Star Power in Major League Baseball:

The Attendance Effects of Starting Pitchers

Dylan Phillips

Syracuse University

SAL 496: Sport Analytics Seminar II

Dr. Rodney Paul

April 21, 2024

Abstract

One of the major sources of income for organizations in the sports realm is through ticket sales. Maximizing this revenue by increasing attendance can lead to an advantage compared to other teams, both in the sport and as a business. This is no different in Major League Baseball. With a 162-game season and 81 home games, learning how to maximize attendance can lead to an unparalleled advantage over other teams. In general, the better a team plays, the more fans will come to the games. Success in baseball has many factors, but one of the most important is the quality of the starting pitcher. If a team is looking to increase attendance, will investing in big-name, successful pitchers in free agency and contract negotiations lead to an increase in attendance? Star power in the NBA has been well-researched, but how does this past research apply to MLB? Using attendance and pitching data from 2021-2023, three ranger random forest models were created, focusing on team aspects, pitcher aspects, and the two together. Variable importance measures found that the home team and the success of both teams playing had more of an effect on tree creation than pitcher success did. In the pitcher model, measures of overall pitching success, such as FIP, took precedence. Overall, the models tended to underpredict attendance, but there were still apparent patterns. On a team basis, the teams that had above-expected attendance were historically successful or were in large markets, like the New York Yankees and Los Angeles Dodgers. On a pitcher basis, attendance was above expected for All-Star level talent, such as Gerrit Cole and Bryce Elder. These models are versatile, as the results and methods applied to create them can be applied to other sports and other positions, and, once optimized, should be accurate predictors of star power.

Introduction

As the modern world continues to become more fast-paced, sports fans are beginning to turn away from America's Pastime in favor of more action-packed sports. The National Football League and the National Basketball Association have taken over the public's eye in recent years, with increased attendance and viewership, while Major League Baseball has struggled. Its slow-paced play has failed to catch the eyes of young fans, and MLB organizations are trying to reverse the downward trend in attendance.

The sports world has changed dramatically in the last decade, as the marketability of players has become an increasingly important factor in viewership, attendance, and profits. The NBA's small roster size and star-centered play have allowed it to blossom into an industry of its own. The NFL's large fanbase and reach have steadied it at the top of North American sports. Although it is shrinking, MLB's reach is as expansive as the NFL's, with a steady hold in Latin America. With this reach, the factors necessary for the growth of MLB's pull in the market are present, but it has fallen behind in the marketing of players.

One determinant in attendance is the players the fans came to see. In baseball, the only way to predict who they will see in a given game is the starting pitcher. Oftentimes, teams will follow a steady rotation for their starters, while batting lineups are often set the day of the game. Because of this, in theory, player effects on attendance will be encapsulated more in the starting pitchers than in any other player. There is no doubt that superstar pitchers help their teams succeed on the field, but how do they help the organization on a game-by-game basis? To see the magnitude and quantify the effect that starting pitchers have on attendance in Major League Baseball, a random forest modeling approach will be conducted, using recent attendance and pitching data.

Literature Review

Because the MLB season is nearly twice as long as other sports, gaining a slight advantage in attendance can contribute to a considerable advantage over other teams in terms of profit. In the past, substantial research has been done on how superstars draw larger crowds in the NBA, which is considered a star-based league. However, the same principles can be used for baseball, and if starting pitcher caliber leads to a bump in attendance. Literature was researched to see how attendance is affected by on-field performance, including researching the direct impact of team payroll, starting pitcher performance, star power, and ticket pricing on attendance.

Team Payroll

Hall, Szymanski, and Zimbalist used the examples of Major League Baseball and English Soccer to study the effect of on-field performance and payroll since both leagues have similar salary structures, however, soccer player markets are much less regulated. To do this, they used linear regression to estimate how win percentage affects payroll, win percentage on the team payroll relative to the league average for the year, and average win percentage over the last ten seasons on the average payroll over the same period. When examined, it was found that team payroll relative to average explained about 24% of the variation in regular season winning percentage (R^2 value of .2364). However, when looking at the average regular season winning percentage over the range of the data (1980-2000), it was found that variation in payroll explained more than 70% of the variation in average winning percentage, with an R^2 value of 0.7067 (Hall, et al. 2002). Additionally, Granger causality tests were run, which found that between 1980 and 1994, there was no evidence of causality between wages and performance, but there is evidence of causality in both directions after the 1994 season. In their research on soccer,

they found greater evidence to support the claim that payroll and performance have a relationship. This is most likely due to the structure of the player markets, with baseball's free agency being much more regulated than the market for players in soccer. Their strategy in using relative payroll instead of total payroll will be used when analyzing relative pitcher payrolls.

Using the results found in the previous study by Hall, Szymanski, and Zimbalist, Nelson and Dennis used simple linear regression to see the true tradeoff between win maximization and profit maximization. To do this, they separated team revenues into parts, such as operating income and gross profit margin (Dennis and Nelson, 2015). They used winning percentages to measure team performance. Although this study does not directly relate to the research question, a good knowledge of how team success both on the field and as a business affects payroll is necessary. Again, because of baseball's long season and the comparative advantage that can be gained from maximizing attendance, Mitchell Woltring examined the relationship between winning percentage and attendance using crosstabs, ANOVA, regression, and logistic regression analyses. Controlling for the year, stadium capacity, and team payroll, all these analyses found a positive relationship between average attendance in the form of percentage of capacity and winning percentage. Regression analysis proved to be the most relevant, finding an R2 value of .242, with significance at the .001 level (Woltring, 2014).

Individual Player Effects

Fishman tested the claim of the Coase Theorem that free agency being introduced in baseball would not influence competitive balance, as the distribution of players would not change. Using a linear model and data from 1950 to 2001, Fishman found a strong relationship between the number of free agents signed in the offseason and the deviation of winning percentages the following season. Within the linear model, the coefficient for the number of free

agents signed was positive and significant at the .01 level, supporting the owner's claims that free agency would harm competitive balance once introduced. Additionally, it was also found that a reverse-order amateur draft had a negative effect on win percentage deviation, meaning it made the league more competitive (Fishman, 2002). From the perspective of a team owner, this study justifies spending during free agency, if the team's goal is to maximize winning.

Diving deeper into the effect players have on attendance, Russo asked if the quality of the starting pitcher influenced attendance at games. To do this, he built a model using a random forest that would predict attendance at games using factors such as opposing team, month, game type (weekday or weekend), and time of day (day or night). He then took these predicted numbers and overlayed them with the actual attendance at that game to find the percentage error. The average differential for each starting pitcher (minimum ten starts) was calculated, and each pitcher was grouped using fWAR. It was found that attendance at games in which a pitcher in the best group (90^{th} percentile fWAR) started was often underestimated by the model, while attendance with pitchers in the above average group (50^{th} – 90^{th} percentile fWAR) was the most accurately predicted. These findings show that the quality of starting pitchers does drive attendance upwards, but only those that are considered the "best" (Russo, 2019).

Ormiston measured starting pitcher star power, which is a metric that measures how big of a star a player is and its effect on in-game attendance. Ormiston found that there is a positive and statistically significant relationship between attendance and the measured star power of both the home and opposing starting pitchers. Ormiston also found that fans' responsiveness to star power declined over the scope of this study, meaning the increase in attendance by these pitchers may become minimal in the future (Ormiston, 1970). Ormiston referenced Rivers and DeSchriver's research in the previous article. Its goal was to create a demand model that

explained the relationship between the fluctuations of team payroll and attendance. Additionally, they also investigated the effect that star players have on attendance. The results found that star players do not have any effect on attendance if they do not contribute to on-field performance (Rivers and DeSchriver, 2015). It also found a negative relationship between payroll variation and attendance, meaning teams are better off spreading out their payroll if their goal is to maximize attendance.

Lewis and Yoon used a created statistic called star power, or SP, to find how players with high star power are developed and how it impacts both ticket demand and overall team performance. Star power is a congregate of salary, individual performance, and awards. In this, they found that a player's salary is not just related to pure performance, but it depends on a player's star power and brand. Additionally, they investigated how teams with a high total star power can drive ticket demand controlling for ticket prices. They did this in many ways, but one interesting finding was game attendance before and after the Dodgers signed Manny Ramirez. The model found that this signing, along with a focus on Ramirez's branding, brought an estimated 4,815 fans to the stadium per game (Lewis and Yoon, 2021).

Ormiston posed the question: do baseball fans prefer homegrown players to be the stars in their systems? Specifically, he studied game-by-game attendance from the dawn of free agency in baseball to the present and found no evidence that fans prefer homegrown pitchers (Ormiston, 2014). It may be quite possible that the effect is purely intangible. In this paper, Ormiston continued to study the role that homegrown pitchers have on attendance. He looked at the attendance effects that these star pitchers have using multiple factors, including preseason prospect rank, draft pick status, and performance. It found a small but statistically significant positive result, but only for particularly elite prospects (Ormiston, 2016).

Star Player Analysis in the NBA

Narrowing in on the effect that star players have on attendance at events, Humphreys and Johnson studied the externalities that "superstars" bring in sports and how this affects attendance. This research was focused on the NBA, but its modeling strategies can be applied to other sports. First, they determined which players were deemed superstars by ranking how many times they appeared in the top five yearly salaries. Its results found that Michael Jordan, considered the biggest star in the league, brought the greatest effect with him. After analysis, it was found that these superstars drew higher attendance on a typical game in which they played than a typical weekend game. This study also dove into novelty effects, examining if a player who stays with a team for many years has the same effect on attendance as a superstar on a new team. The data saw a large increase in attendance in the first year that a player is on a team, and a slow decrease after, suggesting that longer contracts may not have the same payoff as short-term contracts (Humphreys and Johnson, 2019).

