An All-Star Among All-Stars

By Benjamin Kahn

ABSTRACT

My study aimed to discover which Major League Baseball (MLB) players could be considered All-Stars among All-Stars based on their hitting, fielding, and pitching skills as taken over both their entire All-Star season as well as their performance in just their All-Star game that season. Using the Elbow Method, k-Means clustering, Principal Component Analysis (PCA), Hierarchical Clustering, Standardized Scaling, and Scatter Matrices, I was able to come up with a small list of All-Stars in each of the three skill categories who landed within the top 5% of all All-Stars in their respective skills. While that list might not include every one of the greatest players to ever play the game of baseball, it does include some of the players who have performed the best in All-Star Games and as a result, can be considered All-Stars among All-Stars.

INTRODUCTION/PROJECT OVERVIEW AND MOTIVATION

When looking through FiveThirtyEight and Kaggle for potential data sources to use, I knew I wanted to choose something related to sports. For as long as I can remember, I have always had a good mind for math. This natural ability turned into a love when it was combined with sports, one of my other favorite hobbies. Looking at sports through statistics (especially those of an advanced nature) that most fans would ignore makes me feel as if I can understand more about the game being played than they do. It is for that reason that I took so many advanced math classes in high school. It is for that reason that I am at Chapman University, majoring in Data Analytics, when I could have just as easily gone to any other school for math, computer science, or engineering. And it is for that reason that my ultimate career goal in life is to work as an analyst for a professional sports team, be it as a talent scout or as a data scientist providing information, creating databases, and collaborating with those talent scouts.

Now that I had a theme in mind, I had to choose what data source I wanted to use. After taking a quick look at Kaggle and what it had to offer I almost immediately decided to not use any of its offerings. While it clearly had a much wider variety than the FiveThirtyEight GitHub repository, the data sources were a lot messier and also had the potential of being non-reputable.

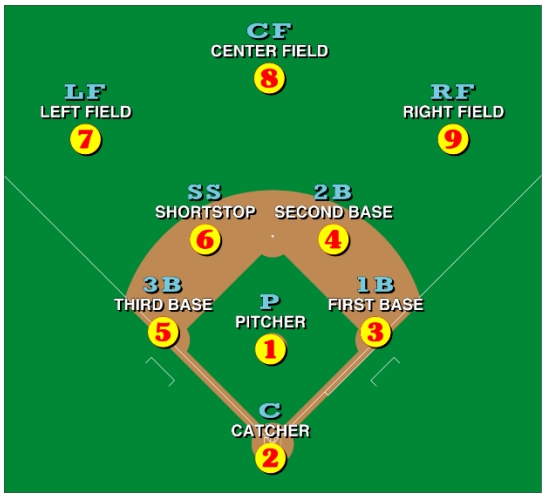
My decision was made easier with the knowledge that FiveThirtyEight had done several articles on sports. While I was disappointed that my first choice of sport, ice hockey, had no representation in the repository, I was happy to select a data source concerning my second favorite sport, baseball.

The database I chose (https://github.com/fivethirtyeight/data/tree/master/mlb-allstar-teams) was used in a FiveThirtyEight story entitled *The Best MLB All-Star Teams Ever*. It was used to determine which year's All-Star team in each league of the MLB would fare the best against every other year's All-Star teams representing each league. Considering that this data set was curated for this specific topic, I had to come up with another way to utilize the data contained within it that wouldn't just revolve around sorting the data points to find what the largest values were.

My project uses the file *allstar_player_talent.csv* contained within that data set in order to determine what makes a player an All-Star among All-Stars. That is, I wanted to figure out who, in a league of players who made All-Star games, would be All-Stars or even would win seasonal trophies such as the Silver Slugger Award (best batter at each position in each league), the Gold Glove Award (best fielder at each position in each league), or the Cy Young Award (best pitcher at each position in each league). I chose to do this project idea because it seemed like something that could feasibly be done with the data processing tools that I have gained over the course of this class, the fact that it seemed to fit in well with the data contained in the set, the topic being something that I was interested in and would enjoy working on, and the fact that doing analyses on topics such as these could mirror analyses I would be expected to be able to perform if I were to get a job in my dream field. I chose to disregard the file *all_star_team_talent.csv* that accompanied the file I chose because it did not offer me anything extra that I could use for this project.

