Aidan Murphy, Luke VanHouten
CSE 163
13 March 2023

Predicting Cy Young Award Votes Based on MLB Pitching Statistics

Summary of research questions

These are the following research questions that we will be following throughout our project:

1. How can we predict how many vote points a top-tier pitcher will receive for the Cy Young Award based on their statistics?
    - We seek to predict the amount of vote points that a highly skilled player (meaning that their pitching statistics are in the top percentiles) will receive for the Cy Young Award, which is given to the pitcher who has the best pitching in either the American or National. Each place in the voting is assigned a given amount of points, such as seven points for first place. The player with the most points wins the award. These statistics that we would use in a machine learning model to answer this question include ERA, WHIP, WAR, SO%, SO, FIP, K/9, BB/9, HR/9, and Wins.
    - We found that Wins Above Replacement (WAR) as well as Wins were the best predictors for the amount of vote points that a top-tier player would receive.
2. How have the development of new statistics and ways of thinking about the sport changed the factors that determine the amount of votes that top-tier players receive for the Cy Young Award?
    - From a layman viewpoint of the award selection, there has been an obvious shift in what statistics are most highlighted for a given pitcher as new statistics are being developed. We seek to find what other older statistics were favored most by voters as opposed to now, along with when the shift occurred from these to more modern statistics such as those mentioned in the first research question. We will use linear regression and analysis of it to answer this question by splitting the data at the year 1990, as this is the beginning of when many new baseball statistics were developed.
    - We found that contrary to the layman viewpoint, there were no real statistical differences in the correlation between the statistics chosen for question one with the vote points given for the Cy Young Award between the datasets split before 1990 and 1990 and after.

3. How can we determine the different pitching statistics thresholds that gain Cy Young Award votes for relief pitchers as opposed to starting pitchers?
    - It is no secret that the vast majority of Cy Young Award vote getters are starting pitchers, as they are much more impactful overall in a game and are more at the forefront of a team. Yet, relief pitchers can get votes too, but the levels of their pitching statistics that impact how many votes they receive (per question one) are surely different as they will only pitch for 1-2 innings at a time instead of 6-7. We seek to identify the statistical differences between top-tier starting and relief pitchers when it comes to Cy Young Award vote-getters.
    - We found that there is a much higher correlation between the statistics outlined in question one with Cy Young Award vote points for starting pitchers than with relief pitchers.

Motivation

Answering the research questions will give us a greater understanding of the trends in Cy Young Award voting. The Cy Young Award is the award for best pitcher in each league of Major League baseball. Pitching in baseball is the aspect of the game that involves delivering the ball to home plate without the batter being able to hit the ball, with the intentions of getting them out to progress the game without the other team scoring any runs. Awards given out in baseball are based on a variety of criteria which have fluctuated throughout the years. Understanding the factors that go into the awards and votes for them and how they have changed over time will give us a better idea of the sport in the past and how it may change. We care about this problem as the Cy Young Award is meant to represent the best pitcher of each season, and in order for the award to be given to its rightful winner we want to see what statistics most contribute to a player winning Cy Young Award vote points.

Dataset

We will be using data from the pybaseball API for baseball statistics, particularly from FanGraphs for sabermetric statistics and the Lahman Baseball Database for awards histories, including the number of votes awarded in the Cy Young voting process. This data will allow us to build a comprehensive dataset that we can feed into our model. In order to install the module, run the command “pip install pybaseball” and look through the documentation for how to access the data with the API. The data we will be using is from the years 1956 to 2016.
Here is a link to the API: https://github.com/jldbc/pybaseball

Another dataset that we will be using is from the Chadwick Bureau in order to find player IDs. Because different data sources use different ways to identify players, we need to utilize a data source that contains all of these IDs for different sources such as FanGraphs, Baseball-Reference, and MLB's internal ID system. A link to this csv file that contains these IDs can be found at this link:
https://drive.google.com/file/d/104pnELf5VSlz4Bxq8P_eWqcDwt8p8hGU/view?usp=sharing
https://mega.nz/file/qQs21QjL#QuogPGa7zuexuc37QeeyBqDfTNNsNQGOnEJkK0zyXCI

Method

Here are the following methods for each of our aforementioned research questions:

1. To answer our first research question, we will first access the awards data from the Lahman baseball database within the pybaseball API. We will select all vote-getters for the Cy Young Award for each league for each season. We will be looking specifically at the vote points share that each player received, as this is the best way to quantify these votes. Because writers can vote for second, third, etc. place winners, each place is assigned a given amount of points, such as seven for first place, four for second place, etc. These vote points are then standardized, meaning that we take the percentage of vote points a player got based on that year's maximum number of vote points awarded, and multiply this value by the maximum number of vote points ever awarded, which is 224. We will then join this data with pitching stats for each player for the given years with data from the Fangraphs data within the module. After this, we will split the data based on testing and training features and labels, respectively. The statistics from this data that we we want to focus on include ERA, WHIP, FIP, SO%, SO, WAR, Wins, K/9, BB/9, and HR/9, and which are all columns in the FanGraphs pitching_stats() function from pybaseball. Here are definitions of each of these, which will be the features of our model:
    - ERA - Earned Run Average
        - This is the average amount of runs that a pitcher gives up per game. A run is earned if it is not the result of an error or anything that is not outside of the pitcher's control. It is averaged to nine innings, with the equation being (9*ER)/IP, with IP being the innings that player pitched. A lower number is better, with an ERA under 3.00 being considered excellent. This is the most famous pitching statistic, although it has the shortcoming of being average to nine innings, while most starters only pitch about six

innings per game, with relievers pitching even less. A lower ERA is more common in relievers than starters, as is the case with WHIP below.