In this study, Slusser found more of the same effect. A player was determined if he had an efficiency rating of 20 or more during the season, and it was found that teams with a superstar should expect an increase in attendance of 405 fans per game. Over an 82-game season, this translates to 16,605 additional fans, or a yearly revenue boost of just over $1.2 million, based on the average NBA ticket price of 73.66 (Slusser, 2021). This study also found that individual player accolades, such as MVP and all-star appearances, have an additional positive effect on attendance. Slusser also mentions the opportunity cost of signing these players: could it be more cost-effective to sign a star player to a smaller contract than what a superstar would demand?

Colin Josselyn looked at Lebron James, who is the modern-day equivalent of Michael Jordan, to see if having him on the team was "worth it." Worth was quantified by three things:

team performance, attendance, and franchise valuation. Other than his recent stint with the Los Angeles Lakers, teams with Lebron James saw an increase in all the studied statistics (Josselyn, 2019). Although this study focuses on basketball, it can still be used to influence a model for baseball's superstar effect.

Aging Curves

Aging curves serve an important role in player contracts. The ability to estimate the effectiveness of a player in their later years can help teams during contract negotiations. Diving into how pitchers perform as they age, Scott Lindholm used fWAR to calculate when pitchers reach their peak, hypothesizing a slight increase, then a sharp decrease in the average fWAR by age after the mid-twenties. Instead, the perceived relationship between age and fWAR was relatively positive, with the highest value at age 45. This result was found due to survivorship bias. More detailed analysis backed this theory, with the number of pitchers in the data set by age peaking at 26, along with the cumulative fWAR by age (Lindholm, 2014). As a pitcher ages, expectations are much higher, leading to high dropout rates in the thirties. Although, in this case, aging curves relay misleading results, finding players that differ from these curves can be a sign of success in the latter years of a contract.

Bill Petti also wrote an article on aging curves, with a focus on starting pitchers. This time, the variable of interest was velocity. It is found that velocity peaks in a pitcher's early twenties, and steadily decreases until their age 26 season, backing Lindholm's hypothesis in the previous study (Petti, 2012). After this, velocity drops at a higher rate. Separating the data into starters and relievers, Petti found a slightly lesser decrease in velocities in relief pitchers, likely because of workload differences. Aging curves using one variable can also be misleading, as a velocity decrease does not necessarily represent a decrease in effectiveness. The research saw a

rise in FIP as relief pitchers age, while it is relatively steady for starters, further showing how survivorship bias can affect aging curve results.

Ticket Pricing and Other Effects on Attendance

Because ticket prices are the most influential determinant in attendance, any model that attempts to predict attendance must include pricing. This study by Young Lee is like other attendance models, in that it looks to find common factors in attendance. However, this looked at how these common factors changed over the years. It was found that in the early years, attendance was influenced less when compared to the modern years, which are more focused on offensive performance. In modern baseball, it was found that not only does the home team influence attendance, but the away team's performance also has a significant effect (Lee, n. d.). This led to a shift in MLB's advertising, investing more in games in which popular, offensively powerful teams were visiting. With the invention of the internet, organizations were able to utilize more information to their advantage and begin using variable ticket pricing. This means that ticket prices change dependent on many factors. Games against better opponents and nights on the weekend will have higher demand, and, in turn, higher ticket prices. Their study found that teams who used variable ticket pricing would see an average increase of $590,000 in additional ticket revenue (Rascher, 2007). There is more to look at in variable ticket pricing, such as how pitching matchups influence pricing. There is also a mention of demand elasticities in ticket pricing, which is an important factor in an attendance model. Each team, or city, will have its own reaction to a rise in the average ticket price.

This study by Scott Thauwald attempted to find which factors were the greatest in determining baseball attendance. The ten independent variables used were regular season wins, ticket price, team payroll, facility age, number of teams within a sixty-mile radius, star players (if

they made it to the previous all-star game), total home runs, playoffs, city population, and household median income. The results of the regression found that payroll and regular season wins were the most significant (Thauwald, n.d.). Although Reifer's research did not seem too reliable, he brought up some decent ideas on what might highly influence per-game attendance in baseball. He found that Opening Day games and games against the Yankees had much higher attendance than the average game (Reifer, 2020). However, this might mean that the opposing teams' payroll and players influence individual game attendance figures more than the home teams, which is not the research question being asked.

Hepper mostly looked to be able to predict MLB game attendance based on several factors, including games back, month, weekday, game time, and win percentage. What was intriguing about his research was how he ran a regression on attendance and the average age of players and found a relatively strong positive correlation between the two. That is, the older the team, the more fans it drew to the ballpark (Hepper, 2017). This may be evidence to support the hypothesis that individual players will affect attendance at a season-long level. This could also support a team's decision to sign an older star to drive up attendance. Cohen looked at the reverse relationship between team spending and attendance: does more fans in the stands have a strong influence on the team's willingness to pay players? To do this, he took team payroll data from 2009-2013 and compared them to the average league payroll. He then did a fixed effects model on team attendance relative to the average for that same period. Cohen does this to control for differences in team strategy. That is, the Red Sox are not being compared to the Pirates, but the Red Sox are compared to the Red Sox in previous years. He found a weak, positive relationship between the two, meaning there is likely more to be studied. Cohen then filters the data to find seasons in which a team's attendance changed drastically between seasons. Using the

example of the 2012 Marlins, when management increased payroll by a factor of .26. The following season saw a spike in attendance, but the team also moved into a new stadium. This shows how many factors can influence attendance on a season-by-season level (Cohen, 2015).

Although the true results of the following study are unknown, it brings forth good research ideas in terms of demand elasticities. Langhorst looked to see how a winning team and high payroll would affect attendance. To do this, he used game attendance and ticket prices within his model. What he found was it was highly dependent on the team. Historically successful teams such as the Yankees and the Phillies will have high ticket demand even in down years, while teams such as the Rays have lower demand despite their recent success (Langhorst, 2014). This introduces an important research question: how do demand elasticities for ticket sales affect team strategies?

The study by Drayer, Shapiro, and Lee on dynamic ticket pricing compares how sports teams price their tickets differently than other larger industries, such as the airline and hotel industries. Instead of prioritizing fan happiness in pricing their tickets, teams are beginning to price using dynamic pricing with the rising costs of operations (Drayer, et al., 2012). Although this study is theoretical in how teams can begin to maximize profits using a more demand-based approach to pricing, it gives some information on how these organizations price their tickets. Shapiro, Drayer, and Dwyer wrote an extension on the previous study on dynamic ticket pricing (DTP) in the sport industry. This study investigated how the popularization of the secondary market for tickets influences primary ticket pricing. It also investigates specific variables in what a fan will pay for a ticket, such as seat location, team performance expectations, and game quality (Shapiro, et al., 2016). This can prove to help create a model that uses pitcher matchups

as a variable in ticket pricing. It also considers fan demographics, which can help create a model for different cities' ticket demand.

To get a more rounded understanding of what influences attendance, Barilla, Gruben, and Levernier investigated how promotions draw people to games. It was found that, on average, promotions brought 1,532 more fans to the stadium than a typical home game (Barilla, et al. 2008). It also saw a small increase in attendance at the typical Friday or Saturday game, as expected. Incorporating promotions into a model may be necessary in predicting attendance for a given game.

The paper by Kirk Wakefield on social influence is less of a data-driven study, but it will help attempt to see the relationship between social behavior and attendance at games. It uses the example of the Pittsburgh Pirates' low attendance and attempts to explain it using social behavior. One thing that is emphasized is family and friends' perceptions of the stadium. A control group is used, and it is found that family perceptions of parts of the gameday experience, such as stadium and concession quality, have a significant effect on attendance (Wakefield, 1995). Chris Harvey described the faults in measurements in fan engagement in the modern day, and how it can be improved through data science. The problem with modern measures is that they leave out important variables and have skewed measures, such as the exclusions of some streaming platforms and the casual fan. MRV Data Science developed a tool that integrates all underlying drivers of engagement which will result in a more accurate and in-depth value for fan engagement (Harvey, 2020). Integrating these findings into a model may be helpful.

Data and Modelling

Jeffrey Näf went into depth about how one can improve a random forest model. The weakness of a random forest is that it is difficult to interpret the effect the variables have on the

final output (Näf, 2023). To combat this, the author suggests using Variable Importance Measures (VIMP). This can help to decrease the number of variables that need to be interpreted in the final model, while not decreasing accuracy at a significant level.

Body

Data Description

To have a full understanding of the models that will be run, knowledge of the dataset I used and its components is fundamental. Before going into depth about each variable, I will walk through how data was collected and its source, some necessary cleaning, and the merging of datasets.

The model I had in mind required both game-by-game data and season statistics. First, I needed information about each MLB game played within the span of research. Ideally, many seasons of data would be used, but, because of the effect COVID-19 had on the 2020 MLB season and its attendance, I decided it would be best to keep the range of data from 2021-2023. This data was scraped from ESPN.com, cycling through the game summary pages. From this page, I was able to find the home and away teams and their respective after the game concluded, listed starting pitchers for each team, game date, first pitch time, and capacity.

Once all three seasons' worth of data was collected, which totaled 7,290 observations, some columns were cleaned for readability. The game date was separated into month, day, and year. This way, the importance of the month the game was played could be measured. Next, the time of the first pitch was separated into hours and minutes, and the minutes variable was ignored in further analysis. For example, a game that started at 7:35 would be considered to have

started at 7. Finally, the main variable of interest, the percentage of capacity for each game was put into decimal form, so a sold-out crowd would be represented as a 1.

The second data set that was needed was pitcher statistics. A well-rounded sample of important performance statistics was necessary to study the full effect that starting pitchers had on attendance. Performance statistics for each pitcher for that season were found on baseball-reference.com. Preferably, pitcher statistics up to the time of each game would have been more accurate when it came to modeling, but they were unavailable. There may be some inaccuracies in the data because of this, as a pitcher's performance on Opening Day would be measured by games that had yet to be played. Other than this caveat, a full analysis of pitcher performance was analyzed.

After collection, pitchers that did not start a game were removed from the dataset. Additionally, many redundant statistics from the Baseball Reference dataset were removed to get a concise analysis of each pitcher. In the end, the pitcher's name, pitching record, number of home starts, and both simple and advanced performance metrics were used. Minor cleaning was necessary for this dataset to prepare for merging with the previous.

