THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

# Projecting MLB Pitcher Performance with Pitch Grades

MALCOLM ZERBE
SPRING 2024

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Data Sciences
with honors in Statistics

Reviewed and approved* by the following:

Matt Slifko
Assistant Teaching Professor of Statistics
Thesis Supervisor

Andrew Wiesner
Teaching Professor of Statistics
Honors Adviser

* Electronic approvals are on file.

# ABSTRACT

Baseball has historically been a game deeply rooted in statistics and numerical performance evaluation. From the batting averages of Ted Williams and home run totals of Hank Aaron to the earned run averages of Clayton Kershaw and on-base percentages of Juan Soto, the way we understand and interpret player performance has continuously evolved. Today, more advanced methods of quantifying player performance and the rise of advanced data mining have ushered in a new era of baseball analytics.

This thesis ventures into the forefront of this new era, focusing on the predictive analysis of starting pitcher performance in Major League Baseball. With the advent of model-based pitch grades estimating both the quality of the physical characteristics of pitches and of the location of pitches, there is an opportunity to enhance our predictive capabilities. This study aims to bridge traditional baseball wisdom with the cutting-edge techniques of machine learning, to see if these new metrics can improve the accuracy of future-season starting pitcher performance predictions. By building and comparing projection models both with and without these new metrics, this research seeks to unveil their true predictive power and potentially redefine how we project pitching performance.

# TABLE OF CONTENTS

LIST OF FIGURES ..... iii

LIST OF TABLES ..... iv

ACKNOWLEDGEMENTS ..... v

Chapter 1 Introduction to Baseball Analytics ..... 1

- The Evolution of Baseball Analytics ..... 1
- Introduction to Model-Based Baseball Metrics ..... 3
- Research Gap and Thesis Objective ..... 7
- Thesis Structure ..... 8

Chapter 2 Related Works ..... 9

- Data Analytics Effects in Major League Baseball ..... 9
- Predicting Baseball Pitcher Efficacy Using Physical Pitch Characteristics ..... 10
- Prediction of Future Offensive Performance of MLB Position Players ..... 11

Chapter 3 Data Sources and Metric Glossary ..... 13

Chapter 4 Data Preprocessing & Exploratory Data Analysis ..... 18

Chapter 5 Modeling ..... 25

Chapter 6 Results and Findings ..... 29

Chapter 7 Conclusion ..... 34

REFERENCES ..... 36

# LIST OF FIGURES

Figure 1. Run Expectancy Matrix ..... 4

Figure 2. RE24 Formula ..... 5

Figure 3. Formulation of botOvr ..... 6

Figure 4. Structure of Final Dataset ..... 19

Figure 5. Reliability of ERA- and FIP- ..... 20

Figure 6. Correlations Between Traditional Metrics, Sabermetrics, and Future FIP- ..... 21

Figure 7. Relationship Between xFIP- and Future FIP- ..... 21

Figure 8. Correlations Model-Based Metrics and Future FIP- ..... 22

Figure 9. Relationship Between botCmd and Future FIP- ..... 23

Figure 10. Correlations Between Features ..... 24

Figure 11. Formula for Standard Error ..... 30

Figure 12. Distribution of Cross-Validation Errors ..... 31

Figure 13. Feature Importances for Model "Without" ..... 32

Figure 14. Feature Importances for Model "With" ..... 33

# LIST OF TABLES

Table 1. Traditional Pitching Metrics ..... 14

Table 2. Early Sabermetrics ..... 15

Table 3. Model-Based Metrics ..... 17

Table 4. General Form of Models ..... 28

Table 5. Cross-Validation Errors ..... 29

## ACKNOWLEDGEMENTS

I would like to thank Dr. Matt Slifko for guiding me throughout this research project. From organizing our weekly discussions to providing invaluable insights into the statistical methods used, this paper would not have been possible without him. I would also like to thank Dr. Andrew Wiesner and Dr. David Hunter for reviewing this thesis.

Most importantly, I want to thank my dad, my grandparents, Nan and Pop, and my Uncle Matt for everything they do for me. They truly are the best support system I could ask for.

# Chapter 1

## Introduction to Baseball Analytics

Baseball's relationship with statistics dates to the game's earliest days, with simple metrics like batting average and runs batted in dominating the scene. The statistical revolution in baseball, however, truly began with the advent of sabermetrics in the late 20th century. Pioneered by Bill James, sabermetrics shifted the focus from traditional, often inaccurate stats to more in-depth metrics that better captured a player's value and performance [9]. This marked the beginning of a new analytical era, where data-driven decision making began to reshape strategies both on and off the field.

## The Evolution of Baseball Analytics

Throughout the evolution of baseball analytics, ERA has remained a cornerstone metric. It represents the average number of earned runs a pitcher allows over nine innings and has been a primary measure of a pitcher's effectiveness. Despite its longstanding use, ERA, like many traditional metrics, offers a limited view, often failing to account for the nuanced aspects of pitching performance. ERA fails to isolate the performance of an individual pitcher as it includes the contributions of the defensive unit playing behind him. A group of pitchers can allow batted balls with identical exit velocities, launch angles, and positions on the field, and the results can range from an out or a fielding error to a single or even an extra-base hit. With the rise of sabermetrics came the introduction of better, more sophisticated attempts to measure pitcher effectiveness. This shift was marked by the adoption of Fielding Independent Pitching (FIP) and

Expected Fielding Independent Pitching (xFIP), which represent a significant leap from traditional ERA.

FIP emerged as an innovative metric, concentrating on elements of performance that a pitcher more directly controls–strikeouts, walks, hit-by-pitches, and home runs–rather than balls in play, whose results are largely random and out of the hands of the pitcher. By isolating these factors, FIP presents a clearer view of a pitcher's performance, independent of the performance of the defense playing behind them. Building upon FIP, xFIP adjusts for the variability of home run rates by normalizing the home run component to a league-average home run-to-fly ball ratio. This adjustment is based on the understanding that pitchers have limited control over home runs once a fly ball is hit.

The ERA-estimators of FIP and xFIP are both scaled to league-average ERA, meaning that the league average FIP and xFIP are identical to league average ERA for any given season. Both metrics also disregard the sequence of events within an inning. For example, allowing a home run after walking two batters results in three earned runs and allowing two walks after allowing a home run only results in one earned run, greatly affecting ERA. FIP and xFIP treat the order of these events as random and thus the performance of these pitchers to be identical.

By making these adjustments, FIP and xFIP better isolate the performance of pitchers. Later shown in Chapter 4, these metrics are much more stable year-to-year than ERA, with FIP and xFIP having stronger correlations to next season FIP and xFIP than ERA does with next season ERA. This suggests that they are stronger measurements of skill, which theoretically should not drastically vary from one season to the next for the league as a whole. These metrics also have higher correlations with next-season ERA than ERA itself. As such, they have become the heart of ERA projection models currently in production, serving as key indicators of a

pitcher's true talent level, distinct from the variability and inconsistency inherent in traditional ERA.

## Introduction to Model-Based Baseball Metrics

The potential of sabermetrics drastically increased with the introduction of Statcast in 2015. Statcast, a system of camera and radar systems used for tracking movement on the field, was installed in all 30 MLB ballparks [9]. The system was further improved upon in 2020 with the incorporation of Hawk-Eye cameras. This state-of-the-art technology has allowed baseball teams, researchers, and fans to measure quantities such as the spin rate and release point of pitches and the exit velocity and launch angle of batted balls.

With the influx of Statcast data, baseball analytics underwent a revolutionary transformation. This wealth of data has enabled the creation of statistics that emphasize a player's approach and process, rather than just the outcomes on the field. Among these are machine learning model-based pitch grades that quantify the quality of "stuff," or physical characteristics, location, and the overall combination of the two. This shift towards a more nuanced understanding of player performance represents a significant evolution in how the game is analyzed.

Stuff grades aim to determine the effectiveness of various combinations of velocity, movement, and release points based on pitches thrown in the past. Using machine learning, stuff grades can account for the non-linear relationships between these variables to determine the expected quality of a given pitch. Location grades evaluate a pitcher's ability to place the ball in regions that historically produce beneficial results, independent of "stuff." Overall pitch grades

combine the two of them, with a separate model estimating the quality of a pitch based both on its physical characteristics and its location.

