WILDCARD WEIGHTING

PLACE NAMES OF GROUP MEMBERS HERE

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 2015–2025, 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

Within the Major League Baseball (MLB) organization, there are two leagues (The American League and the National League), which are further split into three divisions each. At the end of the regular season of 162 games for each team, 12 teams are chosen to advance in the post-season, six from each league. Six of these teams are simply the teams with the best season record within their division. These teams are known as "division winners". The remaining six teams that advance to the postseason are the teams with the best season records out of the remaining teams, three from each league. These teams are known as "wildcards", and are the underdogs of the postseason tournament which culminates in the World Series, where the winners of each league square off to be declared the best team in the MLB.

Recent MLB postseasons have shown an interesting trend: Wild Card teams often perform better than division winners, even though they typically finish the regular season with weaker records. This raises an important question for both baseball fans and analysts:

> *What makes a Wild Card team outperform a division winner in the MLB postseason?*
>
> *How should the fact that a team is a Wild Card factor in to predicting their success?*

These two questions are interesting because they challenge the assumption that regular season strength 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.

Date: 4 October 2024.

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 [Bas25, Fan25]. 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 2015-2025. The data includes variables such as team OOA, team WOBA, bullpen FIP, bullpen freshness, and whether the team is a wildcard. 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 2014-2015, 2) the pitcher logs for each game for each pitcher during the post season from 2014-2015, 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

Tell us what you did when you were cleaning your data and engineering features. Why did you make the choices that you did? What are the consequences of those choices?

(1) **FIP** stands for Fielding Independent Pitching, and is a metric designed to capture how well a pitcher prevents runs independently of the rest of the defense. It is a better metric of a pitcher's skill

than Earned Run Average because it disregards events outside of the pitcher's control.

(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 are called the Bullpen.

(3) **OOA** stands for Outs Above Average and signifies how many more outs a player makes than an average fielder at the same position.

(4) **WOBA** stands for Weighted On-Base Average and signifies how often players make it on base, weighted by how likely they are to score based on the manner in which they made it on base.

We gathered the data we used 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 2014-2015, 2) the pitcher logs for each game for each pitcher during the post season from 2014-2015, and 3) team level statistics (such as outs above average or weighted on-base average) for each team for each year from 2014 - 2025.

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 OOA (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 pitcher's logs (collectively referred to as df) 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. We first converted the Date column to pandas.DateTime. After sorting the df by Pitcher Id and Date, we grouped the df by Pitcher ID and shifted the Date column to create a Previous Game Date column, and subtracted the Previous Game Date column from the Date column to create a Freshness column for each pitcher in each game.

Additional changes to make the pitcher logs more useful included adding a Starter feature that one-hot-encodes whether the pitcher started that game, and a Game ID to uniquely identify each game, based on the date and the two teams playing.

The dataset of team level statistics by year included the following pertinent features: Team, Year, OOA (averaged over the season), WOBA (averaged over the season), and Is-Wildcard. This dataset was nearly immediately useable after cleaning the Team names to match those in the pitcher logs dataset. 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.

This means that in the final dataset, there is a row for each game or match-up. Each row stats for each teams 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

Analyze the data, draw conclusions, and effectively communicate your main observations and results.

- Calculate appropriate summary statistics.
- Use appropriate plotting techniques, visualizations, and other tools and techniques you have learned, to thoughtfully identify and evaluate what the data are telling you,how well suited the data are to answering your problem,
- Reference figures and plots, like Figure **??**.