- WHIP - Walks and Hits Per Innings Pitched
    - This is like ERA, but only looks at walks and hits and is not standardized to nine innings. Walks and hits are two of the most common ways for a batter to get on base, so the pitcher wants to minimize his WHIP as baserunners contribute to earned runs. A WHIP under one is considered extremely exceptional, with a value under 1.2 being above average.
- FIP - Field Independent Pitching
    - This is a statistic very similar to ERA which removes the impact that the rest of the defense has on the pitching. It weights home runs, walks, hit by pitches, and strikeouts highly, and does not consider game events that may have something to deal with defense, such as hits. Defense can impact how hits can score runs if the member of the infield or outfield is not positioned correctly to be able to field the ball, or if they commit an error. FIP is calculated by the equation ((13*HR)+(3*(BB+HBP))-(2*K))/IP + constant, where the constant is the league ERA minus the league averaged value of the first part of the FIP equation. Like ERA, pitchers want to minimize this value, and a lower one is more common in relievers.
- SO (or K) - Strikeout
    - This is when a pitcher throws three strikes to a batter in an at-bat. A strike is when the ball passes through the strike zone formed near the batter's torso without the batter swinging. A strike can also occur if the batter swings at any ball thrown and misses. A foul ball is a strike for the first two strikes, but cannot strike a batter out. An exception to this is a foul tip, which is a foul ball that the batter still catches.
- SO% (or K%) - Strikeout Percentage
    - This is a statistic that standardizes strikeouts. Because strikeouts are a cumulative statistic, pitchers who pitch more will naturally have more strikeouts. A starter who was injured for part of the season or a reliever who simply does not play nearly as many innings as a starter will thus have a lower amount of strikeouts overall. Yet, the amount of strikeouts they throw compared to how much they pitched could be higher than whoever has the most strikeouts, hence why strikeout percentage is

useful. Like with strikeouts themselves, the higher the value for strikeout percentage, the better.

- WAR - Wins Above Replacement
    - This is what is known as a player value statistic, which measures the amount that an individual player contributed to their team winning games. This is done by comparing a given player to what is known as a replacement player, meaning one that is typically just below the minimum skill level to be even a bad starter in the MLB. They are often called AAAA players, as the highest level of the minor leagues is AAA, meaning that a replacement level player is somewhere in between. The WAR value for a player is the number of games the team would not have won had that player been replaced with a replacement level player; in other words, it is the amount of wins that a player directly contributed to a team. The statistic is non-standardized as it is very complex, but we will be using the FanGraphs equation for it, known as fWAR. Exceptional players will typically have at least four-five fWAR a season. The value can also be negative, as a player could have made mistakes that lost their teams games, such as errors or poor performance in general. However, these players are extremely unlikely to net any Cy Young Award votes. This statistic is very commonly thought to be one of the best for determining how good a player is due to its extremely comprehensive nature, although we still want to see if that is reflected in our model.
- Wins
    - This is the original pitching statistic, and measures whether or not the team won the game that a given pitcher pitches. A starting pitcher is awarded a win if a starting pitcher pitched a game where his team is in the lead for at least five innings and their team does not lose the lead for the rest of the game. If the pitcher did not pitch for five innings or did not have the lead throughout his appearance, the win is typically awarded to the relief pitcher who pitched without losing the lead if the offense regained the lead. However, if a reliever had a poor performance yet the offense rallied to maintain the lead, the game scorer may deem the otherwise win-qualifying performance as being “brief and ineffective” and will award the win to a previous pitcher. A key issue with this statistic is

that it does not account for how good the pitcher is within the context of his team. For example, if the starting pitcher pitches a very good game but still gives up one run (usually an acceptable amount), but the batters on the team are simply unable to score any runs off of a statistically worse off pitcher on the opposing team, the pitcher will not be awarded a win. If we look back retrospectively and see that a pitcher did not have many wins, it may not tell us much about how good that pitcher actually was for this reason. On the other hand, a poor pitcher may have a lot of wins simply as a result of a high scoring output of the offense. Thus, it is important to look at other statistics as well as wins. This issue is gone more into detail in question two.

- K/9 (SO/9) - Strikeouts Per Nine Innings
    - This statistic is like strikeout percentage, but is standardized to look like ERA. However, a higher number is better here, as strikeouts are good for pitchers.
- BB/9 - Walks Per Nine Innings
    - This statistic is like ERA, but specifically for walks. A walk is when the batter throws four balls that are not within the strike zone that the batter also does not swing at. A lower number is good here, as walks place runners on base, which can contribute to runs being scored.
- HR/9 - Home Runs Per Nine Innings
    - This statistic is like ERA, but specifically for home runs. A home run is when a fly ball goes over the outfield wall, rendering it not fieldable. The batter then immediately scores one run and advances to home, and any baserunners also advance home, scoring more runs. A lower number is good here, as home runs score the most runs at any given time for any action in baseball and greatly contribute to a pitcher's ERA