To merge two datasets from different sources, a common primary identifier was needed. For this, I used the pitcher's last name, team abbreviation, and year, separated by underscores. With this, I was able to add the statistics collected from Baseball Reference for both listed starting pitchers for each game, connecting pitcher performance and capacity. After this merger, some minor cleaning was necessary for analysis, as pitchers who had appeared for multiple teams in a season had separate observations for their performances with both teams and collectively. To fix this problem, overall performance was analyzed for these instances.

Variables

After cleaning, the dataset contained 7,223 observations and 58 variables. Some observations were deleted, as the listed starting pitcher did not have a significant number of appearances in the Baseball-Reference dataset. To prevent overfitting, a significant number of variables were not used in modeling.

Contained within the dataset scraped from ESPN.com, team and game-specific variables were used. Omitted from this dataset was the first pitch time in minutes, and the date the game was played, as they did not make intuitive sense in modeling. The following is a list of variables and their descriptions that were used in modeling:

capacity: ballpark attendance in percent of maximum capacity (dependent)

home_team_abbrev: abbreviated name of the home team (ex. Boston Red Sox is BOS)

away_team_abbrev: abbreviated name of the away team

home_wins: number of wins the home team had after the game

home_losses: number of losses the home team had after the game

away_wins: number of wins the away team had after the game

away_losses: number of losses the away team had after the game

month: name of the month the game was played in

time_hours: time in hours of first pitch (ex. 7:35 pm first pitch is 7)

The dataset from Baseball Reference containing season-level pitching statistics was trimmed substantially before modeling. After running some initial models, it seemed the away team’s performance that season had little to no effect on the stadium's capacity, measured by variable importance. To create a simpler model, these variables were removed from the dataset.

Additionally, overlapping statistics were removed to limit bias. The following are the remaining pitcher variables:

home_starter_id: home starting pitcher's identifier with team and year (ex. Zac Gallen's 2023 season is Gallen_ARI_2023)

away_starter_id: away starting pitcher's identifier with team and year

home_starter_name: home starting pitcher's last name and team (used in modeling instead of home_starter_id to separate pitcher and season. Additionally, the pitchers who were not in the top fifty in appearances in the data were deemed as "other," as random forest models do not work with categorical variables with more than 51 values.)

away_starter_name: away starting pitcher's last name and team (used in the same way as home_starter_name)

home_starts: number of starts the home starting pitcher had in the season

hp_age: home starting pitcher's age in years

hp_wins: number of pitching wins by the home starting pitcher

hp_losses: number of pitching losses by the home starting pitcher

hp_era: earned run average (ERA) of the home starting pitcher, calculated by

$$9 * \frac{earned\ runs}{innings\ pitched}$$

hp_ip: number of innings pitched by home starting pitcher

hp_eraplus: League Adjusted ERA (ERA+) of the home starting pitcher, calculated by

$$100 * \frac{League\ ERA}{ERA}$$

hp_fip: Fielding Independent Pitching of the home starting pitcher, calculated by

$$\frac{13*home\ runs + 3*walsk - 2*strikeouts}{innings\ pitched} + C$$ where C is a centering constant

hp_whip: Walks and Hits per Inning Pitched (WHIP) of the home starting pitcher, calculated by

$$\frac{walks\ +\ hits}{innings\ pitched}$$

hp_h9: hits per nine innings of the home starting pitcher, calculated by

$$9\ *\ \frac{hits\ allowed}{innings\ pitched}$$

hp_hr9: home runs per nine innings of the home starting pitcher, calculated by

$$9\ *\ \frac{home\ runs\ allowed}{innings\ pitched}$$

hp_bb9: walks per nine innings of the home starting pitcher, calculated by

$$9\ *\ \frac{walks\ allowed}{innings\ pitched}$$

hp_so9: strikeouts per nine innings of the home starting pitcher, calculated by

$$9\ *\ \frac{strikeouts}{innings\ pitched}$$

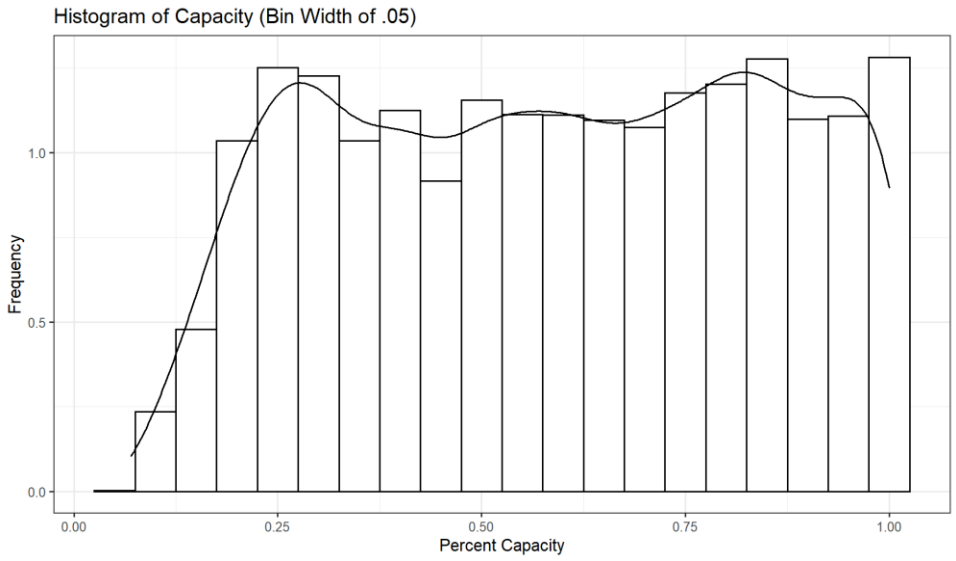
Summary Statistics

Below are the summary statistics of the above variables, first from the team factors dataset, then the pitcher factors dataset.

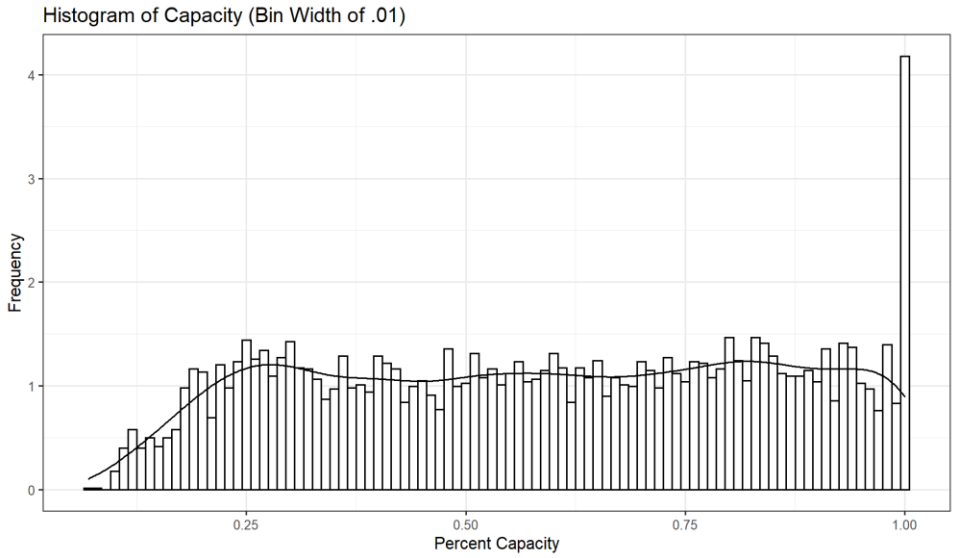
	mean	sd	first_quartile	third_quartile	min	max	skewness	kurtosis
capacity	0.589	0.257	0.360	0.810	0.070	1	-0.056	-1.211
home_wins	40.927	24.840	20	60	0	111	0.246	-0.840
home_losses	40.897	24.790	20	59	0	110	0.261	-0.771
away_wins	40.961	24.828	20	60	0	108	0.222	-0.901
away_losses	40.852	24.716	20	60	0	112	0.255	-0.801
time_hours	5.962	2.829	4	8	1	12	-0.360	-0.806

	mean	sd	first_quartile	third_quartile	min	max	skewness	kurtosis
home_starts	11.275	4.928	8	15	1	27	-0.207	-0.368
hp_age	28.591	4.006	26	31	20	43	0.794	0.393
hp_wins	7.070	4.400	4	10	0	21	0.407	-0.556
hp_losses	6.772	3.488	4	9	0	19	0.350	-0.065
hp_era	4.318	1.466	3.370	4.970	0.560	22.500	2.555	18.345
hp_ip	119.419	51.886	80.100	161.200	1.100	216	-0.290	-0.892
hp_eraplus	106.934	36.829	85	123	21	755	4.201	52.334
hp_fip	4.274	1.071	3.640	4.820	1.240	23.020	2.067	21.145
hp_whip	1.296	0.238	1.145	1.420	0.500	4.500	1.583	10.343
hp_h9	8.641	1.698	7.600	9.500	0	40.500	1.964	23.381
hp_hr9	1.267	0.537	0.900	1.500	0	12.300	3.223	41.397
hp_bb9	3.027	1.096	2.300	3.600	0	13.500	1.546	6.776
hp_so9	8.433	1.807	7.200	9.500	0	15.200	0.197	0.105