Two versions of stuff, location, and pitching grades are publicly available: Stuff+ and botStf, Location+ and botCmd, and Pitching+ and botOvr. While there are nuances in how both versions are calculated, their general purposes are shared. Stuff+ and botStf quantify how "nasty" a pitch is, Location+ and botCmd quantify how effective the location of a pitch is, and Pitching+ and botOvr quantify how well the combination of "nastiness" and location is expected to perform at the professional level.

To better understand these model-based pitch grades, it is crucial to understand what a run value is. In short, every moment of a baseball game has a "base-out state" describing the current number and placement of runners on base and number of outs in the half inning. There are 24 possible combinations of base and out states. The average number of runs scored given the base-out state is known as the run expectancy. Figure 1 shows a basic run expectancy matrix [14], with a base-out state of bases loaded and no outs having the highest run expectancy and bases empty and two outs having the lowest run expectancy.

| Runners | 0 Outs | 1 Out | 2 Outs |
|---|---|---|---|
| Empty | 0.461 | 0.243 | 0.095 |
| 1 _ _ | 0.831 | 0.489 | 0.214 |
| _ 2 _ | 1.068 | 0.644 | 0.305 |
| 1 2 _ | 1.373 | 0.908 | 0.343 |
| _ _ 3 | 1.426 | 0.865 | 0.413 |
| 1 _ 3 | 1.798 | 1.140 | 0.471 |
| _ 2 3 | 1.920 | 1.352 | 0.570 |
| 1 2 3 | 2.282 | 1.520 | 0.736 |

**Figure 1. Run Expectancy Matrix**

Using run expectancy, run values can be generated. Every plate appearance takes an inning from one base-out state to another base-out state. For example, according to Figure 1, a leadoff single would change the run expectancy of the half inning from 0.461, the beginning state of bases empty and no outs, to 0.831, the end state of runner on first and no outs. Figure 2, shows the formula for run expectancy based on 24 base-out states, or RE24, which considers the beginning and end state, as well as the number of runs scored on the play [14]. The hypothetical leadoff single would have had an RE24 of 0.370 (0.831 – 0.461 + 0). The ideas of run expectancies and run values can be extended to base-out-count states, expanding the run expectancy matrix to include all 12 possible pitch counts and changing RE24 to RE288. Various forms of RE288 run values are at the heart of these model-based pitch grades.

$$RE24 = RE\ End\ State - RE\ Beginning\ State + Runs\ Scored$$

**Figure 2. RE24 Formula**

Cameron Grove, the creator of PitchingBot, goes into detail as to how botOvr is calculated [7]. Grove creates a series of five XGBoost classification models for each group of pitch types: fastballs, breaking balls, and offspeed pitches. As seen in Figure 3, for each group of pitches, the models collectively predict the probability of all possible events–called strike, called ball, hit by pitch, swinging strike, foul ball, and 15 types of balls in play. The probability of each event occurring is multiplied by the average run value of that event occurring, resulting in an expected run value of a pitch. The metric is then transformed to a more interpretable scale.

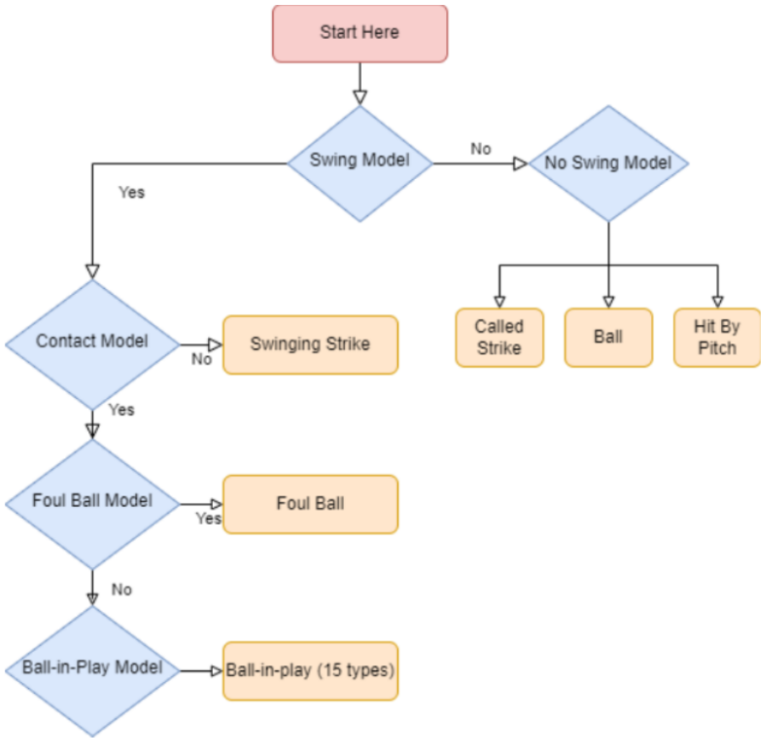


**Figure 3. Formulation of botOvr**

A similar process is done for botStf and botCmd. Eno Sarris and Max Bay, the founders of Stuff+, Location+, and Pitching+, also train their machine learning models using run values [8]. Both these metrics and Grove's aim to identify the velocities, movement profiles, release points, and pitch locations that historically generate results most beneficial to pitchers, like swings and misses and weak ground balls. While the metrics attempt to measure the same skills, Grove uses classification models to estimate probabilities of events, later multiplying these by their respective run values, and Sarris and Bay use regression models to estimate the run values of individual pitches directly.

Using machine learning, these metrics seem to offer a fresh perspective, distinct from other existing metrics. Unlike FIP and xFIP, which are calculated in very similar ways, pitch

quality grades introduce new, process-oriented information. Integrating these grades with existing metrics could potentially yield projections superior to either approach independently.

## Research Gap and Thesis Objective

While the integration of advanced metrics like stuff, location, and overall pitching grades into baseball analytics has provided new dimensions to player evaluation, there remains a research gap in the application of these metrics for predictive purposes. Traditional performance estimators, such as FIP and xFIP, have been the cornerstone of pitcher performance prediction. However, these estimators primarily focus on the outcomes of pitches and do not fully encompass the nuanced aspects captured by the new metrics. This gap highlights a potential area of exploration: how well do these model-based pitch grades metrics, based on the inherent qualities of pitches, predict future pitching performance compared to existing metrics?

The primary objective of this thesis is to bridge this research gap by evaluating the efficacy of stuff, location, and pitching grades as predictive tools for a starting pitcher's future performance, specifically their next-season FIP-, a version of FIP described in Chapter 3 that accounts for the changing league average FIP each year. This study aims to construct and analyze FIP- projection models that incorporate these new metrics with traditional metrics and early sabermetrics, comparing them against models that rely solely on the traditional metrics and early sabermetrics alone. By doing so, the research seeks to determine whether these new model-based metrics offer significant predictive advantages and can redefine our understanding of pitcher projection. The goal is to provide a comprehensive analysis that could reshape how

analysts and teams forecast pitching performance, leveraging the full potential of modern baseball analytics.

## Thesis Structure

To tackle this question as to whether model-based stuff, location, and overall pitching grades can positively contribute to the prediction of a starting pitcher's future FIP-, the thesis takes the following form. Chapter 2 investigates the current literature surrounding the presence of data analytics in professional baseball front offices, using pitch metrics to explain pitcher performance, and projecting MLB performance. Chapter 3 describes the primary data sources, FanGraphs and Baseball Savant, and defines the metrics that will be utilized in future chapters.

Chapter 4 preprocesses the data for modeling before engaging in exploratory data analysis to better understand the relationships between the features and future FIP- as well as the relationships between the features themselves. Chapter 5 uses robust tree-based supervised machine learning techniques to hyperparameter tune and train two final models: one including stuff, location, and pitching grades, and one without them. Null and naïve models are also built as performance baselines for the more advanced models.

