Devan Bose
Jordan Denish
Malcolm Mashig
Christian Rogers

Predicting MLB Pitcher Value

Statistics Capstone Project

Technical Report

University of Virginia

Spring 2020

Introduction

For our capstone project, we wanted to analyze Major League Baseball, hoping to use historical data and various statistics in order to predict player value and make conclusions about the sport. We narrowed our focus to be more specific, ultimately choosing to analyze and project future performance of starting pitchers. Our inspiration for picking this particular position came from the fact that starting pitchers have become increasingly important and prominent among playoff contending teams. The Washington Nationals, the winner of the 2019 World Series, for example, were anchored by three elite, highly-paid starting pitchers in Max Scherzer, Stephen Strasburg, and Patrick Corbin.

Furthermore, the size and length of the contracts that teams are giving to starting pitchers are tremendously large. This past winter, during baseball's free agency, over $1 billion was spent on starting pitchers, the most free agency dollars spent on starting pitchers in baseball history. Free agent Gerrit Cole signed a nine-year, $324 million contract with the New York Yankees, which is the largest deal ever signed by a pitcher. Gerrit Cole is currently 29 years old, which is considered to be one of the prime years for a starting pitcher, but he will be 38 years old at the end of his contract, when most pitchers are already in decline and considering retirement. Only the big market baseball teams like New York, Boston, and Philadelphia have the spending capabilities to afford these massive free agent signings, while the smaller market teams like Tampa Bay and Oakland must find cheap alternatives in order to be able to compete. This relatively recent trend in the importance that the sport is placing on finding valuable starting pitching led us to make it the concentration of our capstone project.

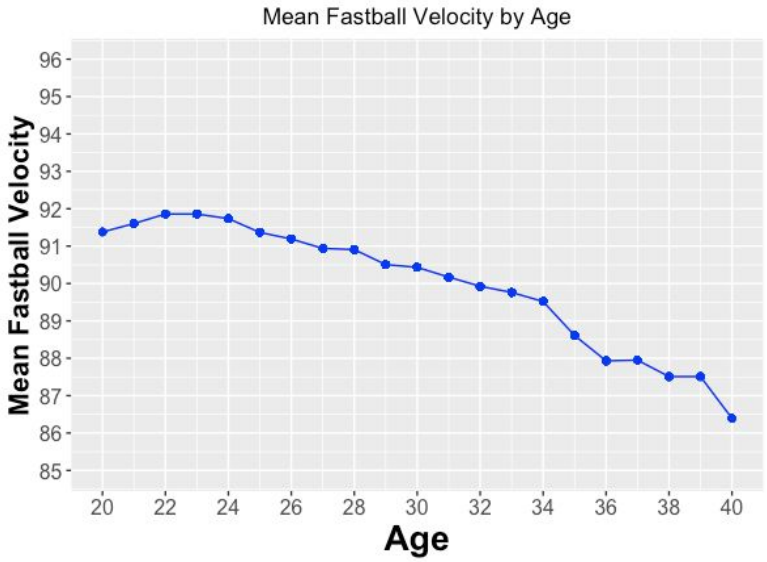
Data Collection

After refining our ideas, we gathered the data relevant to our project. We took the approach of collecting a large amount of data, hoping this would help us further revise and pinpoint a particular question of interest. A website called FanGraphs was very helpful in this data collection process, containing seemingly every imaginable statistic for every player, separated by season. The site also provided the ability for us to filter the data based on position and various other criteria and easily export the data as a CSV formatted file. We decided to harvest all available starting pitcher data, with two significant constraints. First, we limited how far back our data went to the year 2002. Knowing that only somewhat recent data would be useful for drawing conclusions about the future, coupled with the fact that several advanced baseball metrics, such as xFIP and BABIP, have only been tracked since 2002, it made sense to cut the data off at this year. Our other constraint was that we limited the data to the seasons where the pitcher threw at least 80 innings. Though this admittedly is a somewhat arbitrary cutoff, our experimentation found that this excluded artificially short seasons, usually due to injury, while still providing a large amount of data to analyze. Further, after substantial back and

forth consideration about whether to consider injuries, we decided not to pursue this analysis, as the extreme amount of randomness in injuries would likely prevent us from coming away with any meaningful results. Thus, our complete data set included all relevant performance statistics for each season with at least 80 innings pitched for every starting pitcher from 2002-2019.