DATA

Of the fifteen variables that were used by FiveThirtyEight in their original article, I used only nine. OFF600 refers to the amount of runs that the player would provide their team singlehandedly compared to an average player if given 600 at-bats (hitting attempts). DEF600 is the same as OFF600 except for the fact that it refers to the number of runs the player would have saved their team singlehandedly compared to the average player. PITCH200, similarly to



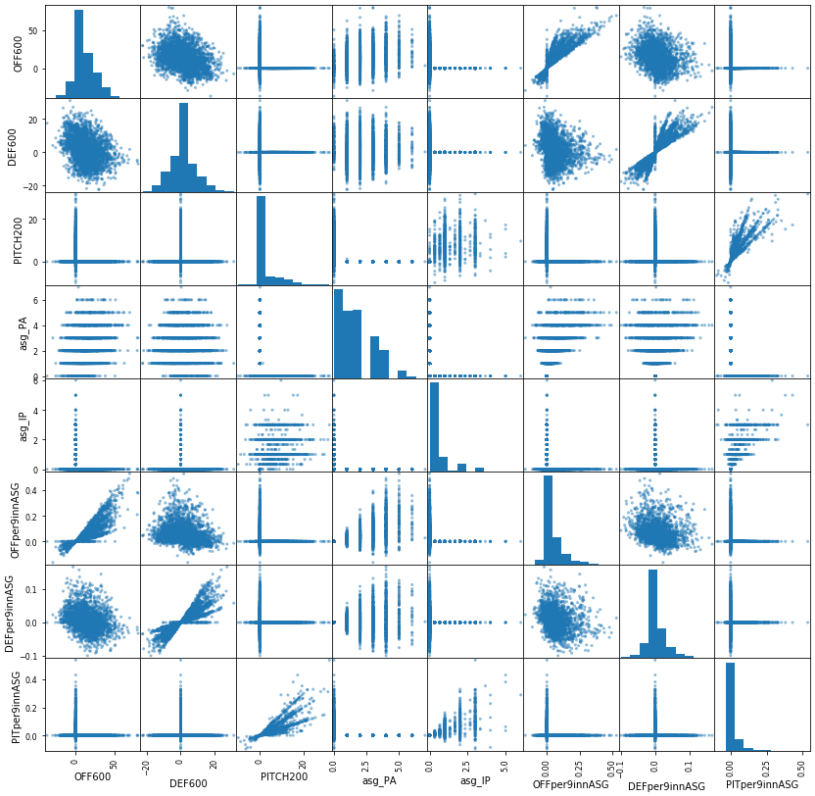
DEF600, details the number of runs a pitcher would have saved their team in 200 pitched innings when compared to the average pitcher. These three statistics are based on how well the player performed during the regular season in the specific year they made an All-Star game. OFFper9innASG, DEFper9innASG, and PITper9innASG match their full season counterparts described earlier except that they only cover a player's run producing or run saving potential over the course of one nine inning All-Star Game as opposed to a full season. These are both important because the season-long statistics measures a player's ability to produce consistently whereas the All-Star Game-specific statistics measure a player's ability when put up against other players of a similar talent level (i.e., All-Stars). The other variable that impacted my methodology was startingPos. This refers to the player's position on the field as designated by their number 1-9 (see graphic). This database also included numbers 0 and 10, which both referred to the Designated Hitter (DH), and a null value to signal that the player was not a starter in that specific All-Star Game. The final two variables I used did not impact my methodology but did help inform my results and conclusions. The variables bbref_ID and yearID helped me determine who the player was and what year's All-Star game their stats were taken from so that I could check to see if my models were outputting as I wanted them to and if the player's season truly was elite.

METHODOLOGY

The first thing I did (as with any data set analysis) was to check the data to ensure that there were no missing or incorrect inputs and to clean it up. Every variable seemed to be correct except for startingPos. When I looked deeper into it, I found that over half of the players had a null value for their position. After doing some research on the FiveThirtyEight GitHub repository that hosted this data set, I discovered that players with a null startingPos value were bench players that were brought on later in the game and did not start. This initially concerned me as I was wondering how I was going to be able to get my database to recognize bench pitchers against bench hitters, but this fear was quelled when I realized that pitchers all had 0 as their score for OFF600, DEF600, OFFper9innASG, and DEFper9innASG, and that batters all had 0 as