Chapter 6 is a comparative analysis between the two final models, as well as the null and naïve models, investigating the out-of-sample RMSE error of each. Variable importance are calculated to determine which, if any, of the new metrics was the most significant in predicting future FIP-. Chapter 7 will draw conclusions as to whether these new metrics provide any useful signal in projecting future FIP- and dive into several ideas as to how to improve this study in future iterations.

# Chapter 2

## Related Works

Having introduced the significance of FIP and other performance estimators in baseball analytics and the potential of new metrics like stuff, location, and pitching grades, Chapter 2 delves into the existing literature within the realm of baseball analytics. It explores previous studies that have applied machine learning techniques to various aspects of baseball, particularly focusing on player performance. While these studies have laid a foundation in the field, they also highlight the notable research gap in pitcher performance prediction, setting the stage for the unique contribution of this thesis in advancing baseball analytics.

### Data Analytics Effects in Major League Baseball

The integration of data analytics in Major League Baseball has significantly altered team strategies and competitive dynamics. A study by Elitzer (2020) [6] underscores this transformation. By constructing a linear least squares regression model, Elitzer aimed to predict MLB team wins in relation to their financial expenditure. The model included variables such as team Wins Above Replacement (WAR), team payroll relative to the league average for the season, and the presence of a data analytics department within their team structure. The study revealed a statistically significant positive relationship between the existence of a data analytics department and the number of team wins per money spent. This finding supports the notion that data analytics departments provide tangible strategic advantages in the MLB.

However, Elitzer's research also indicated a crucial caveat: the advantage of sabermetric innovations diminishes as they become publicly known and widely adopted. This suggests that

the edge gained from data analytics is not static; rather, it requires continuous advancement and innovation to maintain its value.

This study is particularly relevant to the present thesis, which explores the predictive capabilities of advanced metrics like stuff, location, and pitching grades in forecasting pitchers' FIP-. Elitzer's findings highlight the necessity for continual advancement in baseball analytics to sustain competitive benefits. This reinforces the importance of investigating new analytical methods and metrics. By examining whether these advanced metrics can enhance FIP- predictions, this thesis contributes to the evolving landscape of baseball analytics, seeking to provide teams with a fresh competitive edge through innovative statistical analysis.

**Predicting Baseball Pitcher Efficacy Using Physical Pitch Characteristics**

Tejas Oberoi's study (2024) [10], "Predicting Baseball Pitcher Efficacy Using Physical Pitch Characteristics," represents a significant piece in the use of machine learning to predict baseball pitcher performance. The research focuses on sixteen pitch-level features, such as pitch velocity and spin rate, to forecast key pitching metrics including Walks/Hit per Inning Pitched (WHIP), Batting Average Against (BAA), and FIP. By applying neural network and linear regression models, the study examines how each feature influences a pitcher's effectiveness.

Although the study yielded notable results, particularly with the `ballFrequency` feature in predicting WHIP, it found that these features alone accounted for less than 50% of the variance in the pitcher efficacy metrics. This outcome underscores the complexity of pitcher performance and the challenges in capturing it through individual pitch characteristics alone.

Oberoi's approach, focusing on physical pitch characteristics and pitch location frequency, parallels the objectives of this thesis, which aims to understand pitcher success using more nuanced and robust measurements of stuff, location, and overall pitching grades. The findings from Oberoi's research suggest that while physical characteristics and location frequency provide valuable insights, they might not fully encapsulate a pitcher's skill. This thesis extends this exploration by assessing whether combining these advanced metrics with established performance estimators like FIP- and xFIP- enhances the predictability of future FIP-. By doing so, it seeks to contribute a more comprehensive understanding of pitcher performance prediction within the framework of modern baseball analytics.

### Prediction of Future Offensive Performance of MLB Position Players

In the study "Prediction of Future Offensive Performance of MLB Position Players" by Benavidez et al. (2019) [4], the focus was on predicting the future offensive performance (OPS+) of MLB position players using machine learning models. This research, employing techniques like support vector regression and recurrent neural network, aimed to forecast player performance based on past data, demonstrating the potential of data analytics in predicting offensive capabilities in baseball.

While this study represents a significant stride in offensive player analytics, it highlights a notable gap in similar research efforts focused on pitchers. Unlike position players, pitcher performance, especially in terms of FIP-, has not been extensively explored through advanced analytical methods in scholarly literature.

This gap underscores the novelty and necessity of the current thesis, which aims to apply advanced metrics and machine learning techniques to predict FIP- for pitchers. By exploring this under-researched area, this thesis seeks to extend the application of machine learning from offensive performance prediction to pitcher performance prediction, potentially offering new insights into the predictability and evaluation of pitchers in Major League Baseball.

# Chapter 3

## Data Sources and Metric Glossary

Most data for this research can be found on FanGraphs [1], a website that hosts a plethora of professional baseball content, namely blogs and statistics. Data was also collected from MLB's Baseball Savant [2], another well-known source for advanced baseball statistics. Season-level metrics can easily be exported from both sites for further analysis. This research will distinguish these metrics into three groups: traditional metrics, early sabermetrics, and new model-based metrics. To reiterate, the purpose of this thesis is to determine whether the new model-based metrics can improve FIP- projections, which are currently based on traditional metrics and early sabermetrics.

Traditional metrics have existed since the dawn of baseball and are household names to all baseball fans. These are statistics that are spoken about on broadcasts and can be found on the backs of baseball cards. They are all relatively primitive in their calculations, but many still hold very useful information. Table 1 defines many of the most common traditional metrics.

As mentioned in Chapter 1, the limitations of these metrics motivated the creation of early sabermetric measurements, like FIP and xFIP. Definitions for these metrics from PitcherList [11], a blog site about MLB pitching, can be found in Table 2. They were designed to better isolate the contributions of individuals, independent of both the performance of others and of batted ball randomness. As a result, they tend to be more predictive of future on-field results and are at the heart of modern predictive models.

## Table 1. Traditional Pitching Metrics

| *Metrics* | *Definition* |
|---|---|
| **Earned Run Average (ERA)** | The number of earned runs allowed per nine innings pitched |
| **Strikeout Percentage (K%)** | Percentage of batters faced that were struck out |
| **Walk Percentage (BB%)** | Percentage of batters faced that were walked |
| **Strikeout Minus Walk Percentage (K-BB%)** | The difference between K% and BB% |
| **Walks and Hits Per Inning Pitched (WHIP)** | Same as the name, walks and hits per inning pitched |
| **Ground Ball Percentage (GB%)** | Percentage of balls in play that were ground balls |
| **Strike Zone Percentage (Zone%)** | Percentages of pitches thrown in the strike zone, regardless of the pitch call or outcome |
| **First Pitch Strike Percentage (F-Strike%)** | Percentage of batters faced that begin plate appearance with a strike or a ball in play |
| **Swinging Strike Percentage (SwStr%)** | Percentage of pitches thrown that result in a swing and miss |
| **Whiff Percentage (Whiff%)** | Percentage of pitches thrown and swung at that result in a swing and miss |
| **Called Strikes Plus Whiffs (CSW%)** | Percentage of pitches thrown that result in either a called strike or a whiff |
| **Fastball Velocity (vFB)** | Average velocity of fastball, in miles per hour |

**Table 2. Early Sabermetrics**

| *Metrics* | *Definition* |
|---|---|
| **Fielding Independent Pitching (FIP)** | FIP focuses solely on the events a pitcher has the most control over—strikeouts, unintentional walks, hit-by-pitches, and home runs. It entirely removes results on balls hit into the field of play |
| **Expected Fielding Independent Pitching (xFIP)** | xFIP takes a pitcher's FIP, but it uses a projected home-run rate instead of actual home runs allowed. The home-run rate is determined by that season's league HR/FB rate |
| **Skill-Interactive Earned Run Average (SIERA)** | SIERA quantifies a pitcher's performance by trying to eliminate factors the pitcher can't control by himself. For instance, SIERA punishes high-walk pitchers more for each additional walk than regular pitchers. It rewards high-strikeout, high ground-ball, and high-fly-ball pitchers for each additional strikeout, ground ball, and fly ball, respectively |
| **Expected Earned Run Average (xERA)** | Expected ERA, or xERA, is a simple 1:1 translation of Expected Weighted On-Base Average (xwOBA), converted to the ERA scale. xwOBA takes into account the amount of contact (strikeouts, walks, hit by pitch) and the quality of that contact (exit velocity and launch angle) |

It is worth noting that all traditional metrics and early sabermetrics have alternate forms on the "Plus" or "Minus" scales. They compare values for the respective metric to the league average for that metric, forcing the average to be equal to 100. For example, the league average ERA is not the same every year. An ERA of 4.50 may be average one year and below average the next. To account for this, ERA- (Earned Run Average Minus) is used. An ERA of 4.50 with a league average ERA of 4.50 would be given a grade of 100. That same ERA with a lower league average ERA of 3.50 would be assigned a number over 100. Every integer above or below 100

represents a percentage point better or worse than league average in that metric. When the metric is labeled “Minus”, scores lower than 100 are better. When the metric is labeled “Plus”, scores higher than 100 are better. For example, an ERA- of 80 would signify an ERA that is 20% better than league average and a K%+ of 80 would signify a K% that is 20% worse than league average. Additionally, all Minus and Plus metrics are park-adjusted for the run environments.

