Schmidt, 2010) has shown that regular season winning percentage is a relatively weak predictor of postseason outcomes [BS10].

3. Data Set

We use MLB game data from 2014-2025. The data includes variables such as pitcher skill level and team OOA, team WOBA, bullpen FIP, bullpen

Date: 4 October 2024.

freshness, and whether the team is a wildcard (lingo will be explained later in the paper). The data set is strong in that the data is specifically geared towards matchups and understanding the impact of the wildcard on the postseason.

We gathered the data for the project in three main pulls. Our final dataset incorporates data from 1) The pitcher logs for each game for each pitcher during the regular season from 2014-2025, 2) the pitcher logs for each game for each pitcher during the post season from 2014-2025, and 3) team level statistics (such as outs above average or weighted on-base average, whether the team is a wildcard) 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. Therefore any models we train may be incentivized to disregard the wildcard feature.

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

(1) Wild Card teams may enter the later rounds of the postseason carrying more momentum than their division winner counterpart, causing them to win.
(2) Bullpen strength will be more important in the postseason than in the regular season.
(3) Matchup-based factors (pitcher handedness, lineup construction) may be more important than overall regular season strength.

4. Data Cleaning / Feature Engineering

Our dataset required extensive data cleaning and feature engineering. As a precursor to the discussion, we will first define important terms and acronyms.

(1) **FIP** Fielding Independent Pitching is a metric designed to capture how well a pitcher prevents runs independently of the rest of the defense.
(2) **Bullpen** Because pitching strains a pitcher's arm quickly, one pitcher will start and throw about 5-7 innings before being switched out for a *relief pitcher*. The bank of pitchers ready to relieve is called the Bullpen.
(3) **OAA** stands for Outs Above Average and signifies how many more outs a player makes than an average fielder at the same position.
(4) **WOBA** Weighted On-Base Average signifies how often players make it on base, weighted by how likely they are to score conditioned on how they made it on base.

Early on, we decided that we wanted our match-up model to include 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, 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), and whether the team is a wildcard in the playoffs. We also decided to include a feature indicating which team played 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[bas25a]) 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 wildcard 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 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 3 also shows the weakness of logistic correlation between the *differences* between the two teams' features and the results of the game. The results are a 1 if team 1 wins and 0 if team 1 loses. Each of the independent variables is the value for team 2 subtracted from the value for team 2. As epected, the team with higher WOBA and higher OAA is expected to win. Mysteriously, a pitcher with a higher starting FIP than the opponent is inversely correlated with games won.

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. It is ideal for training data.