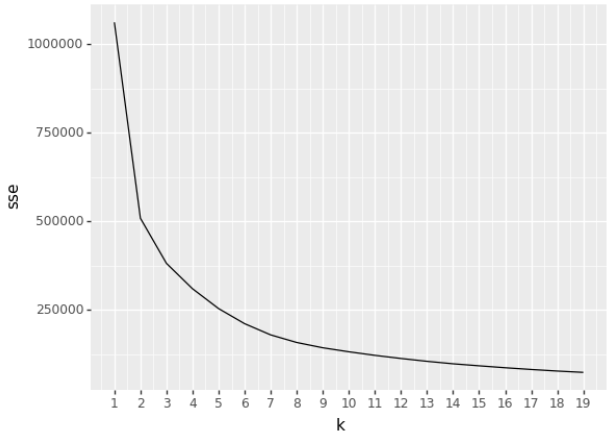
their score for PITCH200 and PITper9innASG. It was during this time that I also discovered that startingPos values of 0 and 10, which are not standard player positions, both referred to designated hitters (players who only hit, not field) and converted all 0 startingPos values to 10 to combine them.

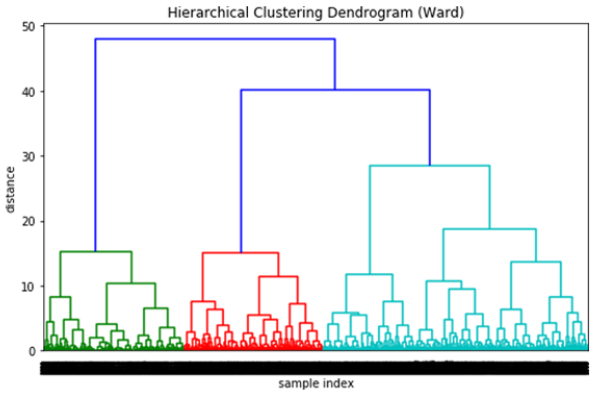
After determining that the data set was in a state that would work best for what I wanted to do, I began to work on it. The first thing I decided to do was to take all of the numerical variables in the dataset (OFF600, DEF600, PITCH200, OFFper9innASG, DEFper9innASG,



PITper9innASG, and two other variables that I eventually discarded) and to put them in a separate data frame in order to make analysis easier. I had determined that the categorical variables in the dataset were most likely not going to affect any conclusion that I could come to. The first thing I did with this new data frame was to create a scatter matrix to determine any possible correlations between offense, defense, or pitching and any of the other variables. Unsurprisingly, I found that an increase in a player's production in a certain area (hitting, fielding, or pitching) over the course of the season generally also led to an increase in their production in an All-Star Game.

The next thing that I decided to do was to see how many different clusters or categories

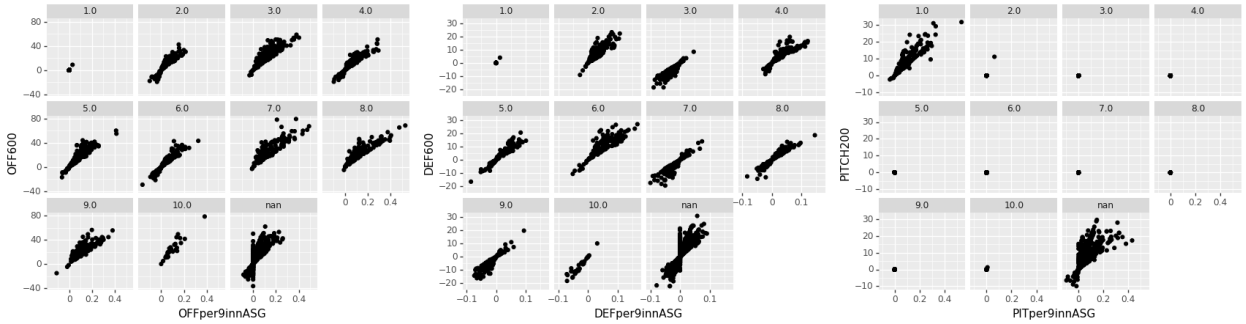




these players could be joined into. I had a hunch that the answer would be three due to the three categories that have already been discussed at length being so prevalent and unique. However, I also knew that hunches can be proven very easily wrong with analytics. I first performed the k-Means Elbow Method on the data into order to hopefully determine how many clusters lay within. The graph suggested that there could be anywhere from 3 to 5 clusters in the data depending on how it was read. This answer left me unsatisfied and wanting a clearer answer. Continuing, I decided to perform Hierarchical Clustering on the same data frame of numerical variables to get a more concrete answer. The first task necessary was to scale the values to make the data usable by PCA. After using standardized scaling on the values and performing a PCA on the variables, I was finally able to create a Hierarchical Clustering Dendrogram. Thankfully, the dendrogram agreed with me that the data should be separated into three clusters. However, I was a bit puzzled as to why one cluster held half of the data while the other two clusters only held a quarter of the data each. Thinking it through brought me to the theory that half of the data is for pitchers and the other half is for batters with one quarter representing hitting and the other quarter representing fielding. This conclusion gave me even more hope that my plan for separating the data into those three categories (hitting, fielding, and pitching) was the correct plan of action.