As mentioned in Chapter 1, Statcast made possible the creation of a whole new suite of metrics: the model-based metrics. These metrics come from a different data source than the early sabermetric statistics. While the early sabermetric statistics can be calculated using box score data (number of strikeouts, walks, hit by pitches, etc.), the model-based metrics come from Hawkeye’s pitch-by-pitch tracking system, capturing important variables like velocity, movement, and location coordinates for every pitch. The model-based metrics and their definitions can be found in Table 3. Stuff+, Location+, and Pitching+ share the same scale as the Plus metrics just mentioned, with an average value of 100, but botStf, botCmd, and botOvr differ slightly. Grove designed these metrics to be on the 20-80 scale, with 20 representing the worst possible grade and 80 the best possible grade, a scale traditional baseball scouts are very familiar with. This thesis tests to see if the new information contained within pitch quality grades can improve player projection capabilities, more so than just the traditional metrics and early sabermetrics alone. The model-based metrics will also be referred to as pitch grades throughout this paper.

## Table 3. Model-Based Metrics

| *Metrics* | *Definition* |
|---|---|
| **Stuff+** | Model-based estimate of quality of "stuff" (velocity, movement, and release points) by Eno Sarris and Max Bay |
| **Location+** | Model-based estimate of quality of location, adjusted by pitch type and pitch count, by Eno Sarris and Max Bay |
| **Pitching+** | Model-based estimate of overall pitch quality ("stuff" and location) by Eno Sarris and Max Bay |
| **botStf** | Model-based estimate of quality of "stuff" (velocity, movement, and release points) by Cameron Grove |
| **botCmd** | Model-based estimate of quality of location, adjusted by pitch type and pitch count, by Cameron Grove |
| **botOvr** | Model-based estimate of overall pitch quality ("stuff" and location) by Cameron Grove |

# Chapter 4

## Data Preprocessing & Exploratory Data Analysis

Before exploring the data, a small amount of feature engineering needed to be performed. As mentioned in Chapter 3, most of the metrics used in this project have either a Minus or Plus form that places the value of the statistic in perspective to the league average of that statistic for the given season. Because the objective is to build a model that predicts future FIP, the later models need to account for the fact that the league average for FIP and many other metrics can significantly change from one season to the next. For this reason, FIP- serves as the target variable, not FIP, and all features will also be on either the Plus or Minus scale to account for changing league averages. K-BB%, SIERA and xERA are unavailable in Minus or Plus form, so they were manually converted to that scale by multiplying the quotient of each player's value for the metric and the league average for the metric in that season by 100.

In this analysis, only seasons in which a starter recorded 60 IP in consecutive seasons were considered. Because the objective of this research is to determine whether these new metrics hold predictive power, the chosen innings threshold removes observations with limited data, ensuring that the question at hand is answered. After filtering observations with sufficient data, 257 data points and 12 features remained for modeling. A glimpse of the final dataset, with a subset of observations and features, can be seen in Figure 4.

| fip_minus_future | era_minus_past1 | fip_minus_past1 | stuff_plus_past1 | location_plus_past1 | pitching_plus_past1 |
|---|---|---|---|---|---|
| 70.30886 | 69.43405 | 71.11521 | 109.89440 | 105.46111 | 109.59886 |
| 78.81276 | 92.28978 | 94.63180 | 95.00400 | 103.33484 | 101.50121 |
| 95.09134 | 68.59669 | 78.86463 | 112.07931 | 103.03480 | 107.38671 |
| 80.12109 | 91.22602 | 77.43127 | 97.33237 | 107.36550 | 105.68247 |
| 66.18429 | 67.44175 | 49.61491 | 131.00123 | 102.02463 | 110.33232 |

**Figure 4. Structure of Final Dataset**

With all metrics of interest calculated, proper exploratory data analysis could be conducted. As a reminder, all data visualizations and descriptive statistics refer to the Plus or Minus versions of each metric and consider only observations with at least 60 IP in consecutive seasons. Firstly, the reliability, or the year-to-year correlation, of ERA- and FIP- is explored to illustrate the decision to predict FIP- rather than ERA-. It is theorized that metrics with less variation from one year to the next are better indicators of player skills. As stated by former MLB Data Scientist Sam Sharpe, “Statistics that demonstrate high correlations from year to year are useful in evaluating player skills since players generally tend to retain their core skills from year to year” [12]. Figure 5 demonstrates the superior stability of FIP-.

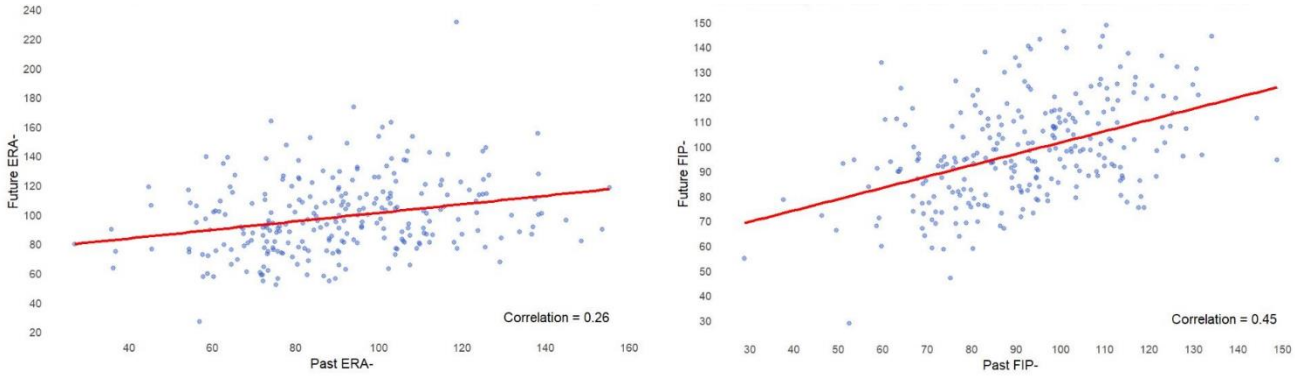


**Figure 5. Reliability of ERA- and FIP-**

With the goal being to predict future FIP-, the relationships between future FIP- and all past metrics were investigated. As shown in the correlation matrix in Figure 6, among the traditional metrics and early sabermetrics, past xFIP- and past SIERA- appear to show the strongest relationships with future performance on the mound. This is not surprising as, as mentioned in Chapter 2, these metrics were created to better represent pitcher skill. As a result, they show stronger correlations with future performance than traditional metrics like ERA- and K-BB%+. Intuitively, one would expect a pitcher to perform closer to their true talent level than anything else.