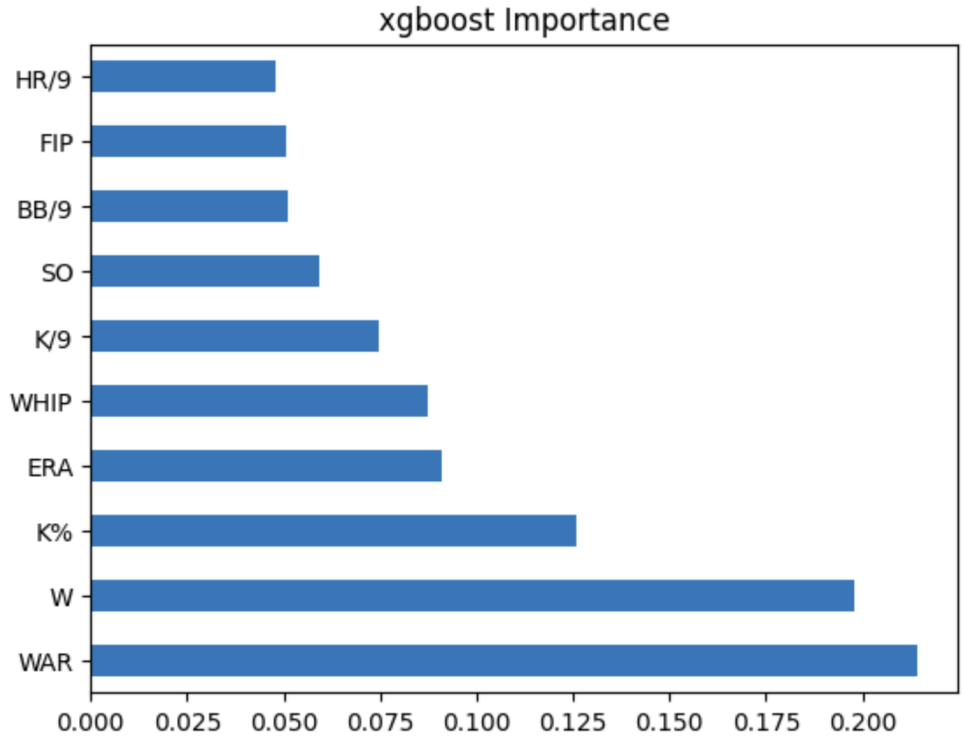
We will not be looking at pitch data statistics such as velocity and pitch movement as these are typically not considered for the award. We will then create a machine learning model to predict Cy Young Award votes based on the top-tier players' (that is, those who receive votes) pitching statistics. We will test different machine learning models for this, such as random forest, XGBoost, and decision trees. Once we have a baseline model created, we will work to reduce error from the model through hyperparameter tuning. We will use feature importance to identify the largest predictors of Cy Young Award votes.

2. With this research question, we will look at our joined data from the first research question method. As stated before, the statistics have columns of the same name from the databases within the pybaseball API. There are a number of different ways to track how the ways of thinking about statistics considered for the Cy Young Award have changed over time alongside the development of new statistics. The first method is to compare the data between dates; we will compare the statistics for vote-getters before and after 1990, because this splits the years of winners in half. The 90's were a decade where baseball statistics started to develop immensely, such as through the creation of stats like WAR, new methodologies for using statistics in baseball operations decisions such as those developed by Billy Beane, and the greater availability of baseball stats such as the Lahman database or baseball-reference in 2000. We will research these changes for our results. Next, we will look at how a linear regression model changes to compare the vote points with each of the statistics discussed in question two when we split it into these two eras (1956-1989 and 1990-2016). We will look at the different linear regression charts for these for visual comparisons, and we will then look at the differences in different statistics between the two, such as the correlation coefficient, R^2, etc. Finally, we will do outside research to figure out what statistics have become more prominent in modern history. We can use this information to add to our previous efforts with more context. We are choosing a linear regression model instead of machine learning because it is much less computationally intensive.
3. For the last question, we will first split the data based on whether or not the pitchers are starting or relief pitchers. Starting pitchers almost always start games, whereas relief pitchers almost never do this and come into the game in later innings to pitch for only a few innings. We will look at the levels of the pitching stats for the top-tier pitchers for each type, and will use basic linear regression to identify correlation between these values and the amount of Cy Young Award vote points, plot these, and then compare the case for starting vs. relief pitchers. We should be able to answer this question by observing the aggregated data tables and plots and make inferences based on these to come to a conclusion about how starting and relief pitchers are different for the purposes of vote-getting for the Cy Young Award, as well as by comparing different statistics related to these regressions, such as .the correlation coefficient, R^2, and more. Like with the previous question we will use regression for its efficiency when compared with ML.

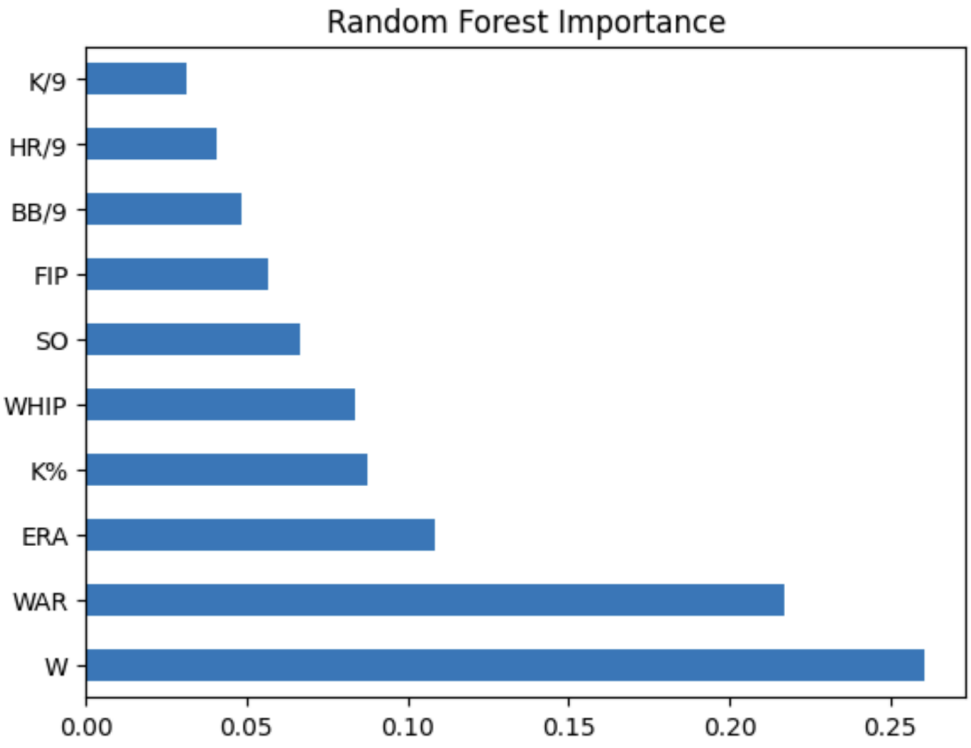
Results

Here are the results of each of our three research questions:

