Quantifying Consistency: Developing New Metrics for MLB Player Valuation

by

David Vapnek

B.S. in Business Analytics and in Computer Science, Economics and Data Science,
Massachusetts Institute of Technology, 2023

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by: David Vapnek
Department of Electrical Engineering and Computer Science
May 10, 2024

Certified by: Anette Hosoi
Professor of Mechanical Engineering and Mathematics, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Quantifying Consistency: Developing New Metrics for MLB Player Valuation

by

David Vapnek

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

ABSTRACT

This study investigates how performance consistency during a player's pre-arbitration years in Major League Baseball (MLB) influences their first-year arbitration salary, offering novel insights for player valuation. Consistency is quantified based on three categories: short-term volatility, in-season adaptability, and environmental consistency (home/away performance). Statistical models, including both OLS and Lasso regressions and Random Forests, show that consistency metrics hold statistically significant explanatory power even when controlling for traditional performance metrics, previous salary, and league conditions. The results also indicate that away performance holds significantly more weight than home performance when determining salary value. These findings suggest that while teams heavily consider known metrics, they also implicitly or explicitly recognize the potential of consistency as a signal of future success. This study contributes to the field by introducing quantifiable consistency measures to highlight a previously under-examined aspect of MLB player value.

Thesis supervisor: Anette Hosoi
Title: Professor of Mechanical Engineering and Mathematics

Acknowledgments

This thesis marks the conclusion of a demanding but rewarding journey, and its completion would not have been possible without the assistance and support of several key individuals. First and foremost, my sincere thanks go to my thesis advisor, Peko Hosoi. Her expertise and guidance were crucial in directing this research, and her positive energy made the process both productive and enjoyable. I also extend my appreciation to Christina Chase, whose insightful perspectives proved crucial in navigating roadblocks and refining my research goals.

I am incredibly grateful to my parents, Gillian and Evan, and my sister, Claire. Their unwavering love and support have been invaluable throughout my academic journey, and I would not be where I am today without them.

Finally, I'd like to thank the many friends I've made here at MIT. Specifically, the camaraderie and support of my fraternity brothers in Delta Tau Delta and my teammates on the MIT lacrosse team have made my time here truly memorable. To everyone who has been part of my MIT experience, I extend my deepest gratitude.

Contents

Title page 1

Abstract 3

Acknowledgments 5

List of Figures 9

List of Tables 11

1 Introduction 13

1.1 Scouting and the Data Revolution 13

1.2 Measuring Consistency 14

1.3 Impact 15

2 Related Work 17

3 Data Creation 19

3.1 Game Data 19

3.2 Salary Data 20

4 Methods 21

4.1 Quantifying Consistency 22

4.1.1 Short-Term Volatility 22

4.1.2 Environmental Consistency 24

4.1.3 In-Season Adaptability 26

4.2 Leveraging Arbitration 28
4.2.1 Isolating Early Years 30
4.2.2 Pooling 31
4.3 Models 32
4.3.1 Linear Regression 34
4.3.2 Lasso Regression 34
4.3.3 Random Forest 36

5 Results **37**
5.1 Linear Regression Results 37
5.2 Lasso Regression Results 40
5.3 Random Forest Results 40

6 Conclusion **45**
6.1 Impact 45
6.2 Future Work 46

A Software Packages **49**

References **51**

List of Figures

4.1 Correlation Heatmaps for Batting Statistics 25

4.2 Correlation Heatmaps for Home and Away Batting Performance 27

4.3 Distribution of Career Years for Reaching Key Arbitration Milestones 31

4.4 Comparison of Pooled and Single-Season Metrics 32

4.5 Correlation Heatmap for Highly Correlated Regressors 35

5.1 Random Forest Feature Importance Plots 43

List of Tables

4.1 Summary Statistics of Short-term Volatility in Slugging Percentage 24
4.2 Summary Statistics of Short-term Volatility with a Three Game Window . . 24
4.3 Summary Statistics of Home and Away Differences 26
4.4 Summary Statistics of First and Second Half Differences 29
4.5 Summary Statistics of Eighth Season Variances 29

5.1 OLS Regression Results 38
5.2 Lasso Regression Results Comparison 41

Chapter 1

Introduction

1.1 Scouting and the Data Revolution

In Major League Baseball (MLB), the difference between successful and unsuccessful teams relies largely on a team's ability to accurately assess the future value of a player. Teams work with limited, and often wildly variant payroll budgets. For example, in the 2022 season, the gap between the largest and smallest payroll was approximately $226 million [1]. This means that teams, especially with smaller budgets, must accurately predict a player's future performance to effectively value them and thus optimize their spending. For decades, organizations relied exclusively on teams of scouts to watch players and evaluate them on five key criteria–none of which include statistics. Rather, scouts were trained to score players on their skills and technique and report back to organizations with a comprehensive report on how they expected those skills to translate into future MLB success [2].

In the early 2000s, there was a monumental shift in the evaluation of MLB players when Billy Beane, the manager of the Oakland Athletics, introduced Moneyball theory. With this theory, the qualitative assessment of scouts was no longer the principal method of evaluation; instead, statistics were the primary source of information. Beane found there were many players with high slugging percentages (SLG) and high on-base percentages (OBP) who were not highly regarded by scouts, but whose performance was clearly strong. With this method, he was able to find undervalued players and create a successful team with a low payroll [2]. After the rest of the league realized their previous scouting methods were inefficient, the

importance of data and analytics in the MLB skyrocketed. Players were valued on their ability to get on base and produce runs, as measured by various traditional statistics.

In tandem with this, the popularity of Sabermetrics also surged. Sabermetrics is a collection of research and advanced statistics from the Society of American Baseball Research [3]. More recently, the introduction of Statcast, a system that quantifies massive amounts of baseball data, including the speed and spin of every batted and thrown ball, has allowed organizations and Sabermetrics researchers alike to quantify a player's performance in nearly every aspect of baseball. Stats like wins above replacement (WAR) and weighted runs created plus (wRC+) can directly and accurately relate a player's performance over a period of time to their value as a player in helping their team win [4].

1.2 Measuring Consistency

Even as baseball statistics and evaluation methods evolve, one thing remains constant: players are evaluated on their average aggregate performance over a fixed period of time. The consistency of a player's output is not well accounted for, even in the most advanced statistics. Though consistency may not affect a player's overall production, it can provide insight into how a player adapts to different circumstances over time. Many players who are drafted in the first round never make it to the major leagues [5], and even those who do can be signed to multi-million dollar contracts in their early 20s before failing to meet expectations as they progress in their careers [6]. Despite the incredible depth of data available, it remains extremely difficult to predict the future performance of young players.

Quantifying consistency adds a new dimension to statistical analysis in baseball. By understanding how production comes about, rather than just the production itself, it may be possible to better predict a player's performance trajectory in the future. Consistency can be especially important as players progress through different levels of baseball, from high school or college, to the minor leagues, to the MLB. At each one of these jumps, players must adapt to new rules, new stadiums, and new levels of skill and athleticism. Analyzing past performance is not sufficient when predicting a player's response to these progressions; instead, understanding how a player has adapted in the past can inform how they will

continue to adapt in the future.

1.3 Impact

Understanding the impact of consistency is not only important for MLB franchises, it is important for the league itself. The MLB is a business that survives by producing an entertaining product: the baseball game. When fans invest their time and money to watch their favorite teams play, they want to see an exciting game with lots of action and scoring. However, the MLB has seen a decline in the proportion of balls in play, meaning plate appearances that end in something other than a walk, home run, or strikeout [7]. Because the entertainment value of a game is tied to the amount of balls in play, the MLB is specifically concerned with increasing this number to keep fans engaged. Recent changes, including banning the shift, a defensive strategy, and introducing a pitch clock, a timer limiting the delay between pitches, have shortened games and increased excitement. This progress is generally seen as a positive for the sport, but the balls in play rate is still lower than optimal. As a result of my research, I will understand if certain high-variance traits are incentivized by teams, resulting in a less interesting game with fewer balls in play.