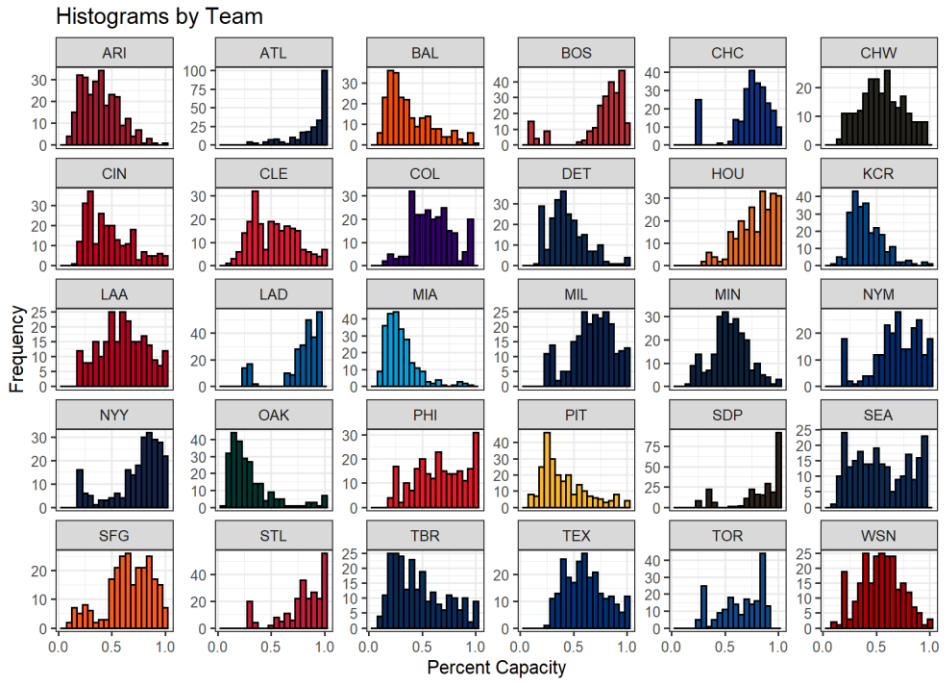
Most of the summary statistics are trivial. Typically, before data analysis, studying the shape and behaviors of all variables is important to correctly model. However, because a random forest model will be used for analysis, the shape of the distribution of the data, especially the dependent variable, is not important. Random forest models are non-parametric, meaning distributions of variables do not affect results.



The histogram shows that capacity looks relatively uniform. Other than the bins between 0 percent and 15 percent, the data has similar frequencies above the 15 percent level. At this bin width, however, we cannot get a complete understanding of the true shape of the dependent variable.



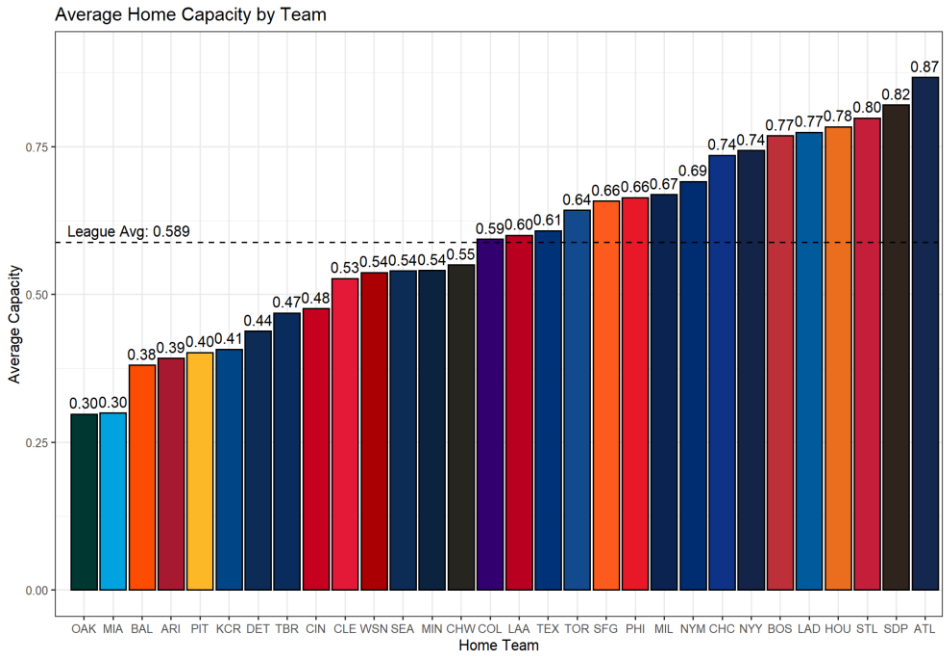
As we can see, decreasing the bin width from .05 to .01 gives a more accurate representation of the shape of the data. There are a high number of sellouts in this dataset, but, based on the shape of the overlayed density curve, we see this does not affect the distribution to a high degree. Before we make estimates of attendance effects using pre-model calculations and visualization, we will look at how the capacity variable behaves when separated by team.



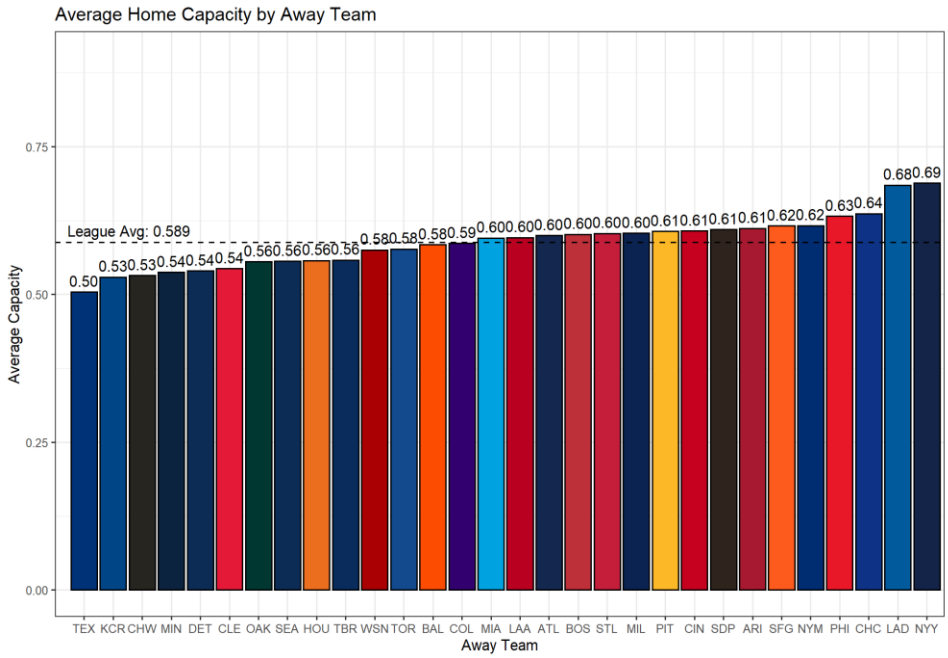
The above visualization shows that the capacity variable has a different distribution for every team, showing the importance that market and performance play on attendance. Teams in large markets, such as the Boston Red Sox, New York Yankees, and Philadelphia Phillies have a left-skewed distribution, while smaller markets like Oakland and Tampa Bay have right-skewness. This pattern is also shown when looking at performance. Successful franchises over the data range are left-skewed (Atlanta Braves and Houston Astros), and perennial losing teams tend to have right-skewed data (Pittsburgh Pirates and Kansas City Royals). From this, we can predict that team performance and market size will have a significant impact on attendance. This should be reflected in the team-based model, with market size being represented by team name, and performance being represented by the home_wins and home_losses variables.

Pre-Model Visualizations

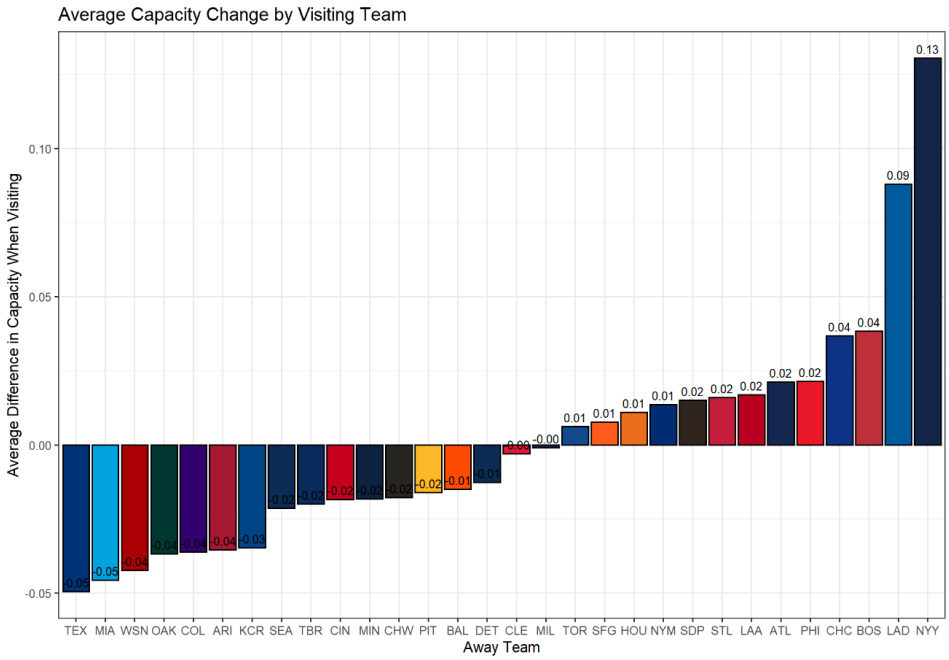
Using simple math and grouping, some estimates of the results of the model could be calculated. My approach for this was to see how both teams and pitchers affect attendance compared to the average for that team. All the pre-model visualizations were done using data from the entire dataset. First, the average home capacity for each team was calculated and is shown below.



As predicted, winning teams in large markets had better average attendance than losing teams in small markets. Oakland and Miami fail to bring fans into the ballpark, while young powerhouses like the Padres and Braves excel in attendance on a game-by-game basis. This could also be predicted by the general shape of the histograms by team. One surprising observation is that both the Red Sox and Yankees are outside the top five in average attendance. These teams have been historically significant in the league, and their large markets provide a perfect opportunity for consistent sellouts. Next, to see which teams draw larger crowds when they are traveling, I calculated the average attendance grouped by the visiting team.