1. When comparing different machine learning models to predict the amount of Cy Young Award vote points that a top-tier player may receive based on the statistics outlined in the methods section for question one, we found that the XGBoost model provided the least amount of error, with approximately a mean absolute error of approximately 35.1775. This is in the context of a range of zero and 224 vote points for the Cy Young Award. The random forest model was a very close second, giving a mean absolute error of approximately 36.4328, and the decision tree model performed the poorest with a mean absolute error of approximately 41.9982. We will now go over each of these, arranged from the least error to the most error. When it came to feature importance, the most important feature for the XGBoost model was WAR, with the next being Wins (W). The next three important features that were substantially less important than the first two were strikeout percentage, ERA, and WHIP, respectively. XGBoost is a type of gradient boosting algorithm that utilizes weak decision trees, training them sequentially under one model. It uses gradients of what is called the loss function to prevent overfitting. The feature importance of all ten features for this model are shown in this chart below, with the most important ones at the bottom with the longest bar going towards the right meaning that it had the highest importance:

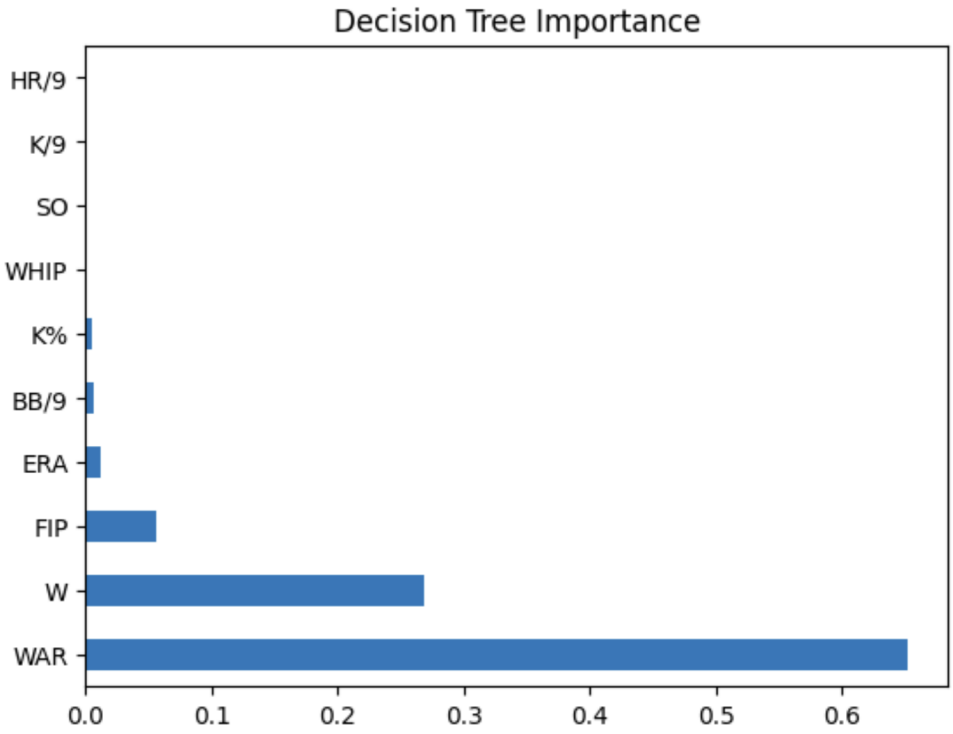


For hyperparameter tuning, it is more about the combination of values that returns the lowest error rather than the value for these specific ones. Hyperparameter tuning for this model included the number of estimators with choices of 100, 500, and 1000, the maximum amount of depth of the booster tree with choices of 2, 4, 6, and 8, the subsample which is the random sample of a percentage of the training data, with values of 0.1, 0.5, and 1, and finally the learning rate or ETA, which controls how fast the model learns, with values of 0.01, 0.1, 0.2, and 0.3. For the Random Forest model, the most important feature this time were wins, with WAR coming in second, swapping the result of the XGBoost model. Also, strikeout percentage and ERA were swapped as well. We can take this to mean that these differences simply come from how we chose the hyperparameters for this model as well as how this model works in general, which is by choosing a random subsample of the data for each tree to train on, hence the name random forest. Here is the feature importance chart, with the same format as the first one:



Hyperparameter tuning for this model included the number of estimators with values of 5, 20, 50, and 100, the maximum depth for the trees with values between 10, 30, 50, 70, and 90, the minimum samples split with values of 2, 6, and 10, which is the minimum number of new branches that come from a node in each tree, and finally the minimum samples for each leaf with values of 1, 3, and 4, where the leaf is the final node for the

prediction. Finally, we had the basic decision tree regression model, which is similar to the previous two models in that it creates decision trees, but it has no special features such as gradient boosting or random sampling of the data for each tree. For the feature performance of this model, only the wins and very specifically WAR mattered, which lines up with the other two models as well as conventional thinking, but provides no other information about the rest of the features used in this specific model. Interestingly, FIP was the only other feature with any sort of noticeable importance, which is quite different from the other two models. Here is the feature importance chart for the decision tree, using the same format as the previous two:



For this model, we used hyperparameters such as whether or not the splitters would be for the best case of data or a random selection of it, the max depth of the trees, with values of 3, 5, 7, 9, and 11. Next, we looked at the minimum samples per leaf as described before, with values of 1, 3, 5, 7, and 9, as well as the amount of fractional weight to give to each of these, with values of 0.1, 0.3, and 0.5. The hyperparameter tuning for all of the above models could have gone a bit better, but we still saw noticeable improvements by about five to twenty vote points worth of error going down between the default models and them after they had been tuned. Overwhelmingly from these three models, the obvious result and thus answer for this research question is that wins and Wins Above Replacement contribute the most to how many vote points a