In the current state of baseball, end results drive key decisions, whether it be the decision to draft a certain player, sign a young star to a long-term deal, or adjust the rules of the game. While this is not inherently a problem, the current statistics only measure a result at the end of a game, series, season, or some other fixed period of time. No matter how advanced stats become, this will always be a limiting factor. This paper introduces a novel approach, whereby instead of seeking to find the newest, more complicated statistic, traditional statistics are investigated with a different perspective.

By quantifying how a player's statistics vary over the course of a fixed window, it is possible to understand how much consistency and adaptability matter for a player, and if certain high-variance behaviors are being incentivized. Pertaining to the team's perspective, this paper seeks to understand if a player's consistency over the beginning of their MLB career can help predict future success. Relating to the league's perspective, this paper investigates if the same statistics being used to evaluate players are incentivizing high-variance plays

which hurt the overall value of the league. In the following chapters, I will first describe my data, then I will explain how I calculate my consistency metrics. Next, I will describe the setup and results of my models, followed by a discussion of their impact.

Chapter 2

Related Work

While there is virtually no work directly covering the topic of consistency in the MLB, there is a large body of work outlining forecasting models for player success, as well as psychological research supporting the idea that consistency and adaptability may influence future success.

The challenge of predicting future success in the MLB is well-documented, specifically with results from the MLB draft showing that historically, only 63% of first-round draft picks even reach the majors [5]. Even when only looking at data from the past decade, only 46.8% of first-round draft picks play in the MLB for three or more seasons [8], showing that it is extremely difficult to discern which players will adapt well to the major league environment. Even with this challenge, evidence shows that the predictive power of a player's minor league statistics increases in tandem with their progression through the minor leagues, and though the value of the data is limited, there is signal in the data among the best players in the minor leagues [9]. Further studies of minor league baseball suggest that the same offensive statistics that are valued in Moneyball theory in the MLB [10] are the most predictive within the minor leagues [11]. This work suggests that there is predictive power in statistics at the point at which players enter the MLB.

More recently, significant emphasis has been placed on using new statistics and methods to predict offensive production among MLB players. Research using machine learning techniques [12], [13] has shown slight improvement in forecasting, while research utilizing newly available Statcast data [14] has also shown slight improvements by isolating factors outside of a player's control. Statcast data has also opened up research into how baseball as a sport

is changing as a whole. Ball tracking data from Statcast shows that pitchers can throw pitches with more spin and speed than ever before, which has led to a consistent decrease in the balls in play rate over time. With this data comes a set of exploratory studies into what can be done to change this trend, including physical changes to the field of play or the ball [7], [15].

There is some work investigating consistency, including an article showing that more consistency on offense led to better wins. Yet in this article, as in many others, consistency is only defined as a standard deviation or variance [16]. Consistency has largely been overlooked, however, especially as new data becomes increasingly available. With the rise of Statcast and advanced metrics, researchers have been able to make marginal improvements to existing forecasting models, yet they all suffer from the same constraint of looking at the cumulative output of a player. Additionally, the league as a whole has a growing interest in improving the quality of its product, aided by this influx of new data. This study seeks to add to the existing research by investigating the consistency in players' hitting statistics as both a powerful forecasting tool and a metric to help determine the cause of concerning trends across the league.

Chapter 3

Data Creation

3.1 Game Data

The primary dataset used for this research comes from the Chadwick Baseball Bureau [17] using raw data from Retrosheet [18]. Retrosheet is a comprehensive database containing event data from every pitch of every game for the past several decades and is regarded as the most complete database of MLB statistics. The Chadwick Baseball Bureau disaggregates Retrosheet data at various levels, reformulating it into more usable datasets. The scale and depth at which baseball statistics are recorded make this study possible, and the specific datasets used for this project include individual batter splits for all major league games from 1995 - 2019. I chose to focus only on batters for this project because offensive production is more quantifiable than defensive performance, especially in the pre-statcast era. The dataset I use has 13,878 individual player seasons across 2839 individual players. Starting with the game-by-game data, I first calculated four key offensive metrics. These metrics include batting average (AVG), on base percentage (OBP), slugging percentage (SLG), and on base + slugging (OPS). These four statistics are considered the main traditional measures of offensive production in baseball. Each statistic is defined as follows:

$$AVG = \frac{\text{hits}}{\text{at bats}} \tag{3.1}$$

$$OBP = \frac{\text{hits} + \text{walks} + \text{hit-by-pitches}}{\text{plate appearances}} \tag{3.2}$$

$$SLG = \frac{1 * 1\text{B} + 2 * 2\text{B} + 3 * 3\text{B} + 4 * \text{HR}}{\text{at bats}} \tag{3.3}$$

$$OPS = OBP + SLG \tag{3.4}$$

3.2 Salary Data

To perform my analysis, I also needed salary data for all players in the dataset. As salary data is sparse, especially data from nearly 30 years ago, I used two different salary databases for my research. I primarily used Cot's Contracts, a salary database which is part of Baseball Prospectus [19]. Cot's Contracts has the length and value of most contracts in the MLB going back decades. To fill in some of the missing values, I supplemented the Cot's Contracts data with data from the Lahman Baseball Database [20], a separate salary database with some additional contract sources. By using both sources, I was able to gather salary data for 84.6 % of player seasons across my 25 years of collected data.

Chapter 4

Methods

This study aims to investigate the economic value of consistency in MLB player performance. In order to perform this analysis, I first aim to quantify consistency. First, I develop a comprehensive definition of consistency, encompassing not only the traditional understanding of consistency as the variability in a player's performance but also factors such as a player's ability to adapt to new environments and maintain their level of play throughout an entire season. After obtaining a set of metrics that quantify a player's consistency, the next step is to find their value in the MLB contract market. The analysis leverages the MLB's unique salary structure. By focusing on a player's pre-arbitration years, where salaries are determined by league-wide standards, I can isolate the impact of consistency on subsequent arbitration salaries. This allows for a direct estimation of the market value placed on player consistency. I use regression models, incorporating appropriate controls, to quantify this relationship and determine the extent to which consistency influences player compensation.

An important caveat in this analysis is that all consistency metrics and regressions are based entirely on offensive performance. This means defensive consistency and pitching consistency are not evaluated, so the study focuses on position players exclusively. While both defensive and pitching consistency are important in baseball, this analysis focuses on offensive performance for several reasons. Firstly, position players are primarily valued for their offensive contributions, and salary is strongly correlated with offensive production. This follows from the idea that fielding skills are expected for all position players, whereas batting skills are more sought after. Another reason for this decision is that offensive statistics are

more readily available than defensive statistics from the pre-statcast era, giving a more robust dataset for offensive consistency calculations. Although defensive skills are vital for position players, offensive performance remains a major differentiating factor, making the focus on offense suitable for understanding the value of consistency among position players.

4.1 Quantifying Consistency

To quantify consistency, I use game-level hitting data from Retrosheet to derive a multi-faceted set of consistency metrics for MLB position players. The first consistency metric I define is short-term volatility, which is the variance in the weighted average of a statistic across a certain number of a player's most recent games. Beyond this intuitive definition of consistency, two novel dimensions are introduced. Firstly, environmental consistency quantifies a player's ability to perform consistently across different stadiums and conditions. Secondly, in-season adaptability measures how well a player maintains their performance throughout the grueling 162-game season. This comprehensive approach allows for the analysis of the economic value of each distinct type of consistency. I measure each type of consistency using the key offensive metrics from modified Retrosheet data, including batting average, on-base percentage, slugging percentage, on-base plus slugging, hits, home runs and runs batted in (RBIs).