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.
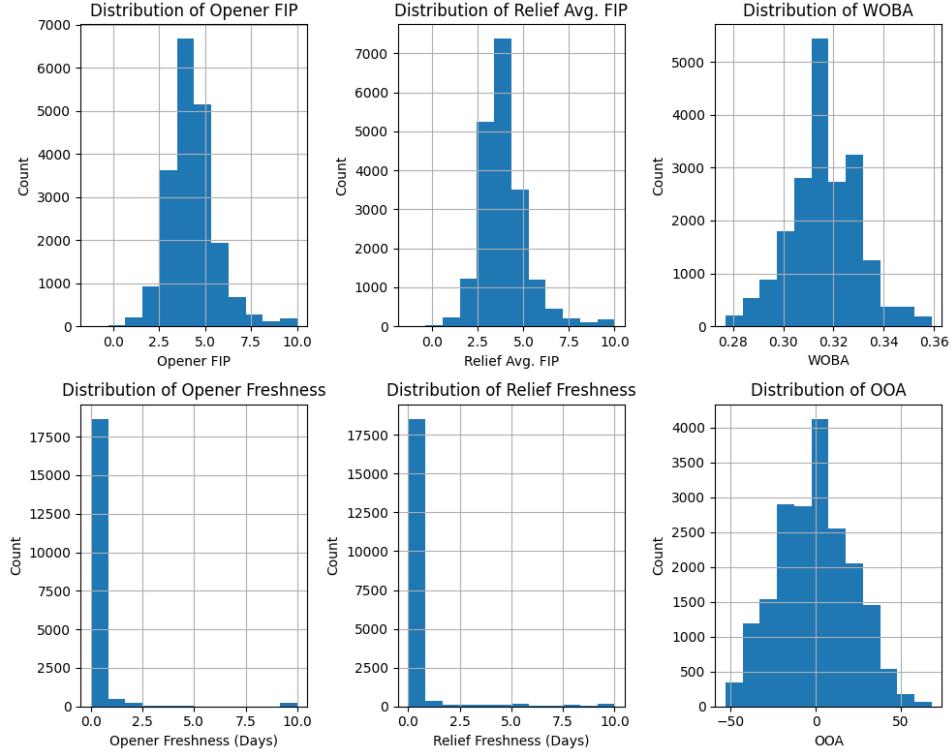

FIGURE 1. Distributions of various of the features of our final dataset. Each of the skill metrics (FIP, WOBA, OOA) are distributed normally, as we would expect, and the freshness is greatly skewed towards 0 days rest.
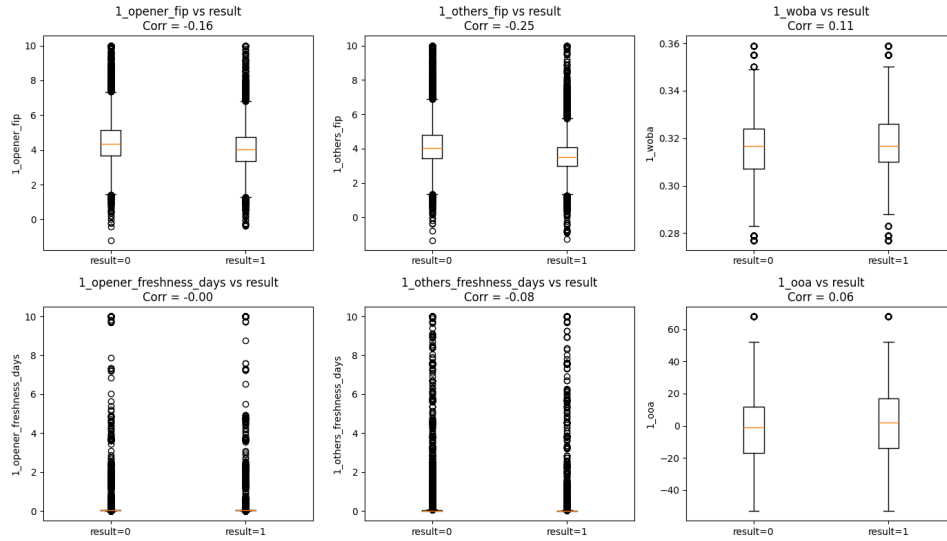
FIGURE 2. This figure shows the distribution of six features, grouped by result (1 for win, 0 for loss). Regardless of feature, the result-grouped distributions are nearly identical for each feature.
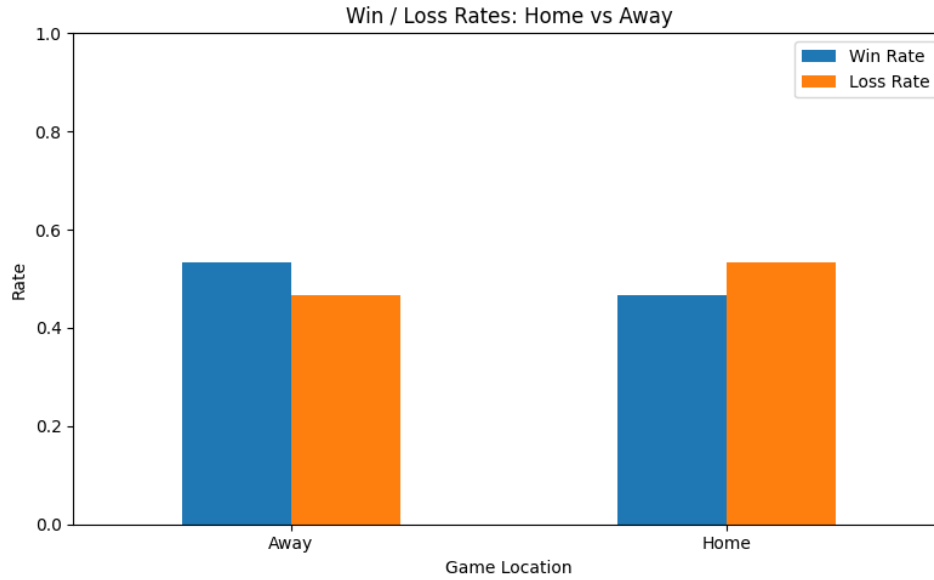