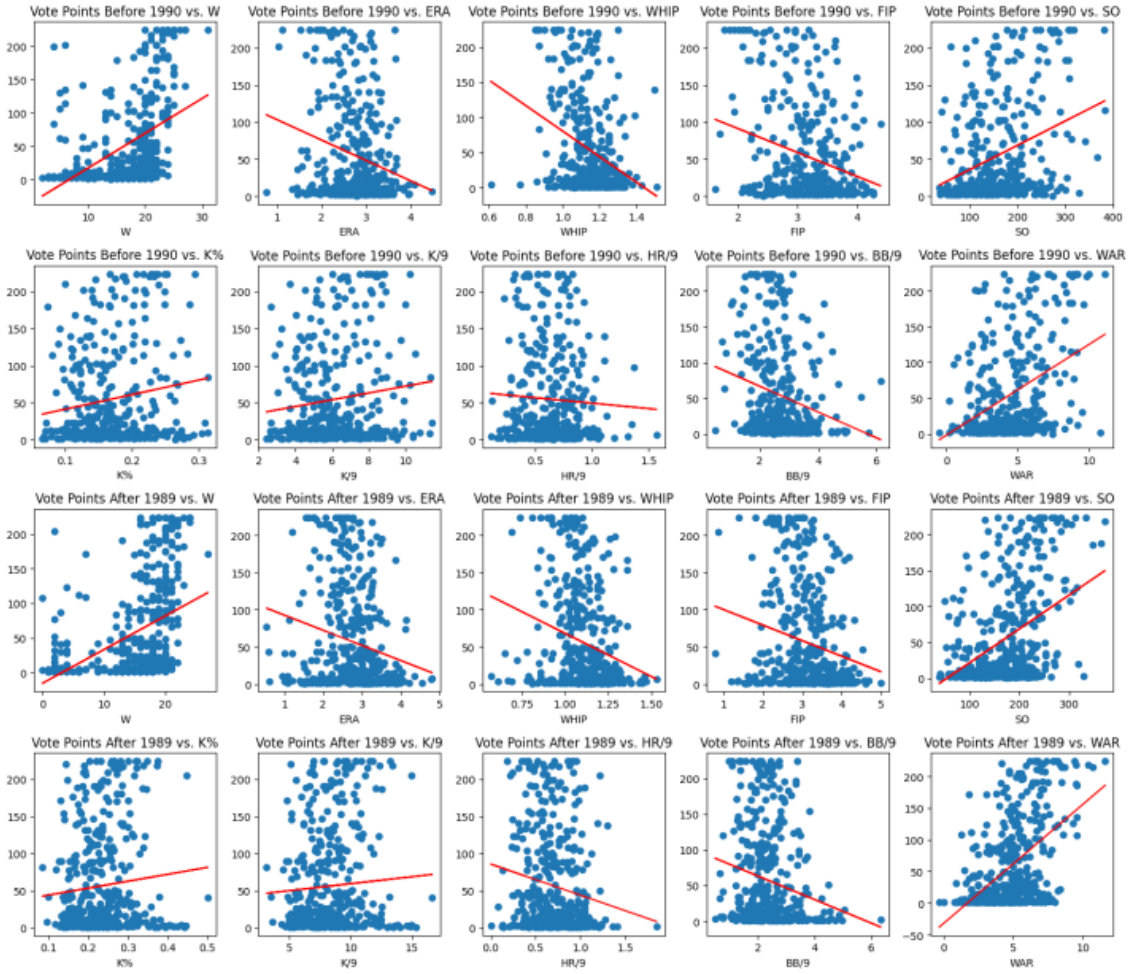
player receives for the Cy Young Award. This makes sense, as winning is the most important thing in baseball or in any team sports, and if one pitcher can individually add a great deal of win's to a team they are most likely to be a very good pitcher. The best players in the league are the most likely to win the awards for their position, and winning, particularly with WAR, shows how good a player is.

2. We did not find any significant impact on the statistics that correlate with the amount of Cy Young Award vote points that a player gets depending on what era they played in. First, we wanted to see if this was even a question worth asking. In order to do this we researched different people's opinions on whether or not the statistics have meaningfully changed. We already have knowledge of the creation of new statistics and ways of thinking about the sport first developed in the 1990's, which is why we split our data for the regression here in order to answer that question. Yet, we wanted to see if other thinkers corroborate the idea that this had any effect on which statistics most correlate with the amount of vote points won. The Society for American Baseball Research (SABR, sabermetrics' namesake) states in a report from 2006 (cited below) that:

> "Team-oriented accomplishments (defined as Wins, Winning Percentage, Saves, and Team Finish) has more often, over the last 50 years in both the National League and the American League, influenced the selection of the Cy Young Award winner than does leadership in individual accomplishments (defined as ERA, Strikeouts, WHIP, and Innings Pitched). There may be some evidence that this trend is changing in the 21st century, although Bartolo Colon would argue against that premise."

This more or less aligns with what we found in our results from question one, although they do not consider WAR. This is likely due to WAR not being nearly as commonly used in 2006. Bartolo "Big Sexy" Colón is brought up as he won the Cy Young Award in 2005 primarily due to his league leading 21 wins. One discussion of WAR vs. wins and the historical debate comes up in the SBNation documentary (cited below) on their YouTube channel Secret Base about Blue Jays star pitcher Dave Stieb in the 1980's and 90's never winning a Cy Young Award despite having very good WAR values to back him up, as well as him losing out to players who had worse WAR than he did. SBNation writers Jon Bois and Alex Rubenstein detail Stieb's rivalry with the infamous Jack Morris, who despite having a much lower WAR than Stieb during this time had more wins and also