4.1.1 Short-Term Volatility

To quantify short-term fluctuations in player performance, this study introduces a metric called short-term volatility. Short-term volatility captures the degree of consistency in a player's offensive statistics by calculating the variance in rolling weighted averages over a predefined window of games. For a given game, the rolling weighted average includes statistics from that game, along with those from the player's most recent games up to the chosen window size. The general formula for calculating a rolling weighted average is below, followed by a specific example formula for calculating the rolling weighted average for slugging percentage across a three game window. Note that the number of attempts used in the weighting will be plate appearances for most statistics and at-bats for batting average

and slugging percentage.

$$\textit{rolling weighted average}_i = \frac{\sum_{j=i-(window-1)}^{i} stat_j * attempts_j}{\sum_{j=i-(window-1)}^{i} attempts_j} \tag{4.1}$$

Example slugging average over a 3-game window:

$$\textit{rolling weighted SLG average}_i = \frac{\sum_{j=i-2}^{i} SLG_j * AB_j}{\sum_{j=i-2}^{i} AB_j} \tag{4.2}$$

For each player, we calculate rolling weighted averages across different-sized game windows. Averages are only included if a player actively participates in a game, resulting in a maximum of $162 - (\text{window} - 1)$ data points per player. A player must participate in at least the window size number of games before the rolling average calculation begins. We then derive short-term volatility for player j by calculating the sample variance as follows:

$$\textit{short-term volatility}_j = \frac{\sum_{i=1}^{n}(rwa_i - \overline{rwa})^2}{n-1} \tag{4.3}$$

Here, rwa is the rolling weighted average for any statistic, and n is the total number of rolling weighted averages calculated for a given player. Experimentation revealed that a smaller window size of three games is optimal for this analysis. Larger windows of 11 or more games caused the rolling averages to converge towards the season average and reduced any variation between data points across a season. Therefore, I selected a three-game moving average as the primary window for the short-term volatility metric. This window aligns with the structure of an MLB season, which is typically broken into three or four-game series. While the rolling metric may include data from multiple series, it effectively captures player performance fluctuations throughout smaller segments of the long season.

After choosing a three-game window, I compared the variability across four key offensive statistics. As expected, OPS and SLG exhibited the largest fluctuations, with much of the variability in OPS attributable to SLG. This aligns with the nature of SLG, which weighs the

Table 4.1: Summary Statistics of Short-term Volatility in Slugging Percentage

window size	count	mean	std	min	25%	50%	75%	max
3	7944	0.07842	0.03404	0.01078	0.05502	0.07258	0.0947	0.3995
7	7944	0.03148	0.01517	0.00334	0.02085	0.02874	0.03896	0.19632
11	7944	0.01899	0.01024	0.00114	0.01175	0.017	0.02383	0.09801
15	7944	0.01316	0.00791	0.00076	0.00761	0.01143	0.01683	0.07519
19	7944	0.00973	0.00643	0.0004	0.00518	0.00824	0.01264	0.05511

Table 4.1: Summary statistics of short-term volatility in slugging percentage across a different window sizes. Window sizes are all in terms of number of games played. The sample includes one data point for each eligible individual player season from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

Table 4.2: Summary Statistics of Short-term Volatility with a Three Game Window

statistic	count	mean	std	min	25%	50%	75%	max
avg	7944	0.02018	0.00622	0.00667	0.0163	0.01901	0.02247	0.07743
slg	7944	0.07842	0.03404	0.01078	0.05502	0.07258	0.0947	0.3995
obp	7944	0.02205	0.00742	0.00604	0.01749	0.02043	0.0245	0.08784
ops	7944	0.15882	0.05718	0.04214	0.12063	0.14841	0.18484	0.7346

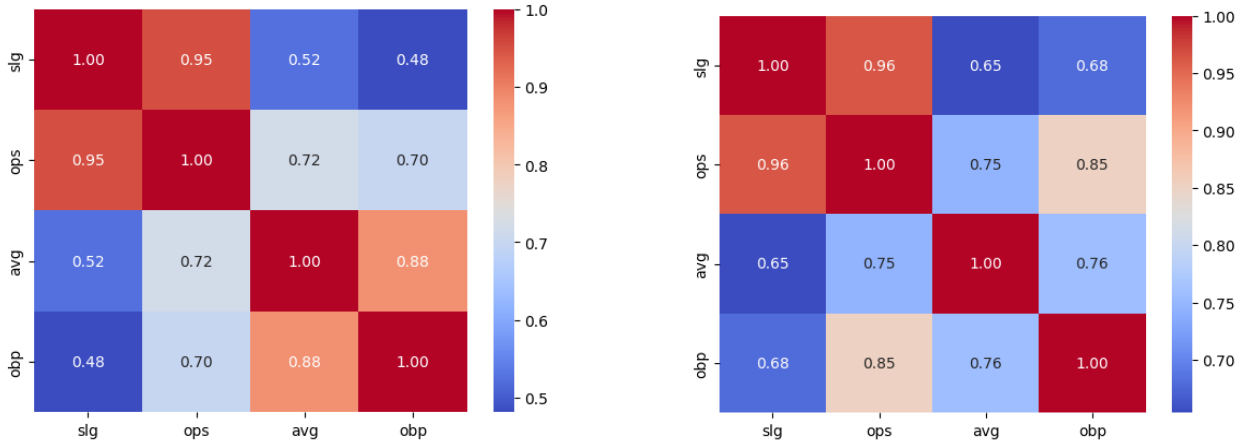
Table 4.2: Summary statistics of short-term volatility across a three game window. Summary statistics are of the variance, not the statistic itself. The sample includes one data point for each eligible individual player season from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

impact of each hit, leading to inherently greater sensitivity to short-term performance swings. Due to its sensitivity, I selected SLG as the primary statistic for the short-term volatility metric. While all four statistics demonstrate strong correlation as shown in Figure 4.1, I explore the impact of each one throughout the analysis to ensure robust findings.

4.1.2 Environmental Consistency

I introduce the concept of environmental consistency, which quantifies how well a player maintains their performance across different stadiums. An MLB season consists of an even split between home and away games, so a player who plays an entire season will play over

Figure 4.1: Correlation Heatmaps for Batting Statistics



(a) Short-term Volatility Metric

(b) Raw Statistics

Figure 4.1: Correlation heatmaps for batting statistics. Figure A is the correlation between the short-term variability metrics of each statistic measured at 3-game windows. Figure B is the correlation between season averages of raw batting statistics.The sample includes all eligible player seasons from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

80 games at their home venue, and between three and seven games at each of more than 20 different away venues. Adding to the difficulty, each stadium in the MLB is different, as no stadiums have the same sightlines, atmosphere, or even field dimensions. Because of this, players can be much more comfortable playing at home than away. A player's ability to adapt to new environments on the road and match, or even exceed, their home performance on the road can inform how a player adapts to new environments.

To calculate environmental consistency for a given statistic, I subtract away performance from home performance. The general formula is below, as well as a specific example for SLG, where *ha diff* is the difference in home and away stats, and $home_i$ is an indicator for if game i was a home game.

$$stat\ ha\ diff = \frac{\sum_{i=1}^{162} stat_i * attempts_i * home_i}{\sum_{i=1}^{162} attempts_i * home_i} - \frac{\sum_{i=1}^{162} stat_i * attempts_i * (1 - home_i)}{\sum_{i=1}^{162} attempts_i * (1 - home_i)} \quad (4.4)$$

$$SLG\ ha\ diff = \frac{\sum_{i=1}^{162} SLG_i * AB_i * home_i}{\sum_{i=1}^{162} AB_i * home_i} - \frac{\sum_{i=1}^{162} SLG_i * AB_i * (1 - home_i)}{\sum_{i=1}^{162} AB_i * (1 - home_i)} \quad (4.5)$$

Table 4.3: Summary Statistics of Home and Away Differences

	count	mean	std	min	25%	50%	75%	max
ha avg diff	7925	0.0012	0.0755	-0.3653	-0.0432	-0.0001	0.0438	0.7315
ha obp diff	7926	0.0006	0.0667	-0.3567	-0.0375	0.0002	0.0375	0.6604
ha slg diff	7925	0.0023	0.1476	-0.6426	-0.0844	-0.0039	0.0817	1.643
ha ops diff	7926	0.0019	0.1815	-0.85	-0.105	-0.0015	0.1045	1.8924
ha hits diff	7944	-39.7709	40.0786	-198	-63	-29	-9	38
ha hr diff	7944	-4.7932	7.0267	-61	-8	-3	0	22
ha rbi diff	7944	-19.8914	22.7703	-138	-31	-14	-4	32