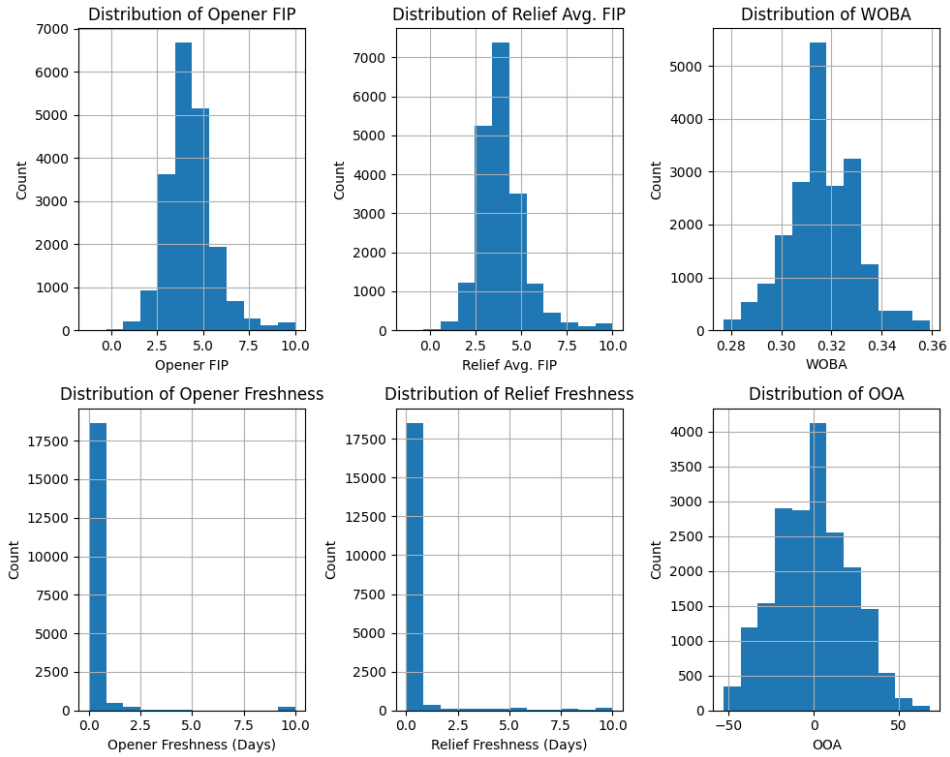

FIGURE 1. Distributions of various of the features of our final dataset. Each of the skill metrics (FIP, WOBA, OOA) are distributed normally, as we would expect, and the freshness is greatly skewed towards 0 days rest.

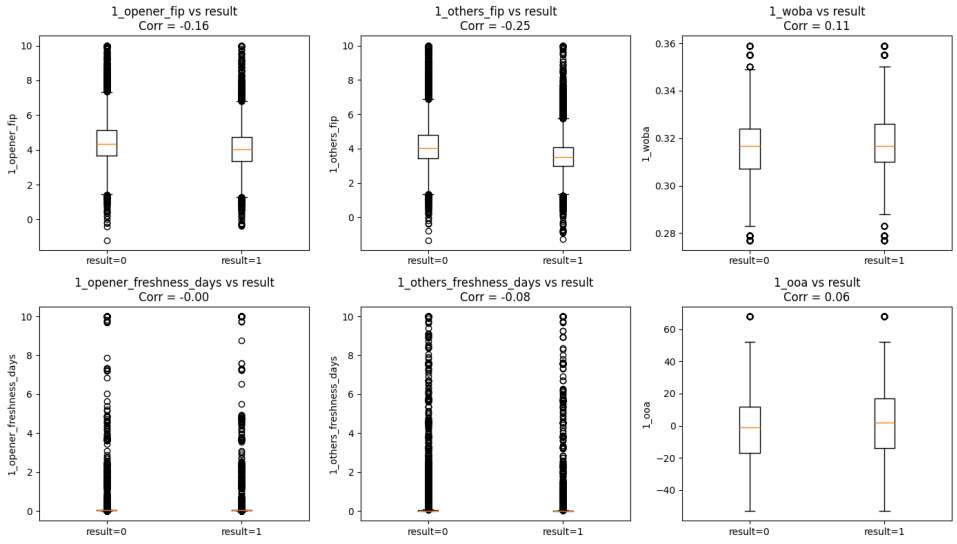

FIGURE 2. This figure shows the distribution of six features, grouped by result (1 for win, 0 for loss). Regardless of feature, the result-grouped distributions are nearly identical for each feature.

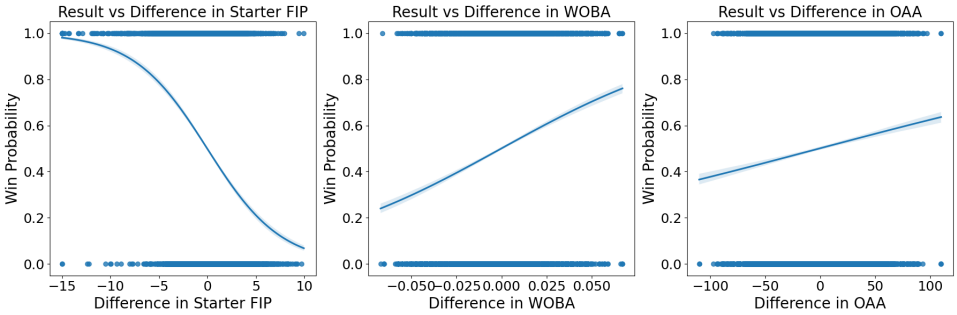

FIGURE 3. These figures plot the result of a game (1 for win, 0 for loss) against the difference between team 1 and team 2 for various features: Starter FIP, WOBA, and OOA. We hypothesized that each feature would correlate with wins, and are surprised that Starter FIP is inversely correlated with wins.