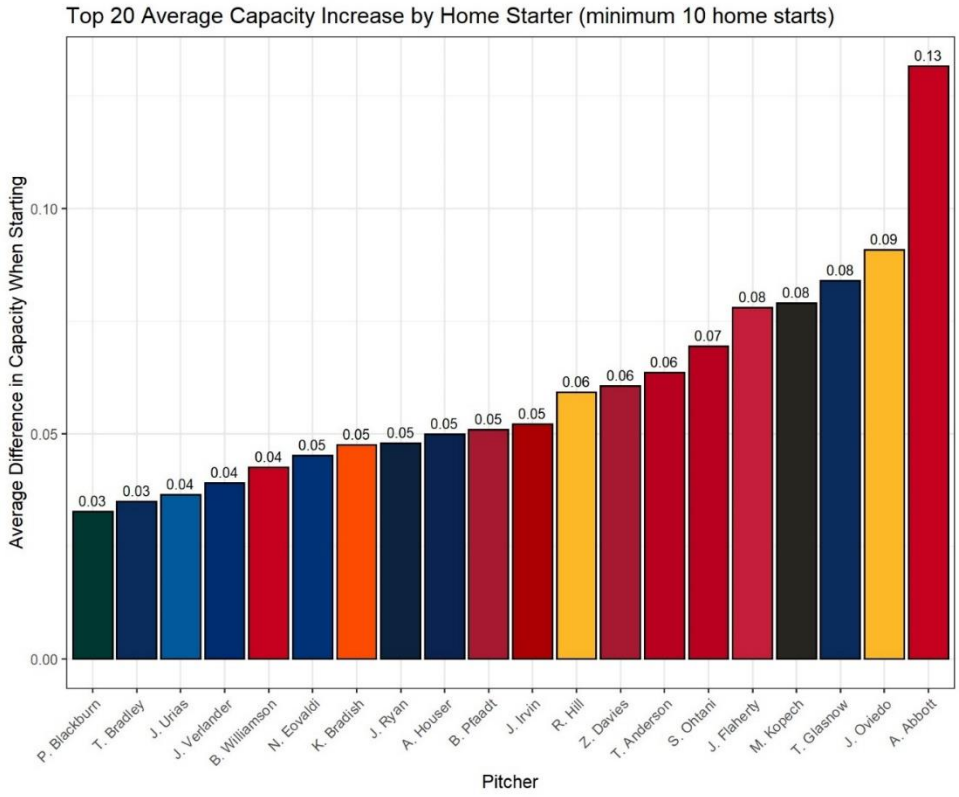


This distribution is more uniform, as every team will play away games in both successful and unsuccessful markets. However, we can still see that when the Yankees and Dodgers come to town, more people seem to go to the games. The next visualization will better show the estimated effect that visiting teams have on attendance. This number was calculated by taking the difference between the capacity of the game and the average attendance number for the home team. Then, grouping by the away team, the average difference was taken.

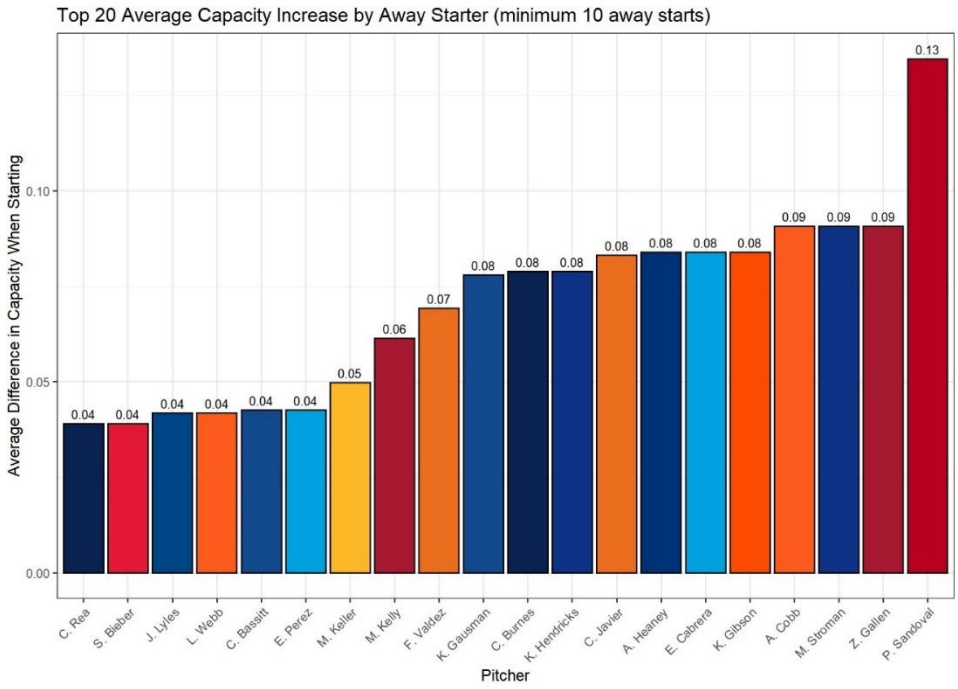


Based on this, the Yankees and Dodgers are fan-favorite teams. When the Yankees visit, stadiums experience a 13 percent increase in attendance, and the Dodgers bring with them a 9 percent boost, on average. This is likely due to these teams becoming global brands and having a steady fan base across the country. Based on these visualizations, we can predict that both home and away teams affect attendance, and the more successful a team is, the more likely it will be to draw large crowds.

Moving away from team effects, starting pitching should have an additional effect on attendance. Other than quality pitching leading to more wins, and, in turn, bringing more fans to the park, high-profile names should bring with them an added boost in attendance. On a normal basis, starting pitchers for a game can be predicted ahead of time, and rotation orders rarely change when compared to batting lineups. Because of this, an attendance boost should be expected when star-caliber pitchers are set to start.



The above chart shows the average difference between the observed capacity for a game and the average capacity for that home team, grouped by starting pitcher. A minimum of ten starts was added to get a decent sample size for each pitcher and to prevent outlier games (Opening Day, holidays, promotions, etc.), from affecting the results. Interestingly, Andrew Abbott of the Cincinnati Reds had the highest average attendance difference after his surprising debut season in 2023. Going down the list, we see big-name pitchers, such as Tyler Glasnow, Jack Flaherty, Justin Verlander, and, of course, Shohei Ohtani.



Focusing now on the effect on attendance the away team's starting pitcher can have, we see similar results. All-Star talent seems to bring more fans than expected, with Shane Bieber, Corbin Burnes, and Zac Gallen present on this list. The top pitcher, Patrick Sandoval, however, performed below average during the range of the data compared to others in the chart. His presence is likely due to fans wanting to see Ohtani and Trout when the Angels visit their hometowns. In the models, it will be interesting to see if the two stars will have the same effect.

Models

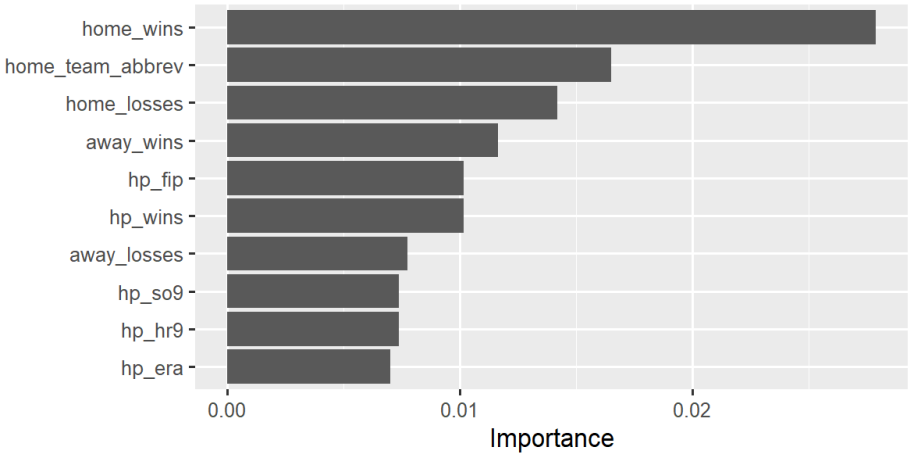
The goal of the model is to measure the externalities that starting pitchers bring and how these "superstars" affect attendance by following a similar approach to Humphreys and Johnson's study of the increases in attendance in the NBA. In their models, they determined which players were deemed superstars by ranking how many times they appeared in the top five of yearly salaries. Due to baseball's unique salary structure and differences in payroll, this strategy would not be as effective in finding the superstars on the mound. For this model, top pitchers were determined by the number of starts that they had in the dataset. If they appeared in

the top 50 on the list, they were deemed superstars. Although this ignores all pitching statistics that would determine the quality of the player, staying healthy as a pitcher is important for many clubs.

Once non-superstar pitchers were filtered out (changed to ‘other’ within the model), the model could be built. Because of the number of categorical and differing behaviors of the data, a random forest model to predict attendance would be the best approach. The dataset was split into train and test datasets, with the 2021 and 2022 seasons being used to train the model. With the 2023 season as the test set, we can see which pitchers brought above-expected numbers during the season using fitted values.

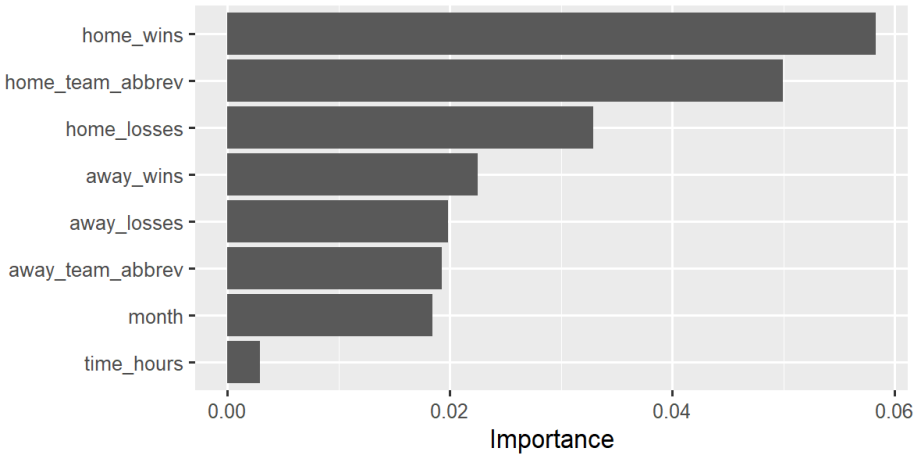
The models were built using a ranger random forest model. The hyperparameters (mtry, min.node.size, and splitrule) were tuned using a grid search approach to reduce the mean-square error (MSE). Variable importance was determined using permutation, and cross-validation was performed ten times for each model.