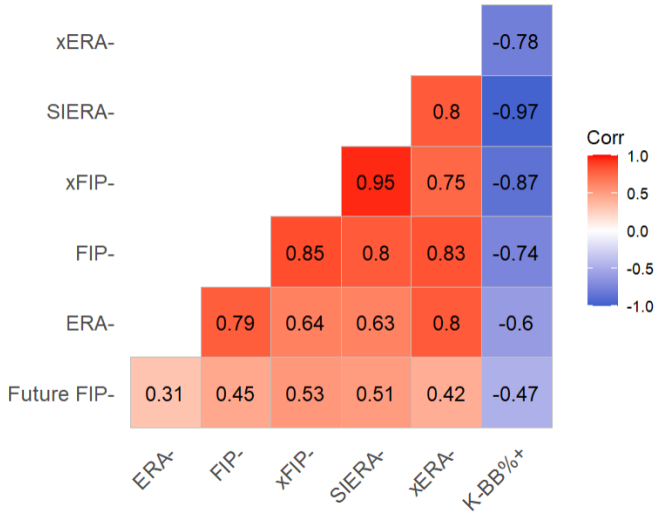


**Figure 6. Correlations Between Traditional Metrics, Sabermetrics, and Future FIP-**

To further explore the relationships between the past metrics and future performance, scatterplots were generated. Figure 7 depicts the strongest relationship of the bunch between past xFIP- and future FIP-. This plot, as well as the others, shows a moderate, linear relationship between past performance and future performance.

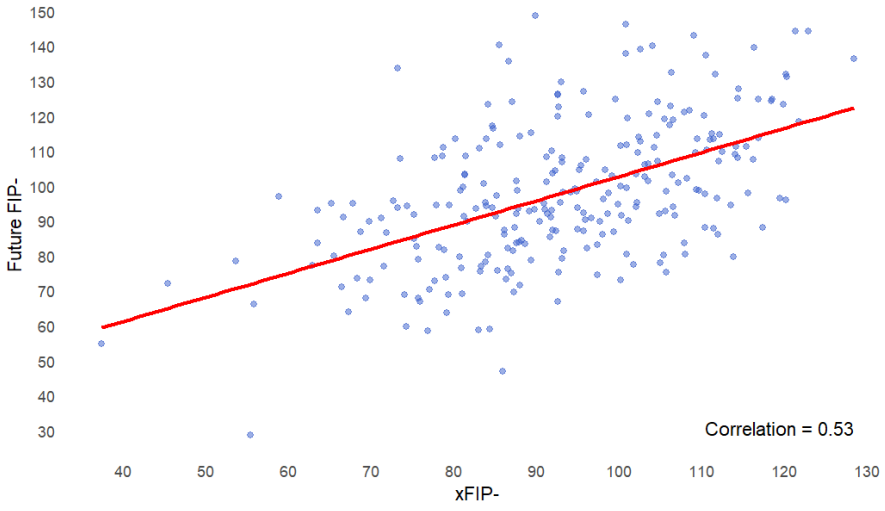


**Figure 7. Relationship Between xFIP- and Future FIP-**

This process was repeated with the new, model-based metrics. The resulting correlations between the model-based metrics and future FIP- are shown in Figure 8. Stuff+, botStf, Pitching+, and botOvr all showed relative promise with correlations ranging from -0.42 to -0.48. On the contrary, Location+ and botCmd each displayed very weak relationships with future performance, with correlations of -0.15 and -0.13, respectively. The random scatter shown in Figure 9 seems to suggest that knowledge of a pitcher's past location skills is not useful for projecting future success.

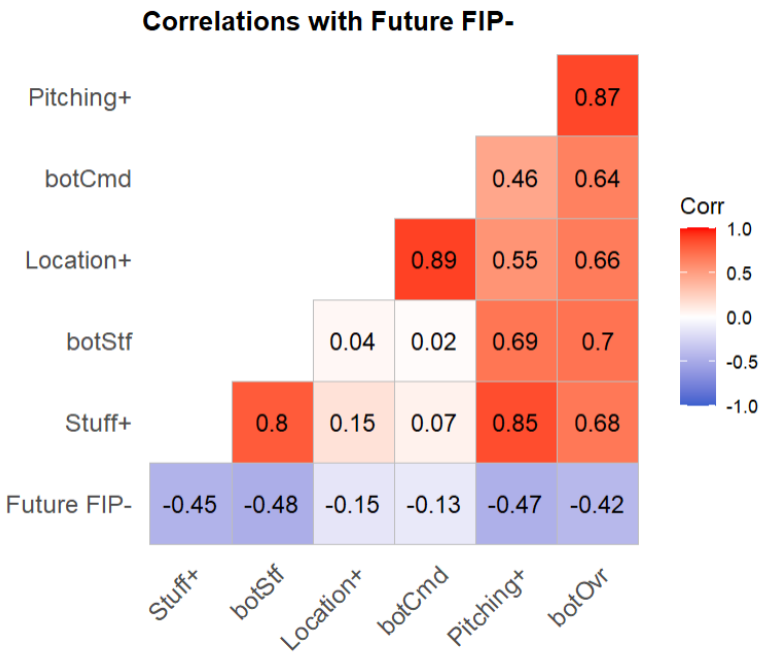


**Figure 8. Correlations Model-Based Metrics and Future FIP-**

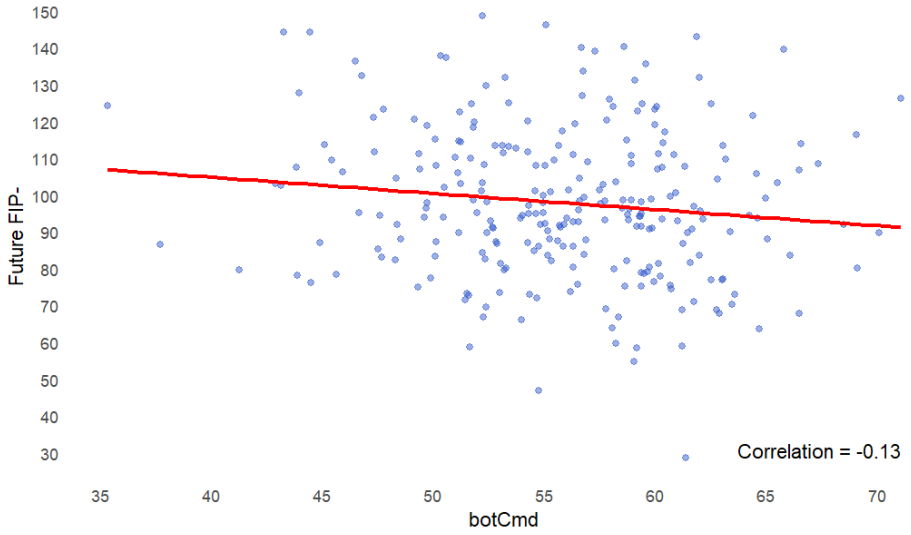


**Figure 9. Relationship Between botCmd and Future FIP-**

It is also important to note the strong multicollinearity concerns between these past metrics, regardless of whether they are traditional metrics, early sabermetric, or model-based metrics. For instance, as shown in Figure 10, xFIP- and SIERA- share an extremely high correlation of 0.95. It is very likely that all these features share similar information and that including xFIP- or SIERA- alone would likely account for all of them. Similarly, the pairs of pitch grades are strongly correlated, ranging from a coefficient of 0.80 (Stuff+ and botStf) to 0.89 (Location+ and botCmd).