6. Learning Algorithms and In-depth Analysis

For our project, we trained both Random Forest and XGBoost models to predict the outcome of a game based on the features of the teams playing. The primary goal of our modeling was not prediction accuracy, but rather to identify which features most strongly influence game outcomes in the postseason. These models work well for our problem because they are able to return feature importance scores that are interpretable and easy to understand.

With different iterations, our best Random Forest model achieved an F1 score of 0.6963, while our best XGBoost model achieved an F1 score of 0.701. The two models had similar feature importance ranking, with the top 15 features being the same for both models. We found that the most important features in the postseason were a team's FIP and freshness for both starting and relief pitchers. This was somewhat surprising to us, as we expected that a team's OOA and WOBA would be more significant in October. However, our results are consistent with baseball intuition: pitching wins games. We also found that being a Wild Card team was not nearly as significant as we expected, ranking last in feature importance. See Figure 4.

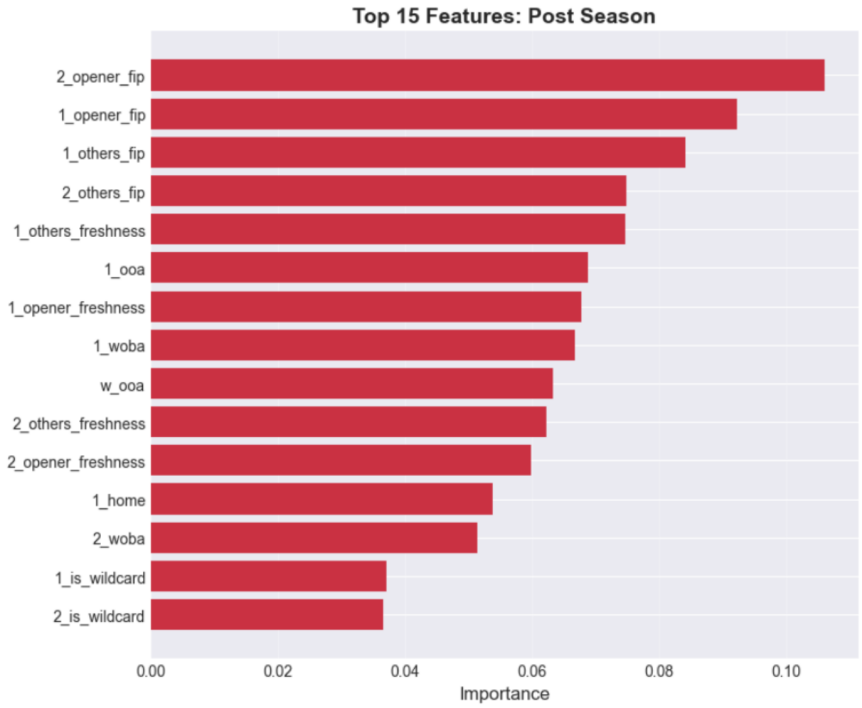

Figure 4. Feature importance for the postseason (XGBoost model)

Taken together, these results suggest what teams should try to optimize for in the postseason. Having a talented, well rested bullpen and a strong starting pitcher appears to be a key ingredient for postseason success. This is easier said than done, as it requires managing the bullpen in a way that does not wear pitchers out over the course of a 162-game season while still winning enough games to reach October in the first place. There has been significant research into optimal bullpen management, but elaborating further is outside the scope of this project.

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

[bas25a] Baseball savant. https://baseballsavant.mlb.com, 2025. Accessed November 25, 2025.

[Bas25b] Baseball Prospectus. Postseason analysis archive. https://www.baseballprospectus.com, 2025. Accessed November 25, 2025.

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

[BS10] David J. Berri and Martin B. Schmidt. The wages of wins and the mlb playoffs. *Journal of Sports Economics*, 2010. Exact citation adapted for course project.

[Fan25] FanGraphs. Postseason research and analysis. https://www.fangraphs.com, 2025. Accessed November 25, 2025.