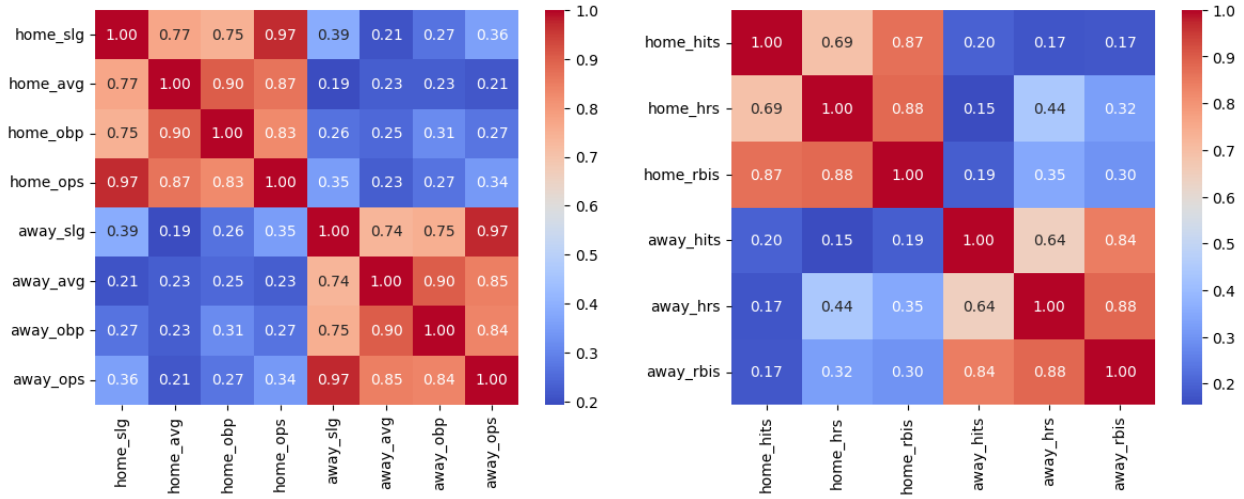
Table 4.3: Summary statistics of home and away differences. Values are calculated by subtracting the away values of a statistic from the home values.The sample includes one data point for each eligible individual player season from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

I also calculate home and away performance differences for hits, home runs, and RBIs, which will all be included in the final models. For all home and away differences, a positive value indicates that a player performs better in that statistic at home, while a negative value signifies better away performance. A value of zero indicates perfect environmental consistency. The distribution of the home and away differences are shown in Table 4.3. As part of creating environmental consistency, it is important to understand how home and away performance are related. Surprisingly, Figure 4.2 reveals that home statistics are minimally correlated with away statistics. There is a very strong correlation between a player's batting statistics within an environment, but the correlation across environments for any given statistic is generally quite weak. This relationship underscores the potential value of understanding a player's ability to adapt to diverse playing environments.

4.1.3 In-Season Adaptability

Finally, to measure long-term consistency within a single season, I introduce the concept of in-season adaptability. This metric goes beyond short-term volatility, focusing instead on how a player's performance evolves throughout the six-month, 162-game MLB season.

Figure 4.2: Correlation Heatmaps for Home and Away Batting Performance



(a) Batting Statistics

(b) Batting Events

Figure 4.2: Correlation heatmaps for home and away batting statistics and events. Figure A is the correlation between the home and away values for common batting statistics. Figure B is the correlation between key home and away batting outcomes. The home or away prefix indicates whetehr stats are taken from a player's home or away games. HRs is home runs and RBIs is runs batted in. The sample includes all eligible player seasons from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

By analyzing if a player maintains their initial performance, experiences improvement, or faces declines, we gain insights into their consistency over time. The in-season adaptability metric quantifies consistency when viewing consistency as the ability to adapt and overcome challenges in pursuit of long-term goals. Players who demonstrate superior adaptability to the physical and mental demands of the long season may possess the resilience necessary for long-term success.

To calculate in-season adaptability, I first segment the MLB season into chronological groups. This enables the creation of metrics that track player performance trends. My initial approach focuses on calculating the difference in key statistics between a player's first and second half of the season by tracking the magnitude of the difference and creating an indicator of improvement or decline. Interestingly, across all four statistics analyzed, 47% of players demonstrated improvement in the second half of the season. More detailed information can be seen in Table 4.5. With improvement over a season accounted for, I can now calculate consistency over the course of a season by segmenting it into eight groups of 20 or 21 games each. This timeframe balances providing significant segments of a season while maintaining sufficient data points for comparison. For each player, I then identify their best and worst performing segments and calculate the difference. This metric captures the degree of variation between a player's peak and low points within a season. Additionally, I calculate the variance across all eight segments. While this is similar in concept to short-term volatility, the approach differs in that this metric uses predefined blocks of a season, and only tracks the variance in a player's overall performance trajectory throughout the season, rather than how much their performance changes on a game-by-game basis. To ensure accuracy, all statistics within each time group are weighted by the respective number of games played in that period. By comparing performance across different time blocks in a season, I am able to quantify each player's in-season adaptability and use it in my later models.

4.2 Leveraging Arbitration

This study focuses on analyzing players early in their MLB careers. Since young players are still developing as athletes, understanding consistency holds particular value in evaluating

Table 4.4: Summary Statistics of First and Second Half Differences

	count	mean	std	min	25%	50%	75%	max
half_diff_avg	7944	0.0005	0.0602	-0.792	-0.0304	0	0.0328	0.7802
half_diff_obp	7944	0.001	0.062	-0.7301	-0.0317	0	0.0351	0.6824
half_diff_slg	7944	0.0018	0.1144	-1.6279	-0.0587	0	0.0651	1.5631
half_diff_ops	7944	0.0029	0.1675	-2.3435	-0.0858	0.0007	0.0957	1.8891

Table 4.4: Summary statistics of first and second differences. Values are calculated by subtracting the second half values of a statistic from the first half values. Halves of a season are defined as the first 81 and last 81 games of the regular season. The sample includes all eligible player seasons from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

Table 4.5: Summary Statistics of Eighth Season Variances

	count	mean	std	min	25%	50%	75%	max
avg var	7944	0.0066	0.0098	0	0.0024	0.004	0.0073	0.1388
obp var	7944	0.0072	0.0099	0	0.0026	0.0043	0.0078	0.1676
slg var	7944	0.0238	0.0456	0.0001	0.0088	0.0154	0.0265	1.8737
ops var	7944	0.0503	0.0769	0.0005	0.0192	0.0324	0.0574	2.7345

Table 4.5: Summary statistics of eighth season variances. Season eighths are determined chronologically using groups of 20 or 21 games. The sample includes all eligible player seasons from 1995-2019. An eligible season is defined as one where a player makes a plate appearance in at least 40 games.

the potential of this group of players. Rules for an MLB player's salary are determined by their service time, which is defined as days spent on an active roster. For a player's first three years, they must be paid the league minimum salary, which is set by the collective bargaining agreement signed by the MLB and the players association. There is some variation due to signing bonuses based on draft position, but all players earn a comparable amount.

Once a player accumulates more than three and less than six years of service time, they enter arbitration. In arbitration, players and teams submit proposed salary figures to an independent arbitration panel. After hearing arguments from both sides, the panel selects either the player's or the team's figure as the binding salary for the upcoming season. Arbitration decisions often consider factors such as the player's past performance, comparable player salaries, and league-wide salary trends. This system provides a unique opportunity to analyze players under a standardized pay structure before their market value is fully established. By aggregating a player's performance over their first three years of service time, I can understand the effect of early career consistency on the market value of a player when they reach arbitration.