Out of curiosity as well as wanting to not rush into anything, I decided to chart every player's categorical scores both over the entire season and just in the All-Star Game separated by their position to see if there was anything I needed to take into account while moving forward. While I found nothing that affected the way that I continued to use the dataset, a few interesting

points could be made. The first is to notice that positions 2 (catcher) and 6 (shortstop) tended to have lower offensive values than the other batting positions. This is because those positions



generally require high defensive skill on the part of the player playing there which often leads to a lack of hitting prowess. This is confirmed in the defensive values chart where those two positions tend to have higher scores than every other position. To that point, positions 3 (first base), and 7, 8, and 9 (outfield positions) seem to have lower defensive scores on average. This is because, traditionally, players at those positions are the best hitters on the team and are put in the positions they are put in because they tend to require less defensive skill than the rest of the infield positions and the catcher.

With confirmation that I had accurately cleaned the data set, that I had the correct number and type of clusters, and that there were no odd positions skewing my data, I could begin performing analyses on the players themselves. The first thing I did was create three new data frames, with one for each category and its specific analytics (season-wide and All-Star Game-specific). This would make sure I could analyze just those categories without any other variables confounding my results.

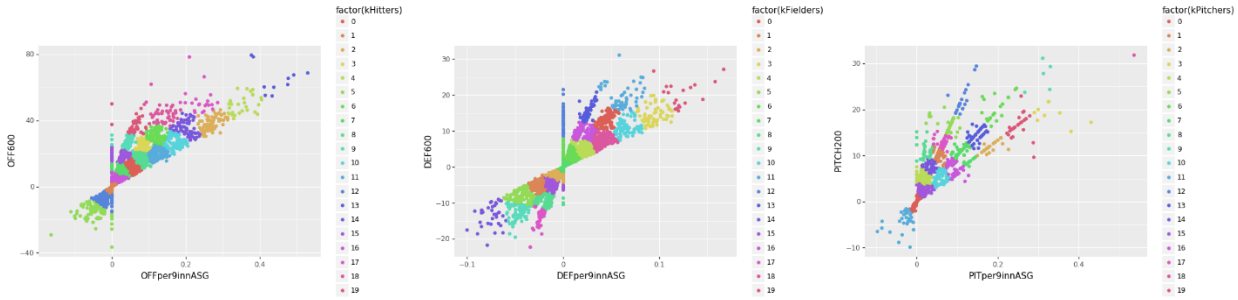
I decided to use k-Means clustering again to determine the top level of players in each category and used hitting skill as a test. I first tried four clusters to put every player into a 25% group. While I was able to prove that my k-Means algorithm worked, there were not enough clusters provided to draw the type of conclusions that I had wished to uncover with my project.

I then tried with twenty clusters to be able to potentially discover the top 5% of players in each category. Luckily, on only my second try, I got a satisfactory number of clusters. This run provided me with three elite hitters (cluster 10), five elite fielders (cluster 17), and six elite fielders (cluster 4). This analysis found that players with an OFF600 level above 75 were considered elite among the elite, that players with a DEF600 level of 25 or greater were All-Stars among All-Stars and that pitchers with PITCH200 ratings higher than 26 were exceptionally gifted. I will go into the specifics of some of these players and whether these findings were

accurate or not in the results section.

As I was finishing up commenting my code, I noticed something about these graphs that might have affected some of my results. The season-long results occurred between -40 and 80

while the All-Star Game-specific results only occurred between -0.2 and 0.6. While my results from the first k-Means analysis seemed accurate, I decided to try again after having used standardized scaling on the data points to be more similar. Several of the players in this analysis were like the first one, but several other players were added, and some were removed. In total, this analysis found ten elite batters (cluster 13), nineteen elite fielders (cluster 19), and six elite pitchers (clusters 18 and 9). These results, when added to the results of the previous k-Means analysis cover many of the greatest hitters, fielders, and pitchers ever to play the sport, which told me that my analysis was at least somewhat accurate. As with the pre-scaled results, I will go into further detail into some of these players and the accuracy of these findings in the results



section.