The first model that was run used both the team-centric and player-centric variables found in the dataset. This model returned an MSE of 0.026 and an R-squared value of 0.592. Below is a plot of the variable importance of this model (ten highest values).



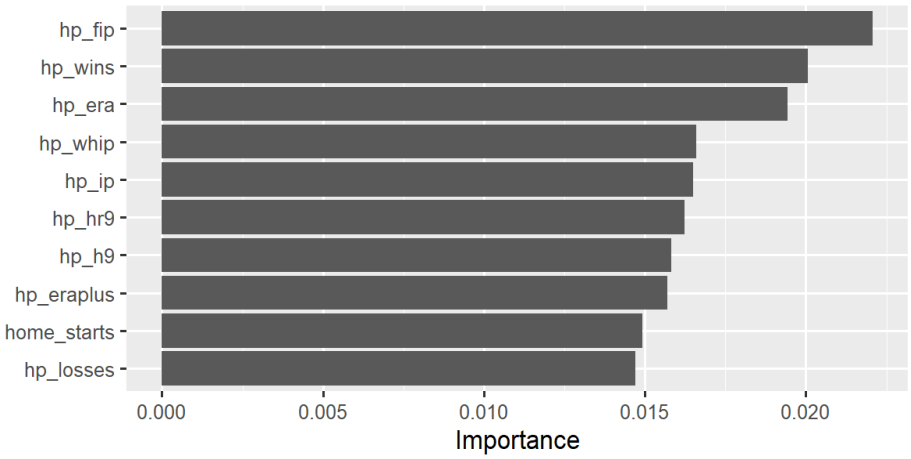
In this model, the team factors seemed to take precedence over pitcher factors when the model was built. Home performance (home_wins) and the team itself (home_team_abbrev) are important drivers in attendance. Surprisingly, the away team did not have a large effect in this model, even though we saw increased attendance when the Yankees and Dodgers visited. To get a better understanding of attendance effects, the above model was split into two, one with team factors and one with pitcher factors.

The team model resulted in an RMSE of 0.015 and an R-squared of 0.76. The higher performance is likely due to the reduction in the number of variables, and, in turn, the minimal noise in the model. The following is the variable importance for this model.



Like the broad model, home team success is the most important in building trees. Additionally, the home team is of high importance, showing the importance of market size on attendance. The time of the first pitch had little to no importance in this model, most likely because most games start around the same time, meaning there is less variation of this variable.

Finally, seeing the effects pitchers have on attendance, the player factors were isolated for a third model. This model returned an MSE of 0.035 and an R-squared of 0.46. The expectations were that the pitcher name variable, both home and away, would be of high importance.



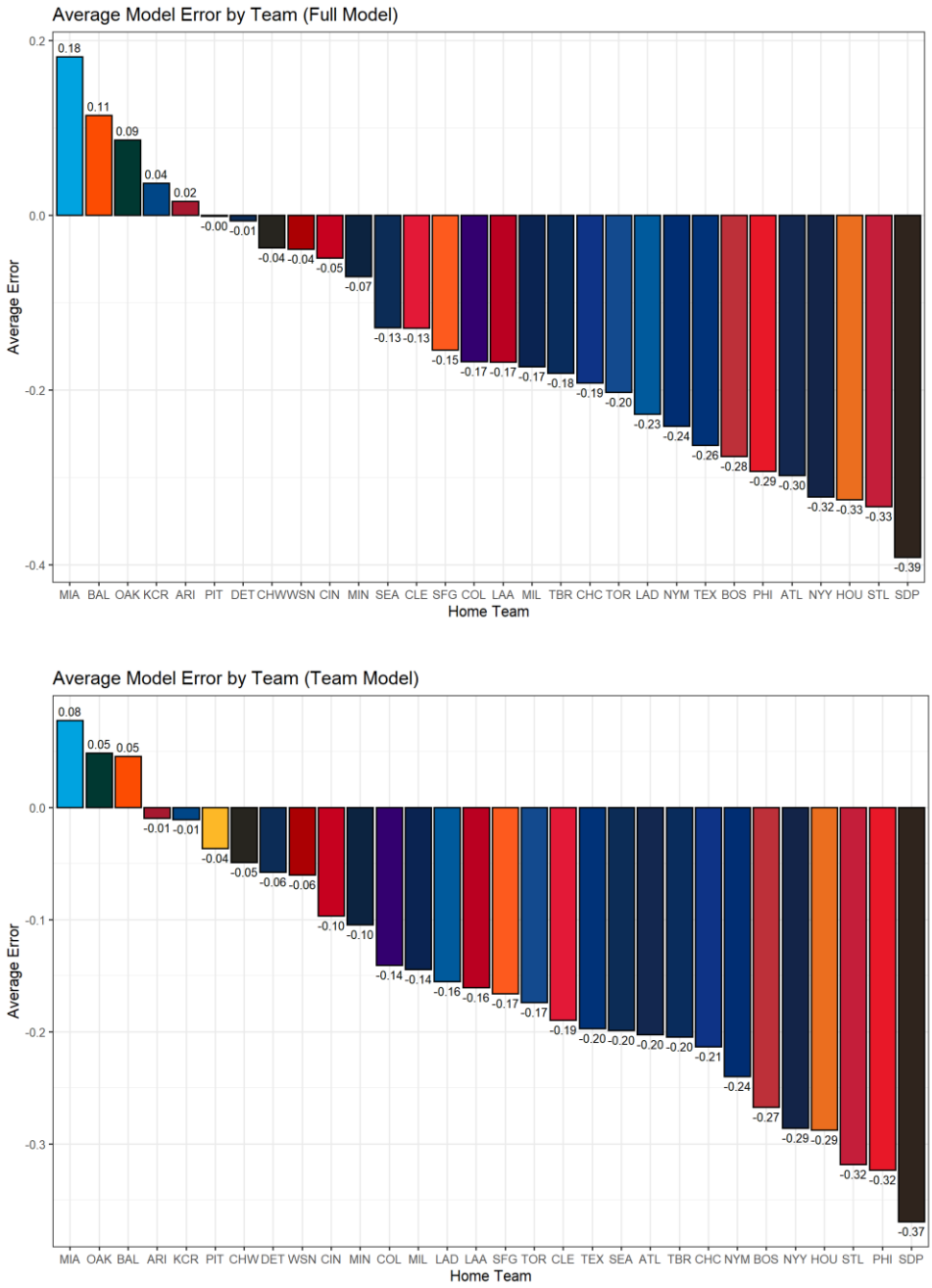
Based on the variable importance plot, the pitcher did not appear in the top ten in importance. Further analysis showed that the home pitcher had a slightly higher importance level than the away pitcher, but the low value is still concerning. This does not mean the pitcher does not have any effect on attendance. We can still use fitted values in the test set to see if there are any patterns in the data.

Looking at the plot, the home pitcher's season performance did affect attendance. Specifically, FIP, which is a good indicator of the true quality of a pitcher, and ERA, a commonly used performance metric are important factors. This pattern is most likely due to better pitching, leading to more wins, which, from what we have seen, is the most important variable according to variable importance values.