Exploratory Data Analysis

Once we imported our data into R, the software used for the entirety of our capstone project, we began exploring the data, looking for any notable or surprising trends. Knowing we wanted to somehow assess the value of pitchers over time, we looked at how various statistics interacted with the age of each starting pitcher throughout their careers. This exploratory analysis led us to focus in on a smaller list of statistics, such as Earned Run Average, fastball velocity, home run rate, and strikeout rate as these statistics appeared to have notable relationships with age. The graph below shows how the mean fastball velocity of starting pitchers in our data set decreases steadily as they age. This process additionally prompted us to begin considering how we would quantify pitcher performance and thus value, the centerpiece of our project which had been ambiguous up until this point.



xFIP - The Best Metric to Evaluate Pitcher Performance

One of the most important decisions we had to make as a group was choosing which metric to evaluate pitchers on, as there are many such statistics that each have strengths and weaknesses. We wanted to choose a metric that was descriptive, in that it encapsulates valuable information, and predictive, in that it is relatively stable year-to-year. The traditional metric for

measuring pitcher success is Earned Run Average (ERA), an easy to understand statistic which simply divides the number of earned runs a pitcher allows by the number of innings he pitches, and then multiplies this average by nine to mimic the length of a baseball game. ERA makes sense as a critical performance measure, as a pitcher's primary goal for his team is to prevent runs, and a pitcher's ERA will show how effective he was. However, ERA has many issues that have caused most baseball analysts to move beyond it. For one, the fielding of the team around a pitcher is not included in the metric, which certainly has an impact on the number of runs that he gives up. A pitcher with a better defense is more likely to have a lower ERA than a pitcher with a worse defense. Another negative aspect of ERA is that it includes no information about the ballpark that the pitcher plays in. For instance, Coors Field of the Colorado Rockies is notoriously easy to hit home runs in, as balls fly farther in the high elevation of Denver. On the other end of the spectrum, Petco Park of the San Diego Padres has deep fences and a spacious outfield which are hostile to hitters. Since the environment that pitchers play in varies so much from pitcher to pitcher, a metric like ERA has major flaws that make it difficult to encapsulate a pitcher's true value.

Another metric that we strongly considered was Wins Above Replacement (WAR), which through a complicated formula attempts to derive the number of wins a player was worth to their team above a replacement level player in a given season. For this metric, a replacement level player means an easy to obtain player that a team could sign during the season. WAR accounts for some of the shortcomings of ERA, but it depends heavily on the number of innings pitched by a player (because it is a cumulative metric). The number of innings pitched can vary wildly depending on a player's coaching as well as his own injuries. We believe that most injuries are fluky and random, meaning that they are extremely difficult to predict with reasonable accuracy.

Through our knowledge of baseball as well as our exploratory data analysis, we eventually decided to use Expected Fielding Independent Pitching (xFIP), which is an improvement upon Fielding Independent Pitching (FIP), as our performance measuring metric. FIP is a metric designed to be a better indicator of pitcher success than ERA, as it only accounts for outcomes that the pitcher exclusively controls, which are walks, strikeouts, home runs, and hit batters. FIP also includes a constant so that it has a similar average each year as ERA, which means that smaller FIP values represent more valuable pitchers. The most important thing about FIP is that by only focusing on the outcomes that a pitcher can control, it attempts to remove the high variance in ERA from season to season. The reason why pitchers' ERAs can fluctuate by large amounts is that luck is a huge factor in the game of baseball. A pitcher can control whether or not the batter can hit the ball, but a pitcher cannot control exactly what happens when the batter hits the ball. For example, a skilled pitcher can make terrific pitches that result in soft contact such as a bloop single, while an unskilled pitcher can have weak pitches that result in a rocket line drive off the bat that is hit directly at the outfielder for an out. This means that a pitcher with the same skills can have very different ERAs in different seasons. With so many

pitchers in baseball, this variance can result in unpredictable outcomes that prevent or allow excess earned runs for a pitcher. Essentially, FIP removes the variance introduced by the flukiness of allowing runs and defense surrounding the pitcher. xFIP takes this metric a step further and also accounts for park differences, by using the league average home run per fly ball percentage for each pitcher. Since we would expect the same pitcher to give up more home runs in Coors Field than in Petco Park, it is unfair to judge all pitchers on the same scale for their home runs allowed. Ultimately, xFIP is far more descriptive of a pitcher's true value as a player than ERA and a stronger indicator of a player's future performance than WAR because it doesn't rely on the number of innings pitched. Therefore, we decided to use xFIP as our dependent variable for modeling pitcher value.