4.2.1 Isolating Early Years

Determining a player's arbitration eligibility is more complicated than finding their third season in MLB. It is based on service time, which accrues daily while a player is on an active roster, meaning when a player is injured or spends time in the minor leagues, they do not accumulate service time. Since my dataset lacks daily roster status, I estimate service time using the games played metric. For each player, I count the number of games they appear in, and unlike the other statistics calculations, I also include games where a player does not record a plate appearance. From here, I scale the number of games into a service time estimate using the number of days and games in the MLB regular season.

With my service time estimates, I can track how much service time a player accrues throughout each season of their career. The MLB defines one year of service time as 172 days spent on an active roster, meaning a player generally reaches arbitration after 516 days, or three years, of service time. However, there is an exception for certain players called the Super Two designation. This designation applies to any player who ranks in the top 22

Figure 4.3: Distribution of Career Years for Reaching Key Arbitration Milestones

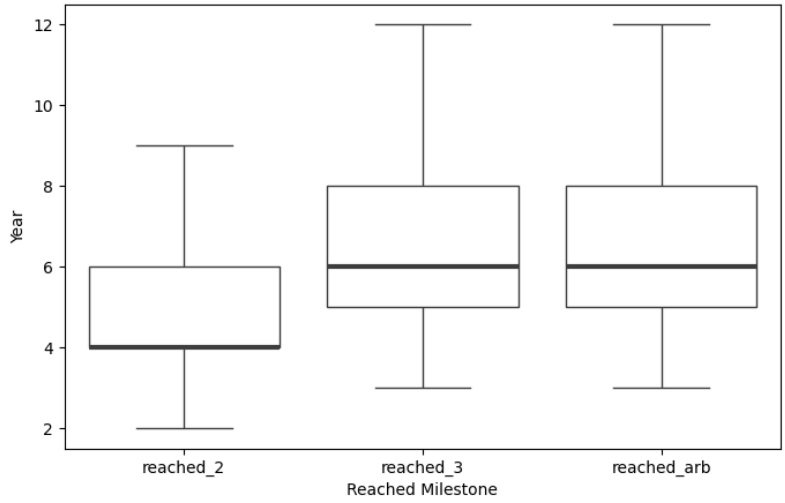


Figure 4.3: Distribution of Career Years for Reaching Key Arbitration Milestones. Box and whisker plots where box spans IQR and thick black line is the median. Whiskers extend to $1.5 * IQR$ in each direction.

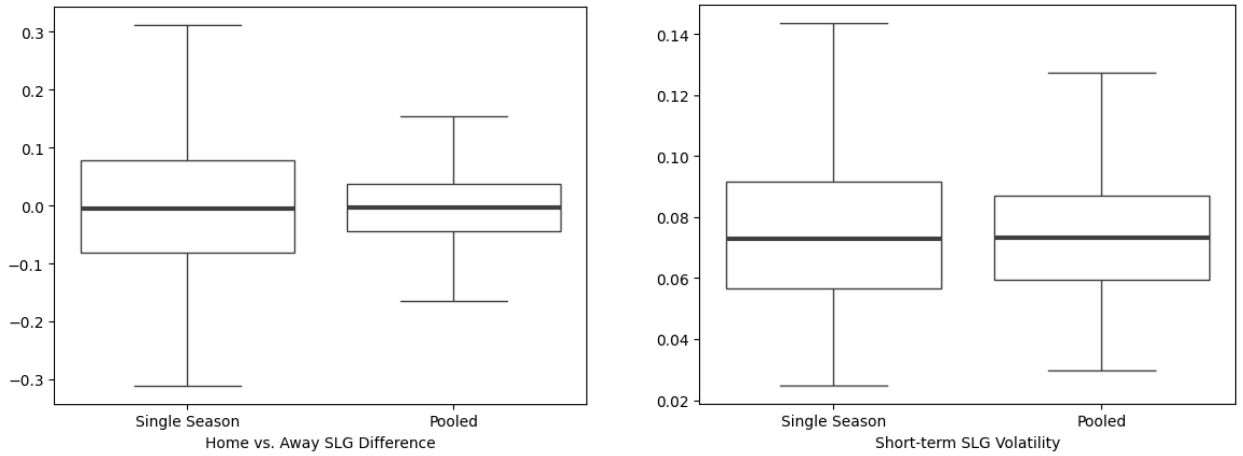
percent of service time among players who have between two and three years of service time and allows them to enter arbitration before reaching three years of service time. Because this threshold changes annually, I manually calculate it based on the definition above and mark players who qualify for the Super Two designation. With the three-year rule and the Super Two designation accounted for, I can find the year in which every player accumulates enough service time to become arbitration eligible for the following year. It should be noted that many players never accrue enough service time to become arbitration eligible. In my dataset, 20.33% of players never reached arbitration and were therefore excluded from the model.

4.2.2 Pooling

After identifying the year where each player becomes eligible for arbitration, I pool their metrics from preceding seasons for a comprehensive view of pre-arbitration performance and consistency. While pooling does increase the sample size for my regression model in this case, as each pooled metric still represents one data point in the model, it still offers several advantages. Pooling provides a more representative view of a player's true pre-arbitration

metrics by accounting for anomaly seasons that may have been influenced by injury, luck, or other unseen factors. Pooling mitigates these fluctuations. Additionally, because arbitration is set at a specified point in a player's career, pooling metrics allows for a more even sample size of data from each player. Pooled metrics are created by taking the weighted average of a given metric across all pre-arbitration years, weighted by the number of plate appearances in a given year.

Figure 4.4: Comparison of Pooled and Single-Season Metrics



(a) Home and Away SLG Difference

(b) Short-term SLG Volatility

Figure 4.4: Comparison of Pooled and Single-Season Metrics. Box and whisker plots where box spans IQR and thick black line is the median. Whiskers extend to $1.5 * IQR$ in each direction.

4.3 Models

Having calculated consistency metrics, determined arbitration eligibility, and pooled pre-arbitration statistics, I construct a sequence of models to investigate the relationship between players' early-career consistency and their initial arbitration salaries. All models employ a player's first-year arbitration salary as the dependent variable. Across all models, I use slugging as the statistic for which I measure performance and consistency. I find slugging percentage to be the strongest statistic as it is a more nuanced version of batting average, allowing for a better representation of offensive output. Regardless, experiments showed that the choice of statistic did not have a large impact on the outcome of the model. Additionally,

I use the pooled statistics for all models, rather than single-season statistics. Because my independent variables had scales that varied by multiple orders of magnitude, I normalized all coefficients to a standard normal distribution. This helps to address multicollinearity in the regression models and ensures that no variables dominate the model due to their magnitude.

The primary independent variables are a player's consistency metrics, alongside their raw performance as measured by home and away splits. Splitting performance metrics across environments allows me to evaluate if performance in one environment is more valuable than another. Additionally, I include a player's final pre-arbitration salary, which allows me to control for any fluctuations within the standardized pay structure due to bonuses and time period. In an effort to mitigate omitted variable bias, I include a series of other controls. To account for the time it takes a player to reach arbitration, I control for career year and total pre-arbitration plate appearances. I also include a control for the Collective Bargaining Agreement (CBA) period to account for changing market conditions over time. Finally, I control for whether a player changed teams in their first year of arbitration to account for payroll disparities across teams. I begin with ordinary least squares (OLS) regression as a baseline. Subsequently, I explore regression models designed to mitigate multicollinearity. Finally, I employ tree-based classifiers to investigate potential non-linear relationships within the data.

Across all models, the sample is restricted to players who entered the league in 1995 or later. Even though there is batting and salary data in these seasons, I am unable to accurately pool pre-arbitration players who began their careers before 1995, so I drop 23.68% of observations. Additionally, I can only include players for which I have salary data, which is 98.7% of my original dataset. Finally, I restrict the data to players who reached arbitration within 8 years of their debut, which is the third quartile. Players above this threshold took an extended time to accrue three years of service time, which diminishes the impact of their consistency as young players. This restriction drops 16.88% of the original data. Some of the players dropped by these restrictions overlap, so after implementing all of my restrictions, I dropped 28.35% of my observations, leaving me with 743 player seasons to analyze.