RESULTS

To assess the accuracy of the predictions that were made in both analyses for every player would be too long for this essay. To show that my predictions were correct and accurate, I will analyze the players who were found in common between both analyses. The hitting analyses found two hitting seasons in common, both from the same player. Barry Bonds in both 2003 and 2004 was found to be elite. Not only has Barry Bonds hit the most home runs of anybody in the history of Major League Baseball (albeit with the help of steroids for a few years), but in those two seasons specifically he won the Silver Slugger Award as the best hitter at his position in the National League and also the MVP Award as the best player in the National League as a whole that season.

While fielding is often overlooked when it comes to both fan excitement and award voters, the Gold Glove Award does recognize the best fielders at each position in each league every year. Neither of the fielding seasons found in both analyses (Bucky Dent in 1980 and

Marty Marion in 1944) won the Gold Glove Award, but both played shortstop, which as described earlier, is one of the two most important and difficult fielding positions in baseball, so for them to receive such high scores at such an intense position is a testament to their skill. In addition, Marty Marion's 1944 season came well before the Gold Glove Award was introduced in 1957.

Like fielders, pitchers are also overlooked when it comes to MVP voting as well as fan hype. The first two of the four seasons found in both analyses came from Lefty Grove in 1936 and 1938. In both seasons, he led the league in ERA (runs given up per 9 inning game) and would have carried his teams practically single-handedly to the World Series had it not been for the juggernaut New York Yankees standing in their way. The Cy Young award, which recognizes the best pitcher in each league yearly, was not introduced until 1956 so there was no way he could have won it. The third season found in common was in 1999, where Pedro Martinez led the league in ERA, wins, and strikeouts and won the Cy Young as well, having received 100% of the first-place votes. In addition, he even came second in MVP voting, only losing it by a small margin, which is incredible for a pitcher. Finally, the last season found in common was Randy Johnson's 2001 campaign. In that season, he led the league in ERA and strikeouts (third most of all-time in a season of anyone in the Modern Era), won the Cy Young, and even won three World Series games for his team, winning a World Series ring and the title of World Series MVP in the process.

Based on all the evidence provided above, I believe that my analyses were very accurate at determining incredible seasons from players based on both their performance in that season and their performance at that year's All-Star Game. While my model is not perfect by any stretch of the imagination (shortcomings discussed in the conclusion), I believe that my analyses have provided quite satisfactory enough answers for which players are All-Stars among All-Stars and what made them so elite.

CONCLUSIONS AND FUTURE APPLICATIONS

One conclusion that I was able to make is that, while looking for an All-Star among All-Stars is a worthy pursuit that every team attempts, it is better to look for All-Stars in general because All-Stars among All-Stars are a once-in-a-generation occurrence. If you focus on only these generational players, you will miss out on several very good All-Star caliber players. In

addition, teams must consider the fact that you cannot build a team around just one player. The team must work well together for any player's true potential to shine. Another thing to consider about the research that I have done is that it does not and can not account for players at such an elite level who are considered five-tool players. A five-tool player is a batter who has an elite talent for hitting for contact, hitting for power, running well, throwing well, and fielding well. Due to the limitations of the variables provided in this dataset, this is impossible for me to predict. Finally, the way I performed this analysis also does not account for either incredible seasons in which the player did not perform well in the All-Star Game that year or All-Star Games in which one player performed incredibly well but had an otherwise pedestrian season by All-Star standards. However, despite all of these caveats, I can safely say that, for what I set out to accomplish in this analysis, I was able to accomplish my goal, especially when you look at how skilled the players that my algorithm recommended in each category were. The predicted players in each category tended to have won not just the award as the best player at their position in their year, but also the MVP award as the best player in the league and possibly even led the lead in certain major categories at their positions or led their team to either a World Series appearance or victory.

As for how this analysis might be used in the future, it is honestly not something that would likely revolutionize the baseball analytics scene as teams have been using analytics like these to inform their personnel decisions for decades. However, this analysis could be used by someone who is trying to make an argument for who is the best player of all time in a certain skill. It could also be used by someone who is trying to decide who their favorite player of all time is, either current or historical, based on how elite of a talent they were/are in both regular season games and All-Star Games.