Research Question

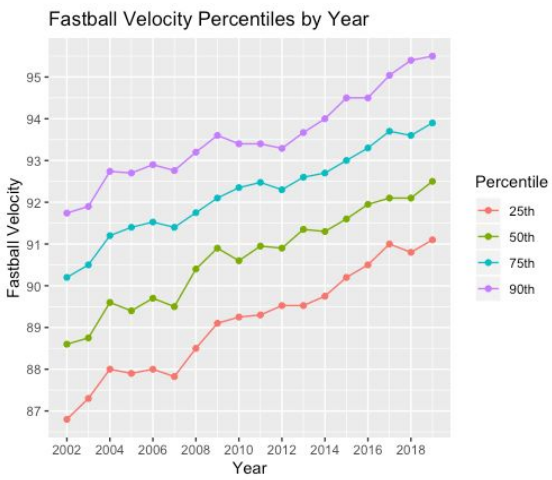
Once we were satisfied and confident in using xFIP as our performance metric, we were able to clearly state our research question: Can we accurately predict an MLB starting pitcher's future xFIP over the next few seasons using various performance statistics from prior seasons in order to estimate relative future value of the pitcher? Simply put, we wanted to create a model that could take any pitcher, along with the statistics of all of their previous seasons, as an input and make projections for the xFIP of their next few seasons. From our main research question, two sub-questions naturally emerged as well. In particular, we wanted to look at how pitchers "aged", or how their performance was affected as they got older, as well as whether all prior seasons were equally predictive or if more recent seasons were more important predictors of future xFIP. Our data set provided us with all of the relevant and necessary information to model xFIP over time, as well as answer our sub-questions. Further, viewing our data as a sample of the population of all starting pitcher seasons with at least 80 innings pitched, both past and future, we will be able to generalize our results to future seasons.

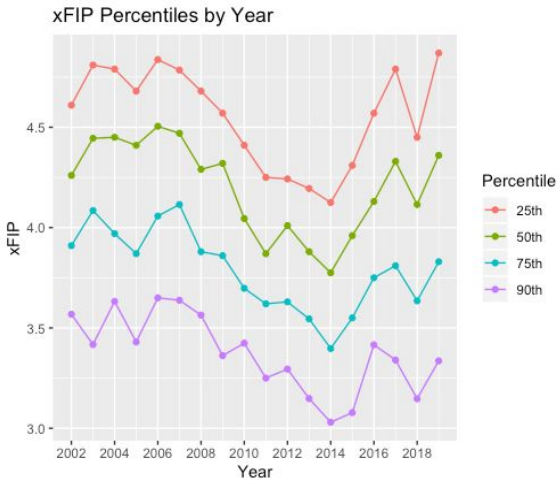
Preliminary Analysis

Next, we conducted some preliminary linear regression using the statistics from a pitcher's previous few years to predict their xFIP in the next year or two in the future. Some examples of statistics we considered as inputs to this preliminary modeling, due to our exploratory analysis and prior knowledge of baseball, were age, fastball velocity, fastball percentage (percentage of all pitches that are fastballs), strikeout rate, walk rate, WHIP, home run rate, and xFIP from previous years. In this modeling, we found that the xFIP from the previous season is the most influential predictor in determining a pitcher's xFIP in the next season. Additionally, our initial models showed that age, fastball velocity, fastball percentage, strikeout rate, and potentially home run rate were significant variables as well. This makes sense because as a pitcher ages out of his prime years, the speed of their pitches typically decreases,

resulting in a dip in performance. Also, while home runs can be somewhat inconsistent, some pitchers might be more prone to hard contact than others.

One important step that we took to improve our regression models was standardizing many of our variables. Baseball as a game has changed substantially even over the last fifteen to twenty years. For example, pitchers are throwing faster than they ever have before, meaning a high fastball velocity in 2002 is merely middle of the pack now. Therefore, we believe that standardizing these variables for each year will minimize the impact of differences between eras and changes in the game, and allow us to use all of our data for model-building.