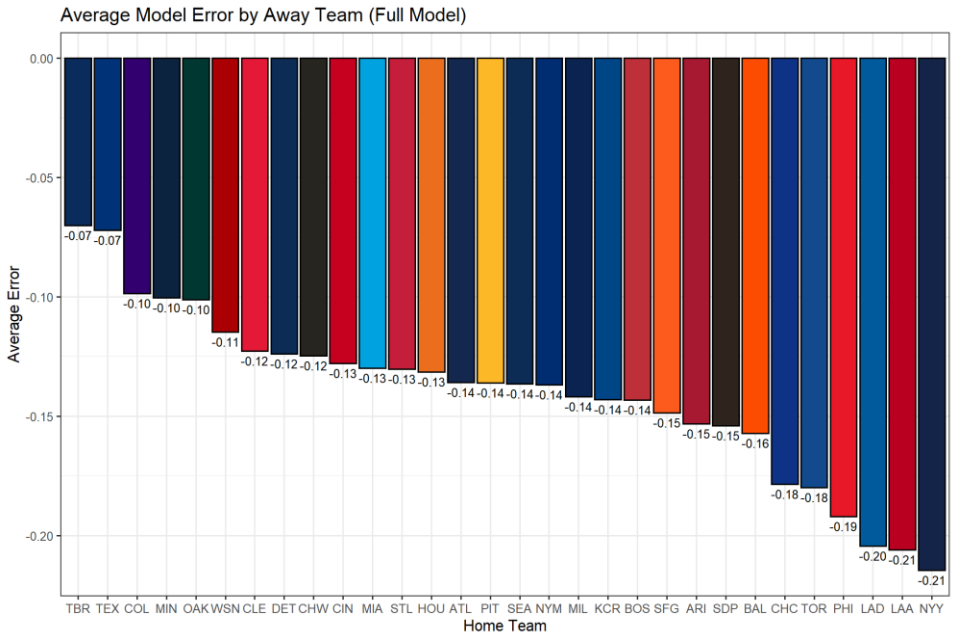
Post-Model Visualizations

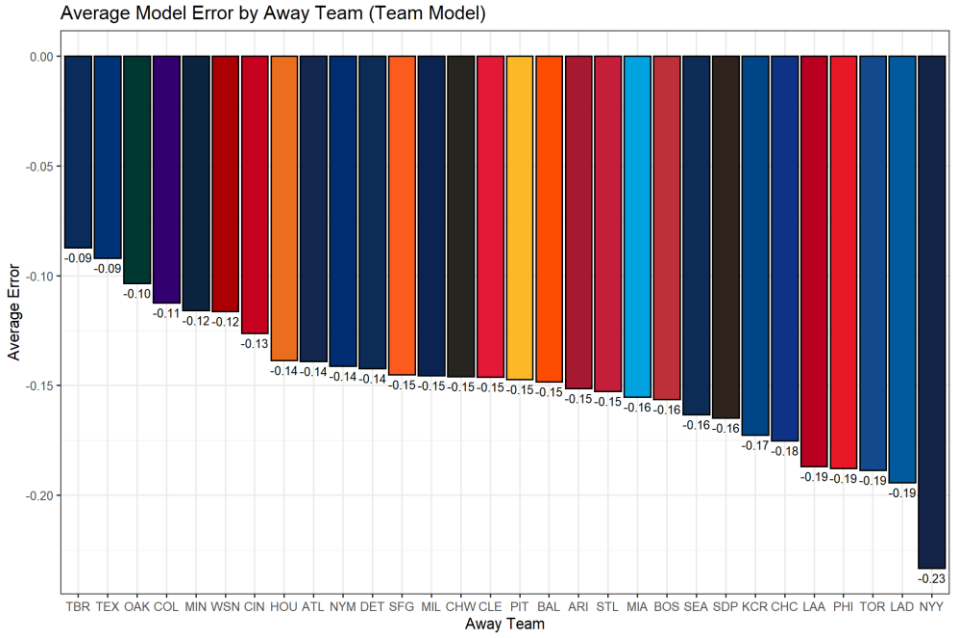
After applying the built random forest models to the test data set, fitted values can be calculated. With these, the accuracy of the model can be judged. To do this, the pre-model visualizations will be recreated, this time using average model error instead of difference from the mean. All three models will be tested using these visualizations.

The following is the average error by home team using the full model and the team factors model.



The results between the two models are similar. In general, the models underestimate capacity, with all but five teams in the full model and three in the team model being overestimated. Although the model's accuracy seems to be off, there is still a pattern. Small market teams (Oakland and Miami) still underperformed in attendance.

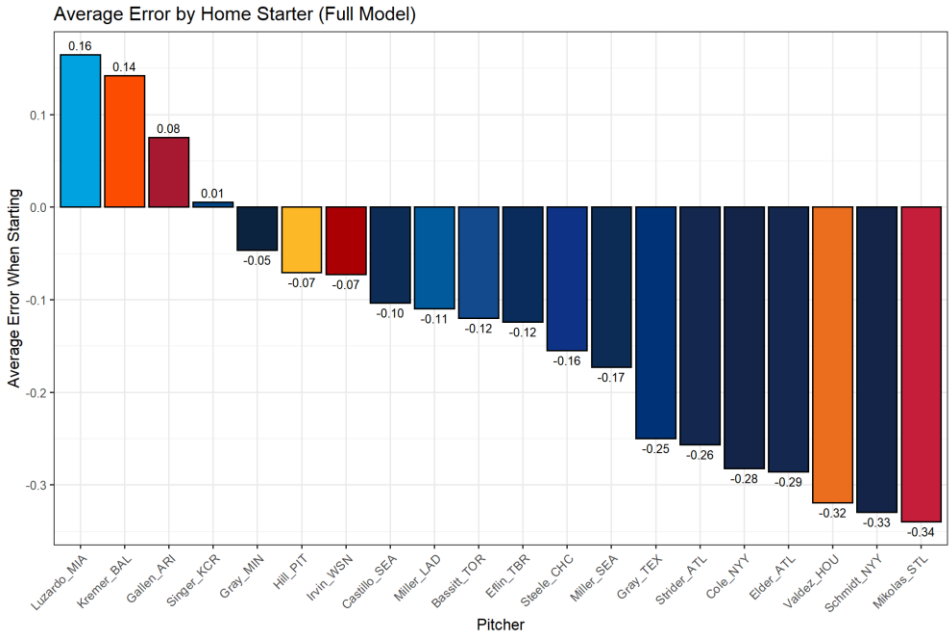


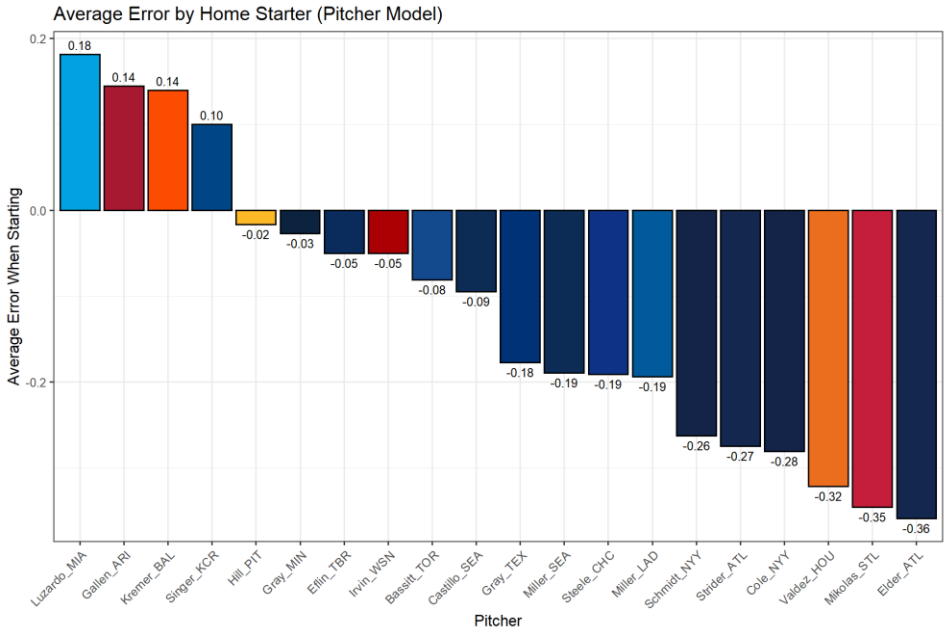


Now, looking at the predictions for the away team, we see the same patterns we saw in the pre-model visualizations. When the Yankees and Dodgers visit, the model underestimates the true attendance at those games. The historically successful teams that have grown to become global brands are often underestimated the most. One exception to this is the Los Angeles Angels, who have had lesser success in the past than other teams whose attendance was

underestimated at their level when they visited. This shows the boost that Mike Trout and Shohei Ohtani bring to the team. These patterns are positive signs that the model has some predictive capabilities.

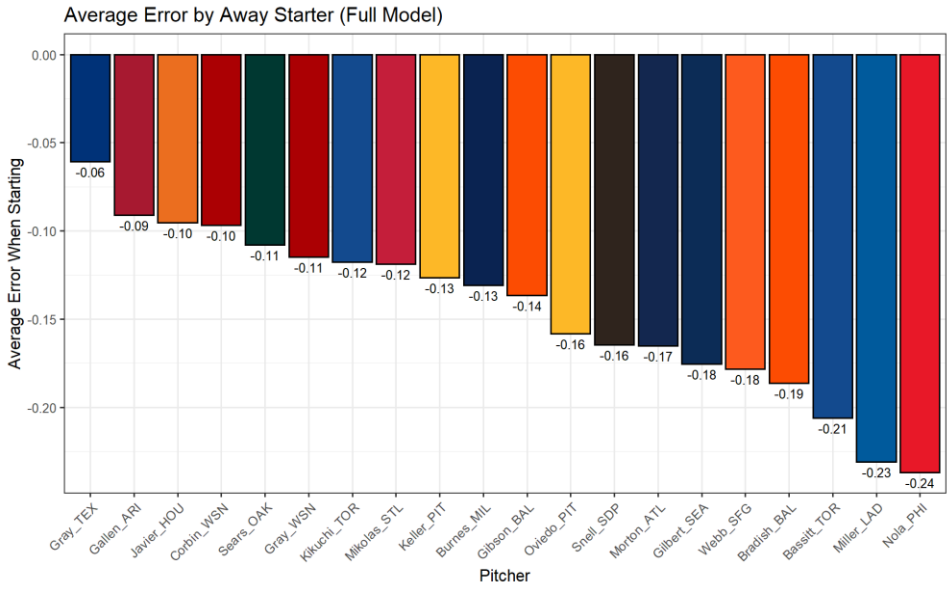
Moving on to pitcher effects, the same process was followed. For the following, the full model is shown first, and pitcher effects second.

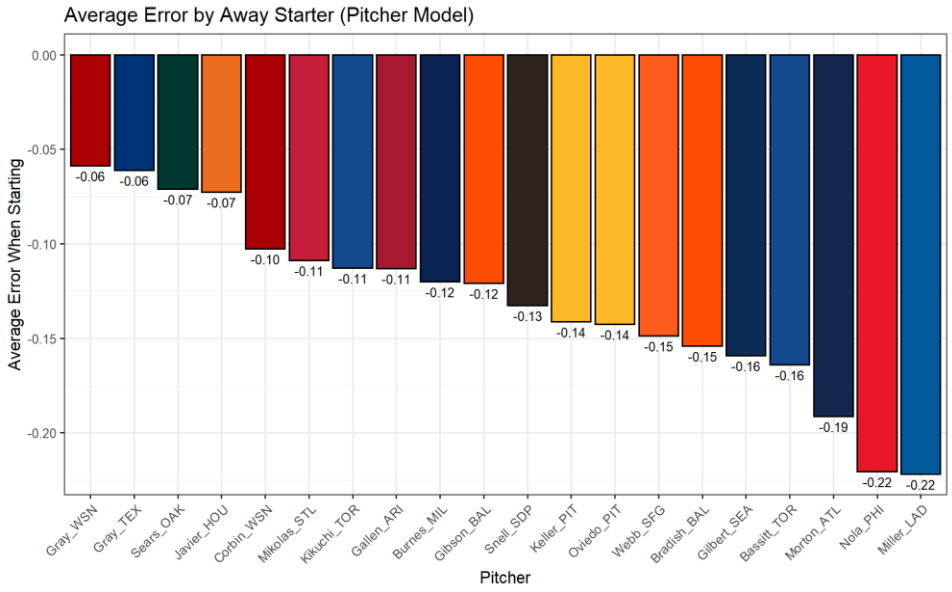




Even in the pitcher effect model, we see the patterns found in the team effects model. The model overpredicts attendance for pitchers in small markets. This could show that the effect of the named starting pitcher is minimal compared to the market size and success of the team.