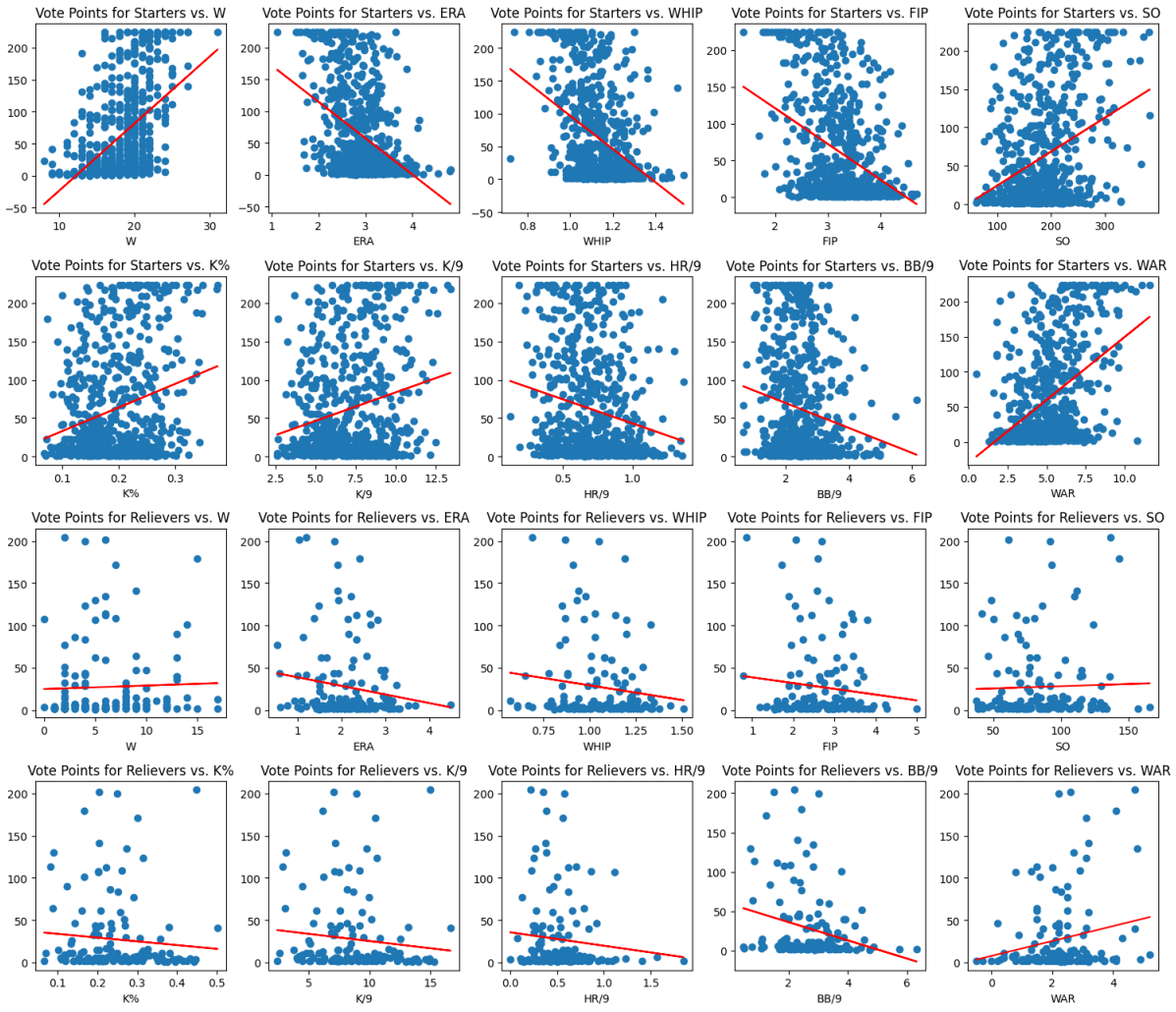
more Cy Young Award vote points. In fact, Stieb led MLB pitchers in WAR by a wide margin between 1980 and 1986, but the best he did when it came to the award was fourth place in 1982. An article written by Jason Catania for Bleach Report similarly details how in recent history some winners of the Cy Young Award have had less wins than previous years' winners, such as Félix Hernández winning the award despite going 13-12. However, in that scenario it was extremely obvious that the offensive capabilities of the Mariners were what hindered Félix as opposed to his pitching, due to the fact that the team went 62-100 for the season, a very bad record. With this research in mind, we sought to identify if there are any differences in the correlation between the statistics from the first question and the vote points. We did not find a meaningful difference in correlations between these eras. The first thing that we did in our analysis is to plot the data comparing the vote points and each of the pitching statistics for both eras to visually identify any differences. Here is that chart:



From a visual inspection, we can see that not much changes between the top two rows and the bottom two rows, being before 1990 and 1990 and after, respectively. Furthermore, the statistics for this regression model that our file prints out agrees with this, with the most significant difference between the two occurring with the HR/9 statistic, which has become more significant over time. Yet, it only has a correlation coefficient that changed by around 0.11, which is still quite low given that it's about -0.16 overall for 1990 and after. We can attribute this change to there being simply more home runs hit since 1990 with the rise of the steroid era in baseball. Thus, writers assigning Cy Young Award vote points may be more likely to take home run hitting into account, meaning that a pitcher with a lower HR/9 may be better off than others. The vertical nature of the data with the vote points having a very large range when compared to the other pitching statistics prevents the R^2 values from being particularly high as well. Based on the inspection of the correlation across these two time periods, we can reach the conclusion that the debate between which pitching statistics have become more significant over time has not been that significant and there has been little change. This is likely to due to the fact that even though there is a large part of flawed statistics such as wins that are dependent on how good a team may be, a pitcher who is very skilled will still have a degree of control over the way the game goes, meaning that the wins a pitcher gets will more often than not speak numbers about how successful they were as a pitcher. We can combine this knowledge of eras not making that much of a difference when it comes to the analysis with our knowledge from the first question to strengthen the idea of wins and WAR both being the largest predictors of the Cy Young Award vote points.