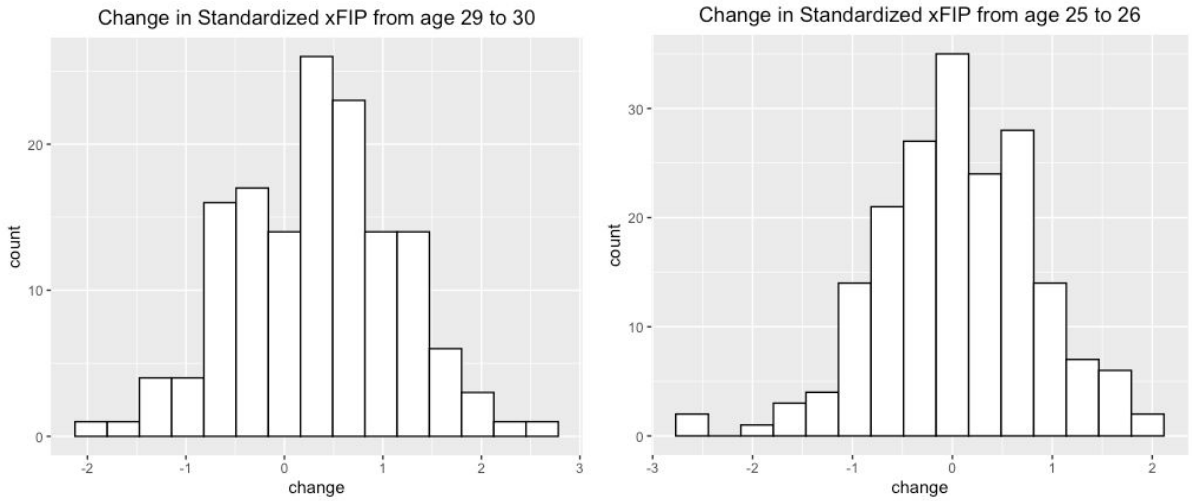
Part of the process of standardizing these variables, however, was making sure that the separations between groups within each season are not changing much. For instance, if the 90th percentile fastball velocity has grown much larger than the 75th percentile fastball velocity, then standardizing will not be capturing that change in the relative difference between the two. Above are graphs for different percentiles for fastball velocity and xFIP over this time period. These plots show that the relative groups have maintained their separations over time, meaning that they rise and fall together at roughly the same rate over time. Therefore, we were confident that standardizing our variables would be an effective tool to account for the changing nature of the game.

Analysis Ideas

After performing initial analysis, our most difficult step was to develop a model that could predict xFIP values multiple seasons in the future. We needed to build a model that would most likely depend on the previous predictions of the model, so we felt that a sort of recursive structure was necessary. The two main models that we considered are below.

Simulation

The basic idea of the simulation method was to look at the average change in standardized xFIP for every two year span for every starting pitcher in our data set for every single age change. We looked at the distribution of standardized xFIP and planned to generate a random value from this distribution to predict standardized xFIP for the next year. The main assumption of this simulation method was that the change in standardized xFIP from year to year for a certain age is normally distributed, as we would be using a normal distribution to determine the year-to-year change in xFIP for each player. The histograms below demonstrate this year-to-year change for two different age groups: 25-26 and 29-30. We looked at the histograms for all possible age ranges and most of them appeared generally normal. While not perfect, we believe these plots demonstrate that the change in xFIP is close enough to being normally distributed that we tried using this method.



We implemented this simulation model, generating a random value from the distribution of standardized xFIP from season to season depending on the pitcher's age value. The theory behind the simulation model was that the most significant predictor for xFIP in our initial exploratory modeling was that same pitcher's xFIP value in the previous season. This simulation model would only include a pitcher's age and most recent xFIP value, generating simulated predictions based on the distribution of changes in xFIP. This means that we used the pitcher's most recent season of data, applying a random change in xFIP from the distribution for that pitcher's age, and adding this change in xFIP to the pitcher's most recent xFIP value. We generated predictions one season at a time, using our most recent xFIP prediction to generate three years of xFIP predictions in the future.

We were able to generate predictions for xFIP that looked reasonable, but quickly realized that the simulation treated every player equivalently regardless of their abilities. Our

projections expected every player to fluctuate around their most recent xFIP value, which did not really add any predictive value. We wanted to add more predictors to the model, so the predictions depended on more than just a player's age and recent xFIP value, and this is where we tried using our lagged linear regression model.

Lagged Linear Regression

The theory of our lagged linear regression xFIP model is derived from the idea that in order to generate predictions for the next three years, we would want to produce incremental predictions. If we are looking at a pitcher's performance in 2019, we will first predict xFIP in 2020 and then use that prediction to generate our second prediction in 2021. Finally, we will use the prediction in 2021 to predict the pitcher's xFIP value in 2022.

Rather than using this same model to predict multiple years in advance, we used various "submodels" to predict each of these inputs for the next year, and then used these submodel predictions as inputs into our main model to predict the next year's xFIP. This is a significant advantage compared to our simulation method, which simulated xFIP values entirely based on a pitcher's age and previous season's xFIP. For example, we could use additional predictors such as fastball velocity and fastball percentage (percentage of pitches that are fastballs) in 2019 to predict 2020's xFIP; then, with our submodels, we could project each of the other unknown predictors that are inputs in our main model in 2020. We would then be able to plug these predicted values into our main model to predict 2021 xFIP and continue the process into the future. This is relevant because there were a large number of explanatory variables that could have potentially been used in our model, such as age, years of experience, or pitcher type, and would be known as predictions were made through the future. We can assume age increases by one each year, but we needed submodels that could estimate values for the explanatory variables that change from year to year and depend on additional inputs, such as fastball velocity.