Finally, the average error for the away starting pitcher is displayed below, again with the full model first.





From these two visualizations, along with the visualizations representing average error by away teams, our model underestimates when the variable pertains to the visiting team. This could be a sign that the model is less accurate with away teams, following the findings of the variable importance charts, which showed away_team_abbrev had much less power in the model building than home_team_abbrev. Nonetheless, the away starting pitcher errors seem to follow the same patterns as home starting pitcher errors, with high-level talent and pitchers on big market teams being the most underestimated.

Conclusion

Using game-level data from the 2021-2023 MLB seasons, random forest modeling was conducted in an attempt to quantify the effects that starting pitchers have on attendance for a given day. Game information, such as the time of the first pitch, the date the game was played, and the two teams playing in the game, were used to build a team-based model, which had an R-squared value of 0.76. Pitchers were analyzed in a separate model, using season-level statistics, such as ERA, FIP, and number of starts, which resulted in an R-squared value of 0.46. A third model was run, using the two sets of variables to create an overarching predictive model, with an R-squared of 0.592.

After running the models, many of the results that were found were consistent with the prediction made using the pre-model visualizations. On a team level, recently successful teams drew larger crowds both at home and when they visited, with historically significant teams having a larger attendance change when visiting. In the pre-model visualizations, this was represented by a high average attendance, and a high level of change in attendance when visiting, and it is shown in the models by high levels of underestimation.

When focusing on the results of the pitcher models, we also see consistency in predictions that were made before running the models. For both the home and away starting pitchers, it seems that fans prefer to see talented pitchers, with a majority of the names in the visualizations being at or near an all-star level. One precaution we must take before estimating the true attendance effects of these pitchers is the variable importance discussed before modeling. In the full model, team-based variables seemed to take precedence in predictive power. This is why many of the pitchers whose attendance was often underestimated are also on more successful teams over the past three seasons. Focusing more on the results of the pitcher model provides much more information in finding which pitchers bring the largest superstar effects.

Discussion

This research has many uses, as it can be utilized by front-office personnel to justify signing a big-name free agent, by box offices to help them estimate a fair ticket price for a game, or by fan experience personnel to know when to expect large crowds. Additionally, the strategies used in creating these models can be used in all sports. If a model can be created to accurately predict the expected attendance of a given game in that sport using broad, team-level data, then quantifying star power effects of a given position (quarterbacks, goalies, etc.) comes down to finding statistics that have high predictive abilities. It will most likely be found if these strategies were to be applied to other sports, that basketball and football stars will have greater influence on attendance, because of the size of the rosters and reach of the sport, respectively.

Although the overall accuracy of the models was below expected, the patterns seen within them are indicative of their potential for successful prediction. Extending the data to

include seasons before the COVID-19 pandemic may lead to greater accuracy, as well as including different pitching statistics, such as changing the pitcher statistics to be their results up to that point in the season. Despite this, the models do show promise, and, with some adjustments, could become accurate predictors of attendance and quantifiers of pitcher attendance effects in Major League Baseball.

References

Barilla, A., Gruben, K. H., & Levernier, W. B. (2008). *The effect of promotions on attendance at Major League Baseball games*. Digital Commons@Georgia Southern. https://digitalcommons.georgiasouthern.edu/marketing-facpubs/74/#:~:text=The%20findings%20of%20this%20study,type%20of%20promotion%20is%20important.

Cohen, G. (2015, January 26). *The effect of attendance on team spending*. Georgetown Sports Analysis. https://georgetownsportsanalysis.wordpress.com/2015/01/26/the-effect-of-attendance-on-team-spending/

Drayer, J., Shapiro, S., & Lee, S. (2012). *Dynamic Ticket Pricing in Sport: An Agenda for Research and Practice*. Old Dominion University. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1032&context=hms_fac_pubs

Fishman, P. (2002). *Competitive balance and free agency in Major League Baseball*. duke.edu. https://sites.duke.edu/djepapers/files/2016/08/fishman.pdf

Hall, S., Szymanski, S., & Zimbalist, A. (2002). *Testing causality between Team Performance and payroll the cases of ...* ResearchGate. https://www.researchgate.net/publication/227351010_Testing_Causality_Between_Team_Performance_and_Payroll_The_Cases_of_Major_League_Baseball_and_English_Soccer

Harvey, C. (2020, March 13). *Measuring true fan engagement and the commercial value of sports*. LinkedIn. https://www.linkedin.com/pulse/measuring-true-fan-engagement-commercial-value-sports-chris-harvey/

Hepper, T. (2017, June 1). *Predicting MLB game attendance*. Medium. https://towardsdatascience.com/predicting-mlb-game-attendance-c36cdc1b8de6

Humphreys, B., & Johnson, C. (2019). *The effect of superstars on game attendance: Evidence from the NBA* ... Sage Journals. https://journals.sagepub.com/doi/abs/10.1177/1527002519885441?journalCode=jsea

Josselyn, C. (2019). *The Lebron Effect: Is a Superstar Worth the Money*. Bridgewater State University. https://vc.bridgew.edu/cgi/viewcontent.cgi?article=1379&context=honors_proj

Langhorst, B. (2014, April 2). *What Do Your Fans Want? Attendance Correlations with Performance, Ticket Prices, and Payroll Factors*. Society for American Baseball Research. https://sabr.org/journal/article/what-do-your-fans-want-attendance-correlations-with-performance-ticket-prices-and-payroll-factors/

Lee, Y. (n.d.). *Common factors in Major League Baseball Game Attendance*. Sage Journals. https://journals.sagepub.com/doi/abs/10.1177/1527002516672061?journalCode=jsea

Lewis, M., & Yoon, Y. (2021, October 21). *An empirical examination of the development and impact of Star Power in Major League Baseball*. Journal of Sports Economics. https://www.academia.edu/59442931/An_Empirical_Examination_of_the_Development_and_Impact_of_Star_Power_in_Major_League_Baseball

Lindholm, S. (2014, February 25). *Pitcher aging curves*. Beyond the Box Score. https://www.beyondtheboxscore.com/2014/2/25/5437902/pitching-aging-curves

Nelson, S., & Dennis, S. (2015). *Performance or Profit: A Dilemma for Major League Baseball*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2627705

Näf, J. (2023, November 3). *Variable importance in random forests*. Medium. https://towardsdatascience.com/variable-importance-in-random-forests-20c6690e44e0

Ormiston, R. (1970, January 1). *Attendance effects of star pitchers in Major League Baseball*. Journal of Sports Economics. https://ideas.repec.org/a/sae/jospec/v15y2014i4p338-364.html

Ormiston, R. (2016). "Hype and Hope: The Effect of Rookies and Top Prospects on MLB Attendance – Society for American Baseball Research" Society for American Baseball Research, 2016, April 20, https://sabr.org/journal/article/hype-and-hope-the-effect-of-rookies-and-top-prospects-on-mlb-attendance/.

Ormiston, R. (2014, April 2). *Do Fans Prefer Homegrown Players? An Analysis of MLB Attendance, 1976–2012*. Society for American Baseball Research. https://sabr.org/journal/article/do-fans-prefer-homegrown-players-an-analysis-of-mlb-attendance-1976-2012/

Petti, B. (2012, May 2). *Pitcher aging curves: Starters and relievers*. FanGraphs Baseball. https://blogs.fangraphs.com/pitcher-aging-curves-starters-and-relievers/

Rascher, D. (2007). *Variable ticket pricing in Major League Baseball - University of San ...* University of San Francisco. https://repository.usfca.edu/cgi/viewcontent.cgi?article=1008&context=ess

Reifer, D. (2020, May 24). *Modeling the Mets' attendance*. Corner Three. https://cornerthree.net/2020/05/24/modeling-the-mets-attendance/

Rivers, D. H., & DeSchriver, T. D. (2015). *Star Players, Payroll Distribution, and Major League Baseball Attendance*. Fit Publishing. https://fitpublishing.com/content/star-players-payroll-distribution-and-major-league-baseball-attendance

Russo, C. (2019, July 2). *Ballpark attendance and starting pitchers*. Fangraphs Community Blog. https://community.fangraphs.com/ballpark-attendance-and-starting-pitchers/

Shapiro, S., Drayer, J., & Dwyer, B. (2016). *Examining consumer perceptions of demand-based ticket pricing in Sport*. Old Dominion University. https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1023&context=hms_fac_pubs

Slusser, A. (2021). *The Star Player Effect: Does the Quality of a Player Impact Attendance in the NBA?*. Cardinal Scholar. https://cardinalscholar.bsu.edu/server/api/core/bitstreams/b69170f7-9b62-4ef8-bf8d-084296cc3cc9/content

Thauwald, S. (n.d.). *The determinants of Major League Baseball attendance*. Colorado College. https://digitalccbeta.coloradocollege.edu/pid/coccc:2895/datastream/OBJ

Wakefield, K. (1995). *The pervasive effects of social influence on sporting event attendance*. ResearchGate. https://www.researchgate.net/publication/249675800_The_pervasive_effects_of_social_influence_on_sporting_event_attendance

Woltring, M. (2014). *Attendance Still Matters in MLB: The Relationship with Winning Percentage*. The Sport Journal. https://thesportjournal.org/article/attendance-still-matters-in-mlb-the-relationship-with-winning-percentage/