3. For our final research question we looked into the size of the difference between starting pitchers and relief pitchers for the correlation between Cy Young Award vote points and the pitching statistics described in the first question. There is no doubt that starting pitchers have dominated the Cy Young Award, with there being 640 vote-getters who were starters as opposed to only 145 vote-getters who were relief pitchers since the award's creation in 1956. Because of the different role that relief pitchers have when compared with starting pitchers, which is to come into the game at times where the context plays a large role in the decision making of the manager to put the relief pitcher into the game, based on things such as the game leverage or the matchup between the pitcher and the better. Most importantly, relief pitchers pitch much less in a game than the starting pitchers who start them, with relief pitchers pitching on average only one to

two innings while many starting pitchers will go for six or seven if they have a good outing. As a result, we could expect that relief pitchers may perform at a more elevated level as their roles are less taxing and they have less chance for injuries. We performed the same statistical tests used in question two for this question, but instead of splitting the data by era we split them by what type of pitcher they are. We found that there is a large statistical difference between the correlations between Cy Young Award vote points and the different pitching statistics for both starting and relief pitchers. Overall, pitching statistics were less correlated with the amount of vote points in general, which can be partially attributed to there being less data for relief pitchers than with starting pitchers. Here is the same grid of regression charts as the last question, but this time split by whether or not the pitcher is a starter or a reliever:



Here, we see that there is much more of a correlation between the different types of pitching statistics for starters than there are with relievers. The largest difference for wins

at a change of -0.45 in the correlation coefficient, followed by strikeout percentage with -0.34. This is interesting as strikeout percentage is commonly known as a statistic for relief pitchers, as it solves the problem that comes from the cumulative nature of the strikeout. The lowest change was with HR/9 at a change of 0.1 in the value for R, which makes sense as home runs are very looked down upon for relief pitchers as they pitch so much less and thus should avoid giving them up more than starters. Furthermore, the fact that many relief pitchers are brought in during situations after the previous pitcher has put runners on base can make home runs even more dangerous through the possibility of there being more runs batted in. We can attribute the difference for statistics such as wins being that, per the previous definition of wins, it is very difficult for relievers to get wins due to the five inning requirement that is only waived if the starting pitcher exited the game without holding the lead. Next, cumulative statistics such as strikeouts also do not have much correlation when compared to the case with starting pitchers as relief pitchers simply pitch less innings. Finally, with the reliever average based statistics with lower correlation coefficients than the starters, we can attribute this to the lesser data, but most importantly that there may be other statistics not utilized in this data that go into why relief pitchers may win the award. Most notably there are saves, which is a stat similar to wins, but specific to closing pitchers to reward them if they maintained a lead in a close game (thus “saving” the game). A final note is that some of these statistics, particularly ERA and WHIP, have a fair share of lower values among relief pitchers. This is due to relievers pitching much less and thus having less of a chance to make mistakes, especially considering that a lot of these statistics are standardized for nine innings, while relievers only pitch one or two. This could make these statistics appear less “special” to a writer voting for the Cy Young Award, as more relievers may have a given very good stat than, say, starting pitchers. A starting pitcher with a 2.40 ERA is going to look much, much better than a reliever with a 2.40 ERA.

Impact and Limitations

The most obvious implication of our results is that pitchers contributing wins to a team, whether that be through WAR or by maintaining enough of a lead throughout the game to net a win, are the most important predictors of Cy Young Award votes. Top-tier pitchers who feel like they have a shot at the Cy Young Award will benefit the most from the knowledge of these results as they will know what pitching statistics to most focus on, like wins and WAR. However, both of these statistics either are not always related to the performance of the pitcher, or are very hard to

control due to the amount of data points that are going into them. Thus, the pitcher should look into other, slightly less important features from our model, such as WHIP, ERA, and strikeout percentage. Relief pitchers should know that they have a very large upward battle as the performances of starting pitchers are much more likely to net them vote points. Another implication of this is that we shouldn't really focus too much on the differences between baseball statistics now and prior to 1990 when evaluating pitchers who received a lot of Cy Young Award vote points, as there aren't too many differences between the two eras contrary to the development and shift of focus to be on new statistics. Players excluded from these results are those who are not very good. Because this is baseball and only a sport, there will be no one harmed by these results.

A potential bias in the data is that it does not account for the quality of the entire team. Being a team sport, statistics such as wins can be very much determined by how good the team the pitcher is on is at things the pitcher cannot control, such as offense. This is partly why wins are much less of a valued statistic for the award after 1990, which can be seen in our analysis. Furthermore, statistics such as ERA, K/9, HR/9, and BB/9 have flaws in that pitchers rarely pitch for all nine innings in a game, instead 6-7 for starters and 1-2 for relievers. These stats being set to nine innings is a product of the early eras of the game (known as the dead ball era) where offensive capabilities were much less and pitchers commonly pitched for entire games. The limitations of our analysis are that running these machine learning models takes a very long time, and thus reducing error through hyperparameter tuning may not be as effective as it could be. Within the context of there being 224 Cy Young Award vote points total that can be given out, an error of around 35 with this is not that bad, but could probably be better. Another limitation is that there is a very large difference between many of the vote values, as many of these top tier pitchers receive only a few votes lower down on the lists that are not weighted the same as, say, first place votes. As first place votes are seven times as important as fifth place votes, there are a great deal more players with less vote points overall, which, particularly in our question two and three results creates very vertical data distributions that may not tell us a ton about the correlations between the vote points and a given statistic. Many of our R and R^2 values were quite small and not the most statistically significant. With these limitations in mind, people should use our conclusions to determine that although our research lines up with conventional thinking that winning games, particularly with Wins Above Replacement, is a very good measure of how good a pitcher is, there still may be inaccuracies with the importance of other more specific statistics. Further research could be done to include wins or WAR to see

how the models change. They should not use our conclusions to determine that one statistic in particular such as wins is the superior statistic for all of baseball analytics, as there were plenty of other baseball stats not used in this analysis for reasons such as complexity and data availability. With this being said, many of the statistics that did end up showing importance through our models do follow conventional logic.