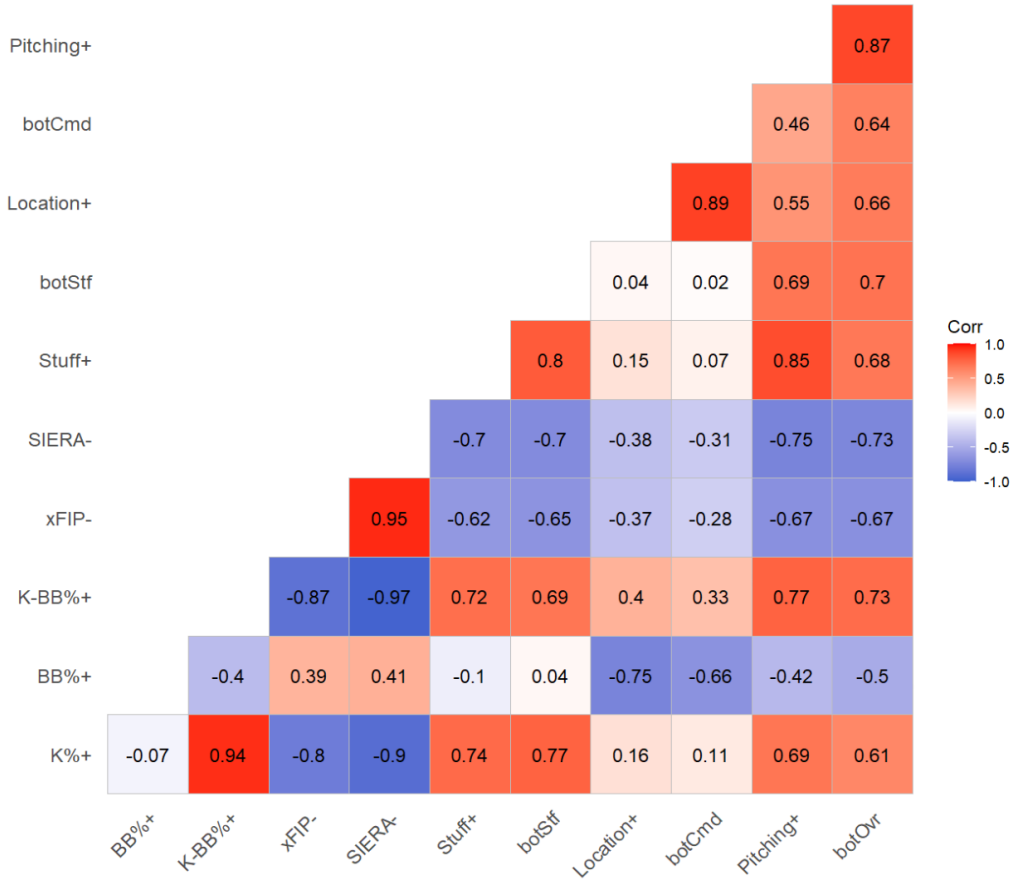


**Figure 10. Correlations Between Features**

This research hypothesized that these metrics, derived from a different data source using an entirely different method, would offer new predictive information. It is possible, however, that this is not the case. As shown in Figure 10, the stuff grades and overall grades, which showed some of the stronger correlations with future performance, are also very strongly correlated with information that already existed, like SIERA- and K-BB%+. Because only linear relationships have been explicitly explored, there is a chance that machine learning algorithms could uncover a more complicated relationship between these features.

# Chapter 5

## Modeling

After exploring the data, advanced modeling methods were tested to determine if the model-based pitch grades improved predictions of future FIP-. With a quantitative response variable, this supervised learning problem required a suitable regression algorithm. Renowned for its generalizability on tabular data and its ability to uncover highly complex and nonlinear relationships, boosted trees from XGBoost were chosen as the regression technique. Additionally, XGBoost is very compatible with generating feature importance scores, which will be helpful for estimating the contributions of the model-based metrics.

In the realm of machine learning, decision trees stand out for their intuitive logic and simplicity, serving as fundamental building blocks for more complex models. At their core, decision trees split data into subsets based on certain decision criteria from the features, creating a tree-like structure of decisions and outcomes. However, when decision trees are combined into a larger ensemble model, the predictive power can significantly increase. XGBoost, short for eXtreme Gradient Boosting, elevates this concept by employing a sophisticated ensemble technique known as gradient boosting, where multiple trees are trained sequentially [5]. Each tree in the sequence aims to correct the errors made by the previous ones, thereby improving the model's accuracy iteratively. This methodical refinement of trees, guided by gradient descent, allows XGBoost to deliver high precision with scalable efficiency.

Within the XGBoost framework, the flexibility to tune hyperparameters opens avenues for significant improvements in model performance. All models were fit using the tidymodels [3]

framework, which includes hyperparameters like `trees`, `min_n`, `tree_depth`, `learn_rate`, `sample_size`, and `mtry`.The `trees` parameter controls the number of trees in the ensemble, directly impacting the model's complexity and potential for overfitting or underfitting. `min_n` represents the minimum sum of instance weight (hessian) needed in a child, influencing the decision to make further partitions in a tree and thus affecting the model depth and complexity. `tree_depth` specifies the maximum depth of a tree, limiting the number of splits in each tree and consequently controlling the model's ability to learn detailed patterns. The `learn_rate` (or learning rate) adjusts the step size at each iteration of the boosting process, balancing the speed and accuracy of learning. `sample_size` determines the fraction of samples to be used for fitting the individual base learners, offering a way to reduce variance by introducing randomness into the model training process. Lastly, `mtry` represents the number of variables randomly sampled at each split when building a tree, contributing to the model's ability to handle feature selection and further diversifying the trees. By tuning these hyperparameters and more, one can fine-tune the XGBoost model's capacity to capture underlying data patterns while preventing overfitting, thereby achieving optimal performance. In this study, only the defined hyperparameters above will be considered for tuning.

Mean Squared Error (MSE) is a widely used evaluation metric in regression analysis, quantifying the average of the squares of the errors or deviations between predicted and actual observations. Specifically, MSE calculates the square of the difference between the predicted values generated by the model and the actual values within the dataset, then averages these squared differences over all observations. This metric is particularly useful because it penalizes larger errors more severely by squaring them, thus emphasizing the need for accuracy in predictions. To enhance the interpretability of the MSE, it can be transformed into the Root

Mean Squared Error (RMSE) by taking the square root of the MSE value. This transformation is beneficial because the RMSE metric is expressed in the same units as the original data, making it easier for practitioners to understand the magnitude of prediction errors. Unlike the squared units of MSE, which can be challenging to interpret in the context of the original data scale, RMSE directly reflects the average distance between the predicted and actual values. Therefore, RMSE provides a more intuitive sense of the average error magnitude, facilitating clearer comparisons between different models or datasets.

By analyzing MSE alongside the tuned hyperparameters, the study aims to identify the optimal configuration that achieves the most accurate predictions while avoiding overfitting. More specifically, we want to investigate the out-of-sample MSE, or the MSE of the model on observations that were not involved in model training. This focus on out-of-sample performance is crucial, as it provides an unbiased assessment of the model's ability to generalize to new, unseen data, ultimately ensuring the reliability and robustness of the predictive insights derived from the model.

Working with an incredibly small dataset (257 observations), it would be difficult to trust the out-of-sample error of any one validation or test set. To account for the limited amount of data, this study utilized 10-fold cross-validation. In this process, 10 separate error estimates were generated and then averaged, providing a more robust estimate of predictive power with every data point being assigned to the validation set exactly once. Bayesian optimization was used to intelligently search and test the parameter space for ideal values, with each hyperparameter setting being tested in cross-validation.

Our goal is to determine if the model-based metrics can improve projections of starting pitcher performance. To answer this question, multiple models were trained–one model for

predicting FIP- using only traditional metrics and early sabermetrics from the season before and one model for predicting FIP- using the traditional metrics, early sabermetrics, and model-based pitch grades. If the model with stuff, location, and overall pitch grades significantly outperforms the model without them, then it can be confidently concluded that these new metrics are useful in projecting the performance of starting pitchers.

In addition to advanced models, however, a null model and a naïve model were crafted. These models serve as performance benchmarks for the future tree-based models. Across the 10 folds, the null model simply used 100, the league average FIP-, as the prediction, simulating what randomly guessing performance would look like in practice. The naïve model took this one step further, using each pitcher's past xFIP-, the most correlated predictor with future FIP-, as the prediction for FIP-. Table 4 shows the general form of all models.

**Table 4. General Form of Models**

| *Model* | *Form* |
|---|---|
| **Null Model** | FIP- ~ 100 |
| **Naïve Model** | FIP- ~ Past xFIP- |
| **Model "Without"** | FIP- ~ Past Traditional Metrics + Past Early Sabermetrics |
| **Model "With"** | FIP- ~ Past Traditional Metrics + Past Early Sabermetrics + Past Model-Based Metrics |

# Chapter 6

## Results and Findings

All four models were evaluated using k-fold cross-validation. Using the 10 error estimates, the average out-of-sample MSE and RMSE errors were calculated for each model. For both MSE and RMSE, lower values signify better model performance, with a perfect model achieving an error of zero, indicating no deviation from the actual values and a stronger ability to minimize prediction errors.