4.3.1 Linear Regression

As the primary goal of this study is to find a concrete value of consistency, I first use an OLS regression to estimate the dollar value of each type of consistency. Because all variables, including the outcome variable, were standardized, the coefficients can be interpreted in terms of changes in standard deviations. This is especially helpful for understanding the consistency metrics, as the raw values do not hold much meaning by themselves. By using the standard deviation of salaries, it is trivial to convert the standard deviation change in salary to a specific number. I introduce four OLS models to estimate the effects of consistency on performance. Models I and III exclude raw performance while Models II and IV include it. Models III and IV include the set of controls mentioned previously while Models I and II do not. Across all four models, I cluster standard errors by team to account for potential heterogeneity in salary structures across MLB teams. This acknowledges that teams may have different internal salary allocation processes or budget constraints, potentially leading to correlations in salaries offered to players within the same team that are not captured by our model's independent variables.

Examination of the OLS models reveals potential multicollinearity among the variables. This is expected due to the inherent correlations within batting statistics. As visualized in Figure 4.5, several regressors exhibit moderate correlations, with the strongest reaching 0.83. To account for this, I further investigate alternative modeling approaches to assess the impact of multicollinearity on the interpretation of my OLS results.

4.3.2 Lasso Regression

In addition to OLS regression, I also use Lasso regression to address potential multicollinearity among the explanatory variables. Lasso regression incorporates a regularization term that shrinks the magnitude of coefficients. I use Lasso to help reduce the risk of overfitting and identify the most important features in the data, given the data has a number of correlated features. The setup for the Lasso model is the same as that of the OLS model, without the clustering of standard errors, as standard errors are not calculated in the same manner for

Figure 4.5: Correlation Heatmap for Highly Correlated Regressors

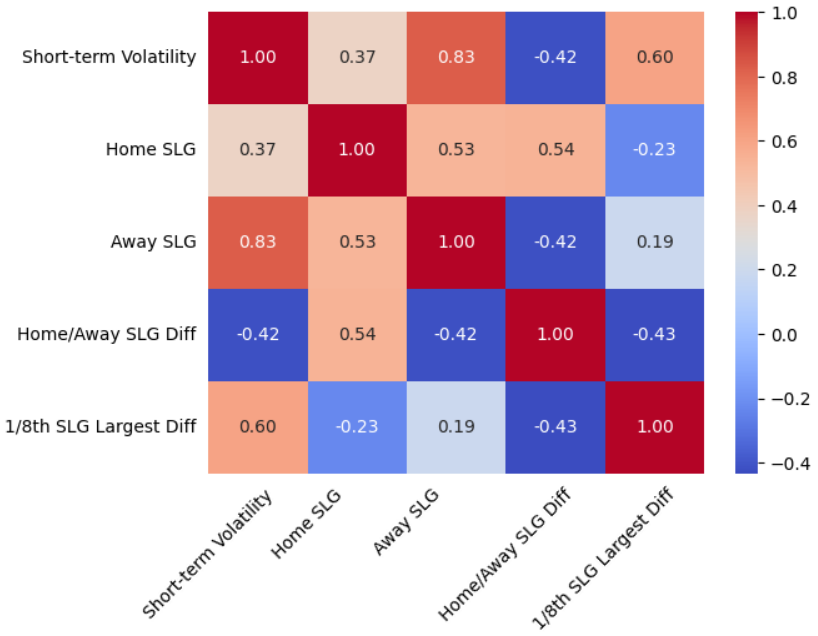


Figure 4.5: Correlation heatmaps for highly correlated regressors. Regressors are only included if they correlate with another regressor at a magnitude of 0.5 or greater. Consistency metrics are all for SLG.

Lasso regression. The equation for the Lasso model is as follows:

$$LASSO(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta}) + \alpha\sum_{j=1}^{m}|\hat{\beta}_j| \tag{4.6}$$

Here, y_i is the outcome variable, or post-arbitration salary, for player i. The estimated salary for player i comes from their explanatory variables x_i^T multiplied by the coefficients $\hat{\beta}$. The second term in this equation is the regularization term, making Lasso unique from OLS. The magnitude of all m coefficients is multiplied by a penalty weight α and added to the error term. Solving the Lasso regression for the set of coefficients involves finding the minimum of $LASSO(\hat{\beta})$. For this model, I used a grid search to obtain an optimal penalty weight of $\alpha = 0.1$.

4.3.3 Random Forest

To explore the potentially complex and non-linear relationships between player consistency and arbitration salaries, I employ a Random Forest model. This technique constructs many decision trees, each trained on a random subset of data and features, and the model's prediction is based on an aggregation of the individual tree outputs. The Random Forest model will not only account for any complex relationships in my data but also help mitigate the effect of any outliers or multicollinearity. Additionally, the model provides feature importance measures, offering insights into which explanatory variables hold the most predictive power in my model.

Chapter 5

Results

Having defined all the models, I now present the results to explore the connection between player consistency and arbitration salaries. I begin by examining the key findings from my main OLS regression models, focusing on the coefficients and their significance. Next, I will compare these results with those obtained from a Lasso regression, specifically exploring how regularization addresses multicollinearity and potentially modifies the identified key predictors. Finally, I will analyze my Random Forest model, comparing its performance to the regression models and investigating its feature importance measures. Exploring a variety of models will reveal which consistency metrics have the largest impact on player salaries.

5.1 Linear Regression Results

First, I present my OLS regression results in Table 5.1. To account for the different scales of explanatory variables, all regressors were standardized. Additionally, I scaled arbitration salaries to hundreds of thousands of dollars for ease of interpretation. I examine four model variations: Models I and III focus solely on consistency metrics, while Models II and IV incorporate performance metrics. Models III and IV progressively include the added set of controls, culminating in Model IV, which yields the highest adjusted R^2 of 0.69. Since Model IV offers the strongest protection against omitted variable bias, it serves as the primary model for coefficient interpretations.

Looking at the consistency metrics in Model IV, three of the four metrics are statistically

Table 5.1: OLS Regression Results

	Model I	Model II	Model III	Model IV
Intercept	34.570***	34.570***	30.794***	30.602***
	(0.569)	(0.515)	(3.063)	(3.210)
Salary	14.984***	13.926***	13.136***	12.617***
	(1.484)	(1.446)	(1.562)	(1.577)
Short-term Volatility				
Short-term SLG Volatility	5.418***	-3.023***	3.672***	-1.637**
	(0.737)	(0.879)	(0.512)	(0.758)
In Season Adaptability				
First/Second Half SLG Difference	-1.860***	-0.875*	-1.205**	-0.795*
	(0.587)	(0.512)	(0.479)	(0.472)
1/8th Season Largest SLG Difference	-4.980***	-2.195***	-0.346	-0.400
	(0.767)	(0.636)	(0.625)	(0.598)
Environmental Consistency				
Home/Away SLG Difference	-0.482	-1.687***	-0.407	-1.166***
	(0.406)	(0.339)	(0.390)	(0.311)
Home SLG		4.090***		2.991***
		(0.337)		(0.370)
Away SLG		6.928***		4.997***
		(0.628)		(0.644)
Controls Added	No	No	Yes	Yes
R-squared	0.523	0.609	0.671	0.695
R-squared Adj.	0.519	0.606	0.665	0.690

Table 5.1: OLS Regression Results. Regressors are normalized and predict arbitration salary in $100,000s. All regressors are pooled from a player's pre-arbitration seasons and use slugging as the statistic of interest. The sample includes all eligible player seasons from 1995-2019, with eligible seasons defined in methods. Added controls include pooled plate appearances, Collective Bargaining Period, career year, and a changed teams indicator. Standard errors are clustered by team.
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

significant. Home/Away Slugging Difference, First/Second Half Slugging Difference, and Short-term Slugging Volatility are significant at 1%, 5%, and 10% levels respectively. Starting with environmental consistency, the coefficient on home and away difference is negative, meaning that improved relative away performance is associated with higher salaries. Specifically, this model shows that a one standard deviation increase in relative away performance across a player's pre-arbitration years is associated with a $116,600 increase in arbitration salary. Though not directly a consistency metric, another interesting result from environmental performance is the difference in the weights of home and away slugging performance. A one standard deviation increase in home slugging is associated with an extra $299,1000, whereas the same increase in away slugging is associated with an extra $499,700, meaning away slugging performance is more than 67% more important when predicting arbitration salaries than home performance. This highlights the surprising importance of away performance in salary valuation, not only in a player's raw away performance but also in the relative strength of their away performance.