Challenge goals

These are the two challenge goals that we will be meeting with this project:

1. **Multiple Datasets**
    - Different baseball datasets focus on different statistics. For example, FanGraphs data for a given pitcher's sabermetric data such as ERA may not have information about the different awards that the player may have. Meanwhile, the Lahman Baseball Database has a dedicated database for every award for every season. This includes information about the amount of votes awarded to the different pitchers in the year. Thus, in order to only look at the pitchers who won the award, or almost won the award, we will need to join these datasets for the purposes of our analysis. This only change in this challenge goal over the course of our project was that we did not use data from Statcast pitch data, such as velocity or pitch movement, as it is not commonly brought up in the context of the Cy Young Award.
2. **Machine Learning**
    - We will be using machine learning as we seek to predict the amount of Cy Young Award vote points based on the pitching statistics of top-tier pitchers. We will consider the hyperparameters for the models and separate features and labels to meaningfully be able to predict Cy Young votes, with the features being the stats of top-tier pitchers and the labels being the votes a top-tier player receives. The final algorithm will attempt to predict the amount of votes based on differing features that we give the program. Finally, the model will be able to be tested against a section of our data that is reserved for testing purposes as opposed to training. There are many different forms of machine learning for predictions, and different methods will return models with different predictions and error. We seek to identify which model out of a selection of them, such as random forest, XGBoost, and decision trees, will give us the lowest error for our prediction.

Work Plan Evaluation

Below are our original work plan details as well as evaluations for each to look at how much they were specifically followed.

1. Join the necessary datasets as well as filter and aggregate the data to become the features and labels to be used for the machine learning algorithm (10 hours). This will be done by mostly Luke in the file and class for the data. The class in this file will have methods for cleaning the data and splitting it up in the different ways needed for the different research questions.
    - This was all quite accurate. Accessing and merging the data was rather streamlined with few issues. It took approximately 10 hours and was done primarily by Luke. A file specifically for the data was created and a class was created inside that. The data cleaning was done in the initializer method of the class, and other methods returned the split data. It took quite a while to run this program as sixty seasons worth of data were being accessed.
2. Adjust or fix any anomalies or missing data within our dataset (3 hours). This will also be mostly done by Luke in the first file in the program.
    - This was rather accurate, and it took approximately 3 hours. The key issue was getting the player IDs into the program, as the API to access these from the Chadwick Bureau has been broken for a few months. Instead, a local source was used. There was no missing data within the awards or FanGraphs data. Within the player ID data, five players' Baseball-Reference IDs did not line up with those in the Lahman database. These players were J.R. Richard, Freddy García, Johan Santana, CC Sabathia, and R.A. Dickey; their IDs were fixed within the data cleaning part of the program.
3. Train and use a machine learning algorithm such as random forest to predict future Cy Young vote totals, how the statistics most important for the prediction have changed over time, and which young players are most likely to eventually win the award (12 hours). This will be mainly done by Aidan in the second file of the program for the model, but Luke will work on this as well as it is a more detailed part of the assignment. The class in this file will be split up for different models or statistical analysis for the different research questions.
    - This was somewhat accurate. The model portion of the project took approximately twelve hours as stated, but likely did not include much of the work on questions two and three to build the linear regression models to see how the

Cy Young Award vote points are correlated with the statistics outlined in question one, specifically split between before 1990 as well as it and after, as well as starting vs. relief pitchers. This work took approximately another ten hours, and was quite streamlined. There were a lot of similarities between the two questions, so some of the code from question two could be used for question three.

An aspect of our work plan we forgot about in our proposal was the writing of this document, any documentation in the program files, as well as the read me file for the project. All of this took about 10 hours. Our workflow for the project will be using GitHub to share our programs with each other and to make changes to what other group members have done. We will use Google Drive and Google Docs to collaborate on this deliverable. We are also using Discord in order to communicate with each other, and may periodically meet in person to get work done in a more face-to-face setting. We will all be using Python locally for working, specifically VSCode.

- GitHub, Discord, VSCode, Google Drive, and Google Docs all worked very well and helped streamline our project

Testing

We created numerous tests for our methods and programs. These included methods in the testing file that run all of the methods in both the PitchingModel and PitchingData classes. We further broke this down for the PitchingData class to have one method to test the data for question one, and another method to test the data for question two. Within the testing file, the time period covered in the test runs is easily manipulated in order to be able to use smaller time periods to reduce run time. We also checked to see if the shape of smaller datasets from calling the function matched the true amount of pitchers who received Cy Young Award votes over a shorter subset of years. These values can be manually counted using the awards page for the given seasons from Baseball-Reference, and we used assert statements to check whether or not it was valid. To test the machine learning algorithms, we returned the error produced by the test and made sure that it was within the bounds of zero to 224 vote points that a pitcher could get for the Cy Young Award, which is the case. All of the data fed to the machine learning methods was accurate and the resulting errors were relatively low considering the very large data set. We ensured that the value for the MAE, with 35, 36, and 41 for the three models respectively, were

Collaboration

This project was completed by Aidan Murphy and Luke VanHouten, with the assistance of our TA mentor Murtaza Ali. Min Heo was in our group to start out, but he had to withdraw from the quarter due to a family emergency. We consulted the pybaseball, XGBoost, and Scikit-Learn documentation for assistance with organizing our data and building our model. These can be found at these links:

https://github.com/jldbc/pybaseball/tree/master/docs

https://xgboost.readthedocs.io/en/stable/python/python_intro.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Furthermore, there were a number of sources that were specifically used for the second research question. These can be found at these links:

https://sabr.org/journal/article/the-cy-young-award-individual-or-team-recognition/

https://youtu.be/wSjf3AjM7jY?t=1586

https://bleacherreport.com/articles/1777341-why-pitcher-wins-should-be-completely-eliminated-from-cy-young-criteria