As shown in Table 5, the null model performed the worst in terms of MSE and RMSE, with the naïve model, model "without" pitch grades, and model "with" pitch grades all providing moderate improvements.

**Table 5. Cross-Validation Errors**

| *Null Model* | *Naïve Model* | *Model "Without"* | *Model "With"* |
|---|---|---|---|
| 412.41 MSE<br>20.31 RMSE | 334.77 MSE<br>18.30 RMSE | 303.49 MSE<br>17.42 RMSE | 293.30 MSE<br>17.13 RMSE |

On average, the null model's prediction of FIP- was off by over 20 points. As a reminder, FIP- is scaled so that every integer below 100 is a percentage point better than league average and every integer above 100 is a percentage point worse than league average. In other words, the predictions of the null model were off by over 20%. The advanced models improved upon this, reducing the prediction error of FIP- to about 17%. Ultimately, the model that included the additional pitch grades achieved the lowest MSE and RMSE. To evaluate the legitimacy of this

improvement over the model without the additional pitch grades, standard errors (SE) were calculated for all models using the 10 estimates of each on the folds. Figure 11 shows the formula for SE, which divides the standard deviation of the 10 MSE values by the square root of the number of folds.

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

**Figure 11. Formula for Standard Error**

Using the standard error, an interval of MSE values was generated for each model. As seen in Figure 12, it seems that the advanced models made significant improvements over the null model, but neither significantly outperformed the naïve model's use of past xFIP- for predictions.

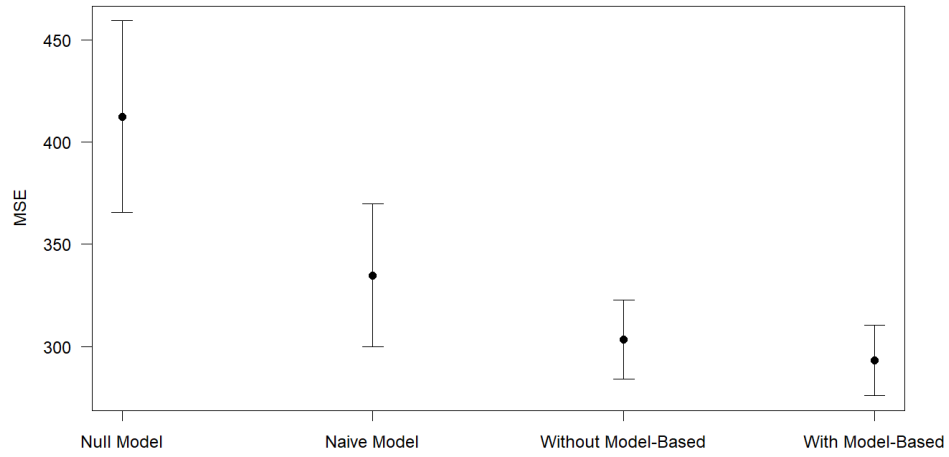


**Figure 12. Distribution of Cross-Validation Errors**

The lack of significant improvement in predictive accuracy following the addition of the pitch grades as model covariates likely stems from the feature multicollinearity discussed in Chapter 4. As previously shown in Figure 10, the stuff grades share a very strong linear relationship with a pitcher's ability to strike batters out and the location grades share a very strong linear relationship with a pitcher's ability to limit walks. It is likely that these metrics are providing information that was already contained within the traditional metrics and early sabermetrics.

Lastly, feature importances were explored for both advanced models. As shown in Figure 13, for the model that contained only the traditional metrics and early sabermetrics, xFIP-, FIP-, and SIERA- achieved the highest scores. On the other hand, ERA- provided the least amount of predictive signal for future FIP-. This lines up with the findings from exploratory data analysis, as past FIP-, xFIP-, and SIERA- were also much more strongly correlated with future FIP- than past ERA- was. This drives home the point that ERA is not a strong indicator of pitcher skill.

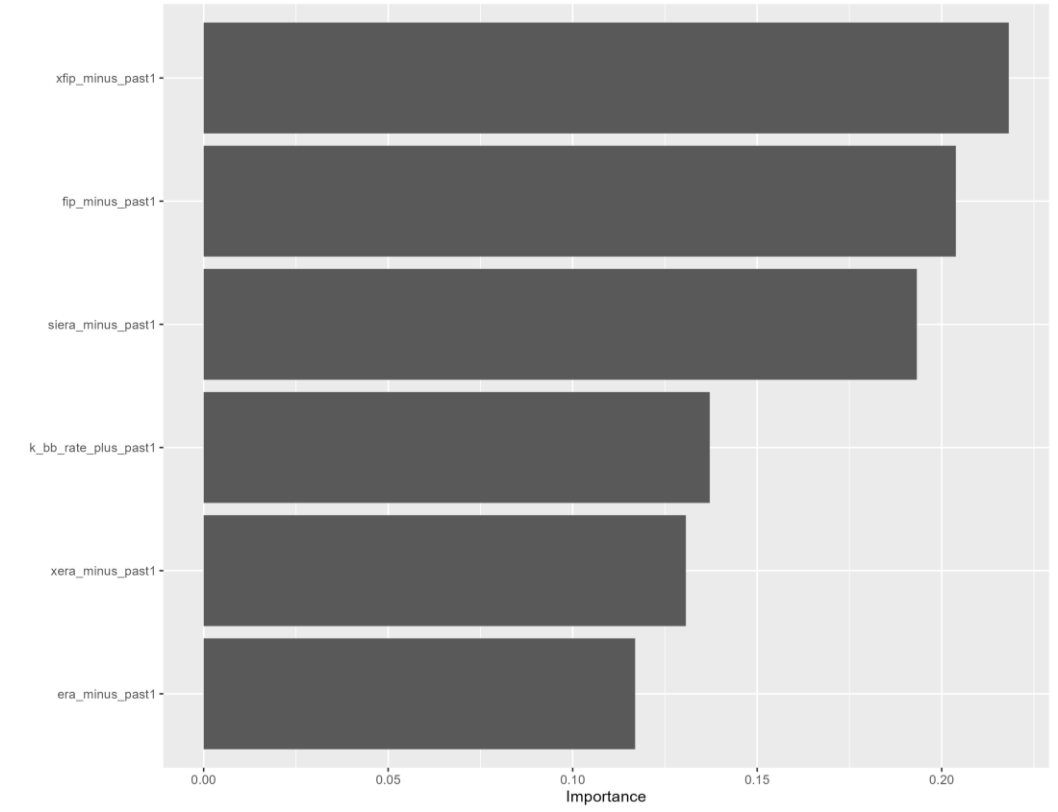


**Figure 13. Feature Importances for Model "Without"**

This idea remains constant after incorporating the model-based pitch grades, with much higher feature importance scores attributed to xFIP- and SIERA- than to ERA-. Additionally, as seen in Figure 14, the weak correlations between the location grades and future FIP- manifested in the two lowest feature importances. Interestingly enough, botStf achieved the highest score, though this metric is highly correlated with some of the other best-performing explanatory variables. For instance, there is a strong correlation coefficient of -0.7 between botStf and SIERA-.