Moving to in-season adaptability, only the First/Second Half Slugging Difference is significant. The model suggests that a one standard deviation improvement in performance in the second half of a season relative to the first half across pre-arbitration years is associated with $79,500 increase in salary. The coefficients on the difference between a player's best and worst 1/8ths of a season are quite large in Models I and II, but disappear once controls are added, suggesting their effect can be attributed to another factor. Even still, this model suggests that consistent improvement over the course of a season early on in a player's career is positively associated with arbitration salaries.

The effect of short-term volatility reveals an interesting interaction between consistency and performance. In the fully controlled model, higher short term volatility is negatively associated with salary. A one standard deviation increase in volatility across a player's pre-arbitration seasons is associated with a decrease of $163,700 in a player's first arbitration salary. When performance is not included in the model, short-term volatility appears to have a positive impact on salary. However, once salary is controlled for, the effect becomes significantly negative. A possible explanation for this is that higher performing players inherently have higher volatility, as their slugging percentage may be very high in games

where they did well. Once performance is accounted for, however, this model shows that less volatility is advantageous for pre-arbitration players.

5.2 Lasso Regression Results

While the OLS regression provides valuable insights, Lasso regression plays a crucial complementary role. By introducing regularization, Lasso mitigates overfitting concerns in the OLS model, strengthening the reliability of the findings. Furthermore, Lasso's feature selection highlights the most important consistency metrics for predicting salary. Analyzing both approaches together gives a better understanding of the relationship between consistency and salary, leading to more confident conclusions. Table 5.2 compares coefficients between OLS Model IV and my Lasso model. While Lasso doesn't provide standard errors or R^2 in the same manner as OLS, it highlights key variables and potential redundancies. Notably, Lasso shrunk the coefficients of Short-term Volatility and Home/Away Slugging Difference to zero, suggesting their predictive power might be partially explained by other included variables. It is possible that both effects were attributed to home and away raw slugging performance, as the performance coefficients were not absorbed. Interestingly, the coefficient on away SLG increased slightly, while the coefficient on home SLG decreased. This further supports the idea that away performance is a more important determiner of arbitration salary than home performance. Both coefficients on in season adaptability metrics increased slightly, implying that their effect could not be explained well by any other regressors. The Lasso analysis reinforces the significance of consistent away performance in salary valuation and suggests that though there may be some redundancy in the consistency metrics, their effect is important for predicting arbitration salaries.

5.3 Random Forest Results

Finally, I present results from a random forest model. This model shows concrete feature importance plots and captures any non-linear or complex relationships within the data. The random forest model produced an R^2 value of 0.7235, which was better than any of the OLS

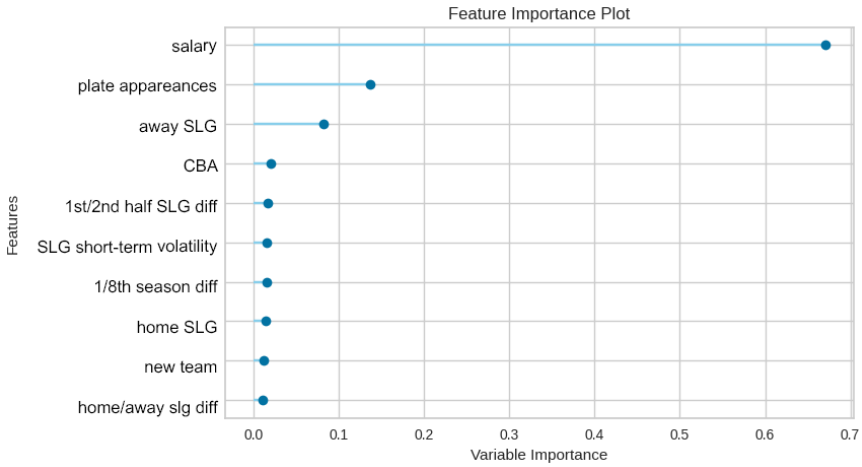
Table 5.2: Lasso Regression Results Comparison

	Lasso	Model IV
Intercept	31.988	30.602
Salary	12.827	12.617
Short-term Volatility		
Short-term SLG Volatility	0.000	-1.637
In Season Adaptability		
First/Second Half SLG Difference	-0.926	-0.795
1/8th Season Largest SLG Difference	-0.674	-0.400
Environmental Consistency		
Home/Away SLG Difference	0.000	-1.166
Home SLG	1.037	2.991
Away SLG	5.011	4.997
Controls Added	Yes	Yes

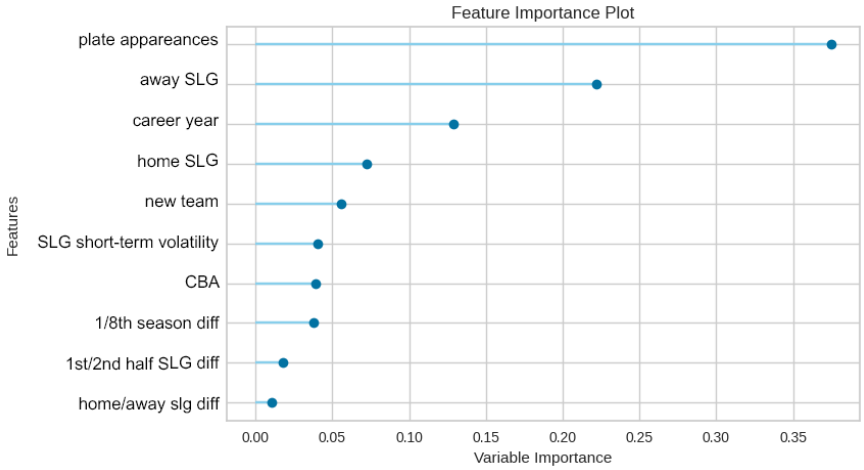
Table 5.2: Lasso Regression Results Comparison. Regressors are normalized and predict arbitration salary in $100,000s. All regressors are pooled from a player's pre-arbitration seasons and use SLG as the statistic of interest. The sample includes all eligible player seasons from 1995-2019, with eligible seasons defined in methods. Added controls include pooled plate appearances, Collective Bargaining Period, career year, and a changed teams indicator. Standard errors and R^2 are not available from the Lasso model. The penalty weight for the lasso model is $\alpha = 0.1$.

models, implying that there may be some complex relationships in the data not captured by OLS. The feature importance plot can be seen on the left side of Figure 5.1 and shows that the final pre-arbitration salary was by far the most important predictor of post-arbitration salaries. This makes intuitive sense, as the salary variable accounts for not only time period, but also perceived value through a player's signing bonus. Because salary was weighted so heavily, I created a secondary model without salary included as a regressor. Instead, I normalized by pre-arbitration salary to find the total change in salary after a player entered arbitration. Without salary included, the model did not perform nearly as well, but it was much easier to interpret relative feature importance, as can be seen on the right side of Figure 5.1. Beyond salary, other important features were away slugging and total pre-arbitration plate appearances. Again, away performance was much more important than home performance in the Random Forest model. The consistency metrics were all similarly important, with short-term volatility being the most important.