FIGURE 3. Surprisingly, teams are slightly more likely to win while playing away games, and slightly more likely to lose at home games.
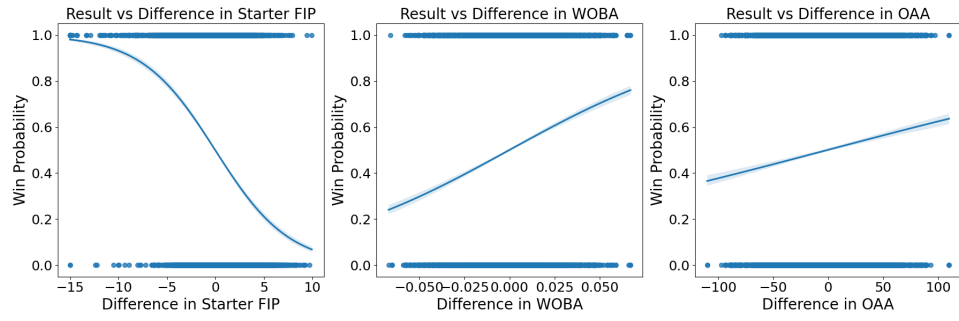
FIGURE 4. These figures plot the result of a game (1 for win, 0 for loss) against the difference between team 1 and team 2 for various features: Starter FIP, WOBA, and OOA. We hypothesized that each feature would correlate with wins, and are surprised that Starter FIP is inversely correlated with wins.

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 5.
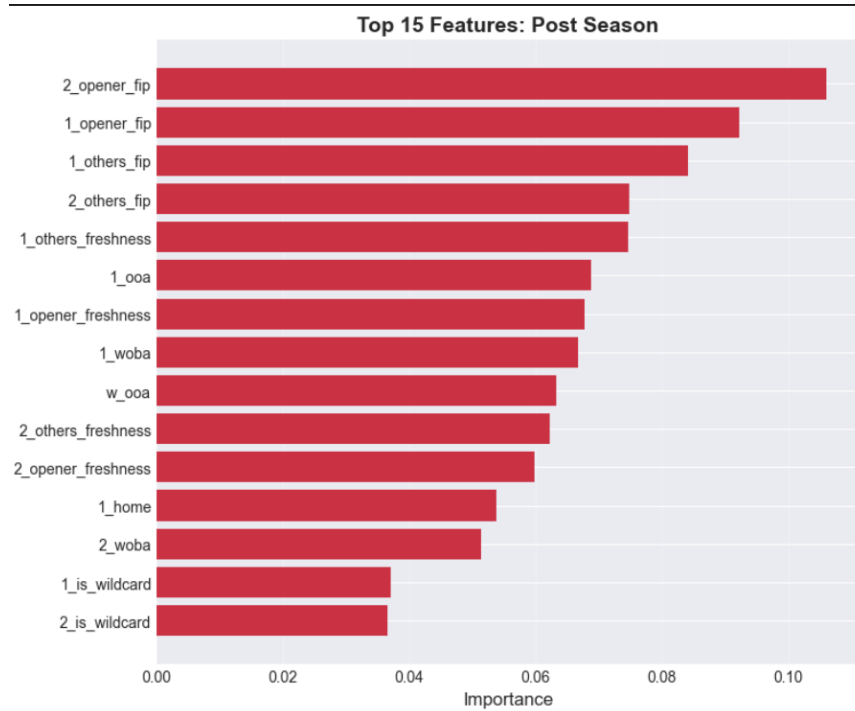

FIGURE 5. Feature importance for the postseason

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 hypthesis that Wild Card teams would perform better than division winners was not supported by our findings.

References

[Bas25] Baseball Prospectus. Postseason analysis archive. `https://www.baseballprospectus.com`, 2025. Accessed November 25, 2025.

[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.