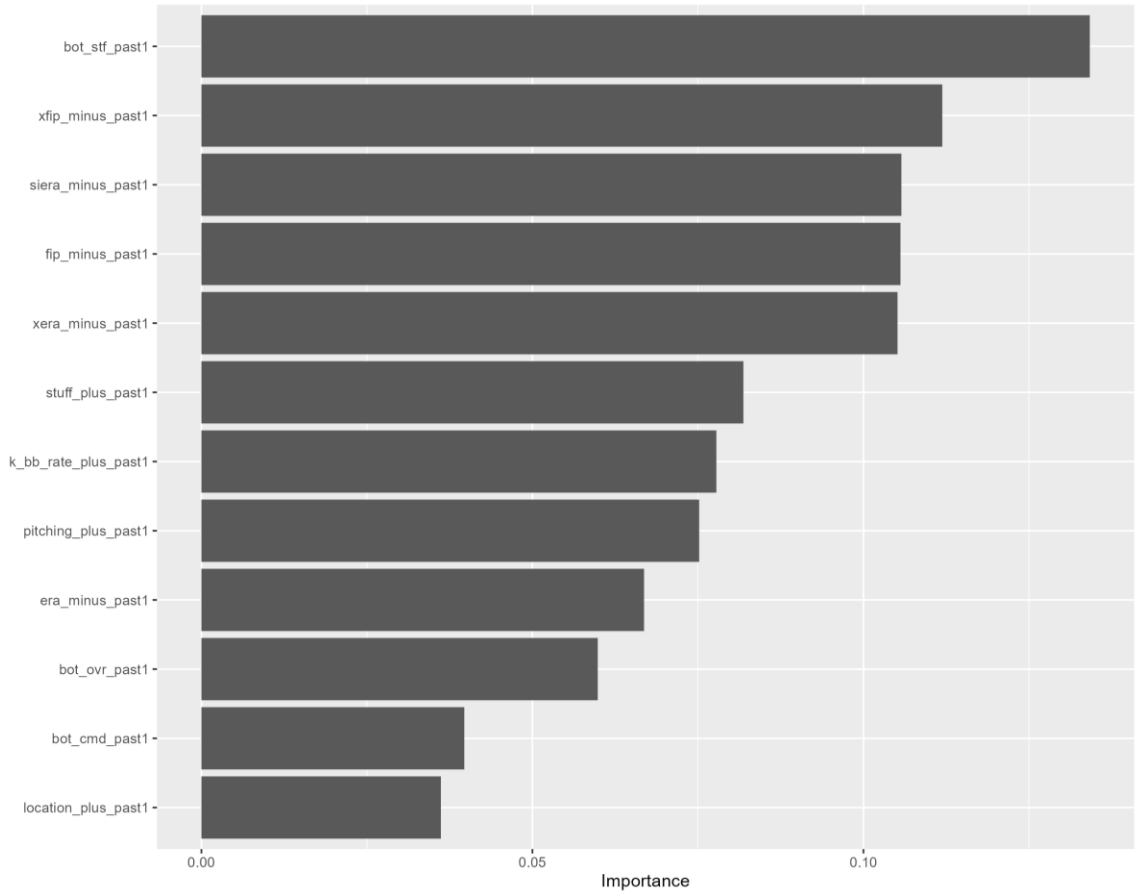


**Figure 14. Feature Importances for Model "With"**

# Chapter 7

## Conclusion

Forecasting pitcher performance is a challenging task. When predicting future FIP-, it is difficult to outperform merely using last season's xFIP-, which served as the naïve model in this paper. All advanced modeling techniques, both with and without the model-based pitch grades, failed to drastically outperform the naïve model. This is not surprising given what we know about the Marcel the Monkey Forecasting System. This method is named based on its primitive calculations, predicting future performance based on a weighted average of past performance [13]. Marcels tends to perform similarly to complex projection systems despite its simplicity. The performances of the XGBoost models with and without the pitch grades were similar to each other. Based on the preliminary exploratory data analysis and enforced by the results and intuition, it is likely that the information within "stuff", location, and overall pitching grades is already captured by existing metrics like K-BB%+ and xFIP-.

There are several avenues for future research regarding using pitch grades to improve pitcher performance projections. Because publicly available pitch grades only date back to 2020, the size of the data set was small (257 observations). However, the pitch data needed to calculate pitch grades, including the velocity, movement, and location of pitches, goes as far back as 2015. To increase the size of the dataset and to become increasingly confident in the conclusion, pitch grade models can be built from scratch and applied to all additional pitching seasons, expanding the observations from 2020-2023 to 2015-2023.

Secondly, this research focused only on pitchers with at least 60 IP in consecutive seasons. It is plausible that pitch grades provide much more predictive power for pitchers with very limited playing time in previous seasons. For example, according to Owen McGrattan, Stuff+ takes roughly 80 pitches to stabilize [8], which is much less than existing metrics like K-BB% and xFIP, which may take over half a season to become accurate representations of pitcher skill. Location+ and Pitching+ tend to settle at

around 400 pitches [8]. While these metrics may provide similar information as the previously existing ones, their unique ability to become accurate in small sample sizes may vastly improve projection models that also consider pitchers with limited past playing time, such as rookies, recent call-ups, and relief pitchers.

The pitch grades retain key uses despite their lack of additional predictive power. These metrics are crucial to player development, the field of improving athletes with appropriate training regiments. While metrics like xFIP and SIERA are just as useful as pitch grades in a predictive setting, they fall short in a developmental setting. Training improvements to xFIP and SIERA are not exactly intuitive. However, breaking down these statistics into more granular metrics like stuff and location grades allows teams and coaches to directly train them. For instance, using Stuff+, teams can identify combinations of velocity, movement, and release points that are expected to perform well and improve them directly. Because the pitch grades are so related to the traditional metrics and early sabermetrics, improving the quality of a pitcher's individual pitches is likely to improve the traditional metrics and early sabermetrics as well.

Ultimately, the pitch grades did not add significant predictive accuracy to the advanced models. It is likely that the skills encapsulated by these metrics are already captured by existing metrics. However, there are still several use cases for them. Because pitch grades become accurate much quicker than the traditional metrics and early sabermetrics, future research may show that pitch grades do provide predictive power when making projections for pitchers with limited past playing time. Pitch grades also provide coaches and teams with a more intuitive way to improve players.

## REFERENCES

[1] (n.d.). FanGraphs Baseball | Baseball Statistics and Analysis. Retrieved March 23, 2024, from https://www.fangraphs.com

[2] (n.d.). Baseball Savant: Statcast, Trending MLB Players and Visualizations. Retrieved March 24, 2024, from https://baseballsavant.mlb.com/

[3] (n.d.). tidymodels. Retrieved March 25, 2024, from https://www.tidymodels.org/

[4] Benavidez, S. (2019). Prediction of Future Offensive Performance of MLB Position Players.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[6] Elitzur, R. (2020). Data analytics effects in major league baseball. *Omega*, *90*, 1-56. doi:10.1016/j.omega.2018.11.010

[7] Grove, C. (2023, March 10). *PitchingBot Pitch Modeling Primer | Sabermetrics Library*. Sabermetrics Library. Retrieved March 23, 2024, from https://library.fangraphs.com/pitching/pitchingbot-pitch-modeling-primer/

[8] McGrattan, O. (2023, March 10). *Stuff+, Location+, and Pitching+ Primer | Sabermetrics Library*. Sabermetrics Library. Retrieved March 23, 2024, from https://library.fangraphs.com/pitching/stuff-location-and-pitching-primer/

[9] Mizels, J., Erickson, B., & Chalmers, P. (2022). Current state of data and analytics research in baseball. *Current Reviews in Musculoskeletal Medicine, 15*(4), 283-290. doi: 10.1007/s12178-022-09763-6

[10] Oberoi, T., & Saarinen, S. (2024). Predicting baseball pitcher efficacy using physical pitch characteristics. *Journal of Emerging Investigators*.

[11] Richards, D. (2020, April 9). *The Relative Value of FIP, xFIP, SIERA, and xERA Pt. II*. Pitcher List. Retrieved March 23, 2024, from https://pitcherlist.com/the-relative-value-of-fip-xfip-siera-and-xera-pt-ii/

[12] Sharpe, S. (2019, September 20). *An Introduction to Expected Weighted On-Base Average (xwOBA)*. MLB Technology Blog. Retrieved March 24, 2024, from https://technology.mlblogs.com/an-introduction-to-expected-weighted-on-baseaverage-xwoba-29d6070ba52b

[13] Tango, T. (2012). *Marcel 2012*. Tango on Baseball. Retrieved April 2, 2024, from https://tangotiger.net/marcel/

[14] Weinberg, N. (2014, July 30). *RE24 | Sabermetrics Library*. FanGraphs Library. Retrieved March 24, 2024, from https://library.fangraphs.com/misc/re24/