Figure 5.1: Random Forest Feature Importance Plots



(a) Salary Included



(b) Salary Excluded

Figure 5.1: Random Forest Feature Importance Plots. X-axis is relative feature importance in each model. The models are identical except for including or excluding salary.

Chapter 6

Conclusion

6.1 Impact

The set of models in this paper demonstrates that while performance metrics, previous salary, and prevailing league conditions are the most robust predictors of a player's first-year arbitration salary, consistency metrics also hold explanatory power. The significance of these consistency measures is particularly noteworthy, as they quantify an aspect of player value that has not previously been measured. Specifically, the short-term volatility metrics show that less volatile players experience better outcomes, supporting the common belief that it is beneficial to perform at a consistent level day to day. The results show that teams may implicitly favor players who can be trusted to maintain a steady level of play. Building on this, the in-season adaptability results show that players who improve throughout a season are favored. For young players just entering the MLB, the ability to not only maintain but improve their level of play throughout a season holds value amongst teams. Perhaps most surprising are the effects of environmental consistency. A young player's ability to perform on the road holds significant value in determining their future contracts. Players that match or exceed their home performance end up with higher earnings, and the weight placed on away performance exceeds that which is placed on home performance by over 67%.

Overall, these findings suggest that while teams prioritize known statistics and market forces when making salary decisions, they also either implicitly or explicitly acknowledge the potential for consistency and adaptability to translate into future success, as evidenced

by the premium placed on those who demonstrate these traits during their pre-arbitration years. It is not clear if teams currently look at consistency when valuing players or if it is implicitly worked into their valuations through other metrics; however, in the broader baseball community, the mentions and metrics of consistency are sparse. This suggests that this aspect of a player's performance does not receive as much attention as it should. Early career consistency has statistically significant impacts on future outcomes, and understanding how a player's performance varies across games, environments and seasons gives valuable insights into their future career trajectories.

This study also contributes to the understanding of broader trends within the MLB. With the worrying decline of balls in play, the MLB seeks to understand if this is a result of pitcher skill improvement, or if higher-variance plays are being encouraged by teams. While both may be true in some regard, this study finds that teams do not place value on generally inconsistent players. Teams prefer players who can be trusted to perform at a stable level on any day, in any environment, and at any point throughout a season. Teams do not reward highly volatile players with larger contracts; rather, they do the opposite. Though the game of baseball is trending towards fewer balls in play, this analysis suggests that it is not the result of teams incentivizing high-variance batters.

6.2 Future Work

This study introduces a significant new dimension to player valuation, demonstrating consistency as a valuable predictor of career outcomes. Even with this finding, there is still much work to be done on the topic. First, this paper's analysis of consistency is limited to batting. Future work on this topic could include a study of consistency in pitchers, or incorporating fielding into a player's valuation using the advanced statistics that have become increasingly available. Another method of analysis could be using more advanced offensive statistics when creating a player's consistency profile. While advanced data was too limited for the timeframe in which I did my analysis, it could be useful for understanding consistency in more recent MLB seasons. Additional work could be done using the same techniques from this paper on the data from developmental stages of baseball, including high-school, college,

or minor league data. Here, there is much more variability in the level of play, meaning raw statistics are inherently less valuable. Understanding consistency in players at these levels could provide useful insights into a player's potential for major league success.

This study contributes to the understanding of player valuation in multiple ways. It introduces three definitions of consistency and creates novel consistency metrics that capture each facet of a player's early-career performance patterns. Then, by incorporating these alongside standard performance measures, it highlights how both raw performance and the consistency of that performance factor into salary determination. By introducing consistency as an important metric in player valuation, this paper paves the way for more studies to explore the nuances of these metrics so that they can be better understood. This knowledge will contribute to teams' valuations of players, the league's analysis of broader trends, and ultimately a better understanding of baseball as a whole.

Appendix A

Software Packages

All code for this thessis was written in python. Included below is a list of packages that were used to complete my research.

- **Numpy:** Used for numerical computation with arrays and matrices
- **Pandas:** Used for analyzing and manipulating data structures
- **Matplotlib:** Used for creating a variety of plots
- **Seaborn:** Used for data visualizations
- **Statmodels:** Used for statistical models and testing
- **Scikit-learn:** Used for machine learning algorithms
- **PyCaret:** Used to compare a variety of machine learning algorithms

References

[1] J. Rogers, *MLB is a sport divided by historic payroll disparity – so what's next? — espn.com*, https://www.espn.com/mlb/story/_/id/37775153/mlb-divided-historic-payroll-disparity-next, [Accessed 30-04-2024], 2023.

[2] E. Wassermann, D. R. Czech, M. J. Wilson, and A. B. Joyner, *An examination of the moneyball theory: A baseball statistical analysis*, Mar. 2015. URL: https://thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/.

[3] B. Harris, *A sabermetric primer: Understanding advanced baseball metrics*, Feb. 2018. URL: https://theathletic.com/255898/2018/02/28/a-sabermetric-primer-understanding-advanced-baseball-metrics/.

[4] *Mlb gloassary*. URL: https://www.mlb.com/glossary/statcast.

[5] S. J. Spurr, "The baseball draft: A study of the ability to find talent," *Journal of Sports Economics*, vol. 1, no. 1, pp. 66–85, 2000. DOI: 10.1177/152700250000100106.

[6] A. Dorney, *How do mlb rookie contracts work?* Nov. 2022. URL: https://www.sportskeeda.com/baseball/news-how-mlb-rookie-contracts-work.

[7] E. Sarris, *How would we increase balls in play?* Aug. 2017. URL: https://blogs.fangraphs.com/how-would-we-increase-balls-in-play/.

[8] R. T. Karcher, "The chances of a drafted baseball player making the major leagues: A quantitative study," *Baseball Research Journal*, no. Spring 2017, 2017.

[9] N. Longley and G. Wong, "The speed of human capital formation in the baseball industry: The information value of minor-league performance in predicting major-league performance," *Managerial and Decision Economics*, vol. 32, no. 3, pp. 193–204, 2011. DOI: 10.1002/mde.1526.

[10] J. M. Congdon-Hohman and J. A. Lanning, "Beyond moneyball," *Journal of Sports Economics*, vol. 19, no. 7, pp. 1046–1061, 2017. DOI: 10.1177/1527002517704019.

[11] G. Chandler and G. Stevens, "An exploratory study of minor league baseball statistics," *Journal of Quantitative Analysis in Sports*, vol. 8, no. 4, 2012. DOI: 10.1515/1559-0410.1445.

[12] H.-C. Sun, T.-Y. Lin, and Y.-L. Tsai, Jun. 2022. URL: https://arxiv.org/pdf/2206.09654.pdf.

[13] C. Heaton and P. Mitra, Mar. 2022. URL: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/622770fbd0f4027cbec673bd_Using%20Machine%20Learning%20to%20Describe%20How%20Players%20Impact%20the%20Game%20in%20the%20MLB%202.pdf.

[14] S. R. Bailey, J. Loeppky, and T. B. Swartz, "The prediction of batting averages in major league baseball," *Stats*, vol. 3, no. 2, pp. 84–93, 2020. DOI: 10.3390/stats3020008.

[15] M. Petriello, *Panel explores ideas to improve contact*, Feb. 2023. URL: https://www.mlb.com/news/sabr-three-ways-to-get-more-balls-in-play-c239668382.

[16] J. Hwang, *The case for consistency metrics in sports — visualnoise.substack.com*, https://visualnoise.substack.com/p/does-consistency-lead-to-more-wins, [Accessed 30-04-2024], 2020.

[17] D. T. Turocy, *Chadwick baseball bureau*, https://github.com/chadwickbureau/retrosplits?tab=readme-ov-file, Accessed: 2023-12-01.

[18] Retrosheet, https://www.retrosheet.org/, Accessed: 2023-12-01.

[19] *Cot's baseball contracts*, https://legacy.baseballprospectus.com/compensation/cots/, [Accessed 30-04-2024].

[20] S. Lahman, *Lahman database*, https://seanlahmnan.com, [Accessed 30-04-2024].