Ultimately, after looking at the large amounts of variation in the predictions from our simulation model and understanding that the simulation model treated all players, even those of varying levels, equivalently, we chose to pursue this lagged linear regression model. The lagged linear regression eventually became our final model for this project.

Model Building

In building our future xFIP prediction model, we used several predictors including a pitcher's age, his previous xFIP values, his fastball velocity, and his fastball percentage. These predictors describe our full model that will predict a pitcher's **standardized** xFIP value in any future season given last year's values for his age, fastball velocity, fastball percentage, and xFIP. However, we do not know what a pitcher's fastball velocity and fastball percentage will be in the

future, so we must use "submodels" to predict those values as well. We used these submodels to project our predictors in future seasons, generating one season and one xFIP prediction at a time.

We started with a "submodel" that predicts **standardized** fastball velocity based on the age of the pitcher and their fastball velocity in the season prior. The model has the following form:

Submodel 1

$$FBv = B_0 + B_1(lagFBv) + B_2(age) + B_3(young) + B_4(prime) + B_5(age * young) + B_6(age * prime)$$

$$FBv = 0.318 + 0.935(lagFBv) + -0.016(age) + 0.508(young) + 0.278(prime) + -0.020(age * young) + -0.009(age * prime)$$

FBv represents the standardized fastball velocity that we want to predict (for the upcoming season). *LagFBv* represents standardized fastball velocity for the year prior. *Age* represents the age of the pitcher for the year of the prediction. Since age increases by one for each pitcher each year, we always know age in the year of the prediction. *Young* represents a dummy variable that is 1 when the player is younger than 25 years old in the year of prediction and 0 otherwise. *Prime* represents a second dummy variable that is 1 when the player is between the ages of 25 and 31 and 0 otherwise. If both *young* and *prime* are 0, then the pitcher is above 31 and classified as *old*.

Next, we moved onto a "submodel" that predicts **standardized** fastball percentage based on the age of the pitcher and their fastball percentage in the season prior. The model has the following form:

Submodel 2

$$FBp = B_0 + B_1(lagFBp) + B_2(age) + B_3(young) + B_4(prime) + B_5(age * young) + B_6(age * prime)$$

$$FBp = -0.094 + 0.866(lagFBp) + 0.001(age) - 0.079(young) + 0.249(prime) + 0.006(age * young) - 0.007(age * prime)$$

FBp represents the standardized fastball percentage that we want to predict. *LagFBp* represents standardized fastball percentage for the year prior. *Age*, *young,* and *prime* have the same meaning as in the fastball velocity submodel.

With the ability to make predictions for fastball velocity and fastball percentage for pitchers in a future year based on the latest available year, we have all the predictors necessary to make predictions for xFIP. We can now create one main xFIP model with the following form that predicts **standardized** xFIP.

Model 1

$$xFIP = B_0 + B_1(lagxFIP) + B_2(age) + B_3(predictedFBv) + B_4(predictedFBp)$$

$$xFIP = -0.388 + 0.639(lagxFIP) + 0.014(age) + -0.115(predictedFBv) + -0.0025(predictedFBp)$$

xFIP represents the standardized xFIP that we want to predict for the next season. *PredictedFBv* represents the standardized fastball velocity we predict with the fastball velocity submodel.

PredictedFBp represents the standardized fastball percentage we predict with the fastball percentage submodel. *Age* has the same meaning as in the submodels. *LagxFIP* represents the standardized xFIP from the year prior.

With this modeling process framework, we were able to predict standardized xFIP for a pitcher multiple years in the future. Once again, we predicted one year in the future, and then used the predicted xFIP value as input for the *lagxFIP* variable in order to predict the xFIP two years in the future. Each new year out that we predicted, we increased age by one and used the predictions for the last year (fastball velocity, fastball percentage, and xFIP) as input for the next round of submodel and model predictions. We planned to use the model to generate xFIP predictions for the next three seasons, 2020, 2021, and 2022. Going three seasons into the future paints a meaningful picture of the pitcher's projected performance, and predicting any further resulted in increasingly reduced accuracy and higher variance.

Generating Predictions

After creating our regression model to predict xFIP, before we could use it for prediction of three seasons into the future, we wanted to test the model on the data that we already had. Essentially, we ran the model for every player with at least two consecutive seasons. For players with more than two consecutive seasons, we could assess our prediction for additional seasons in the future. However, we needed to un-standardize the xFIP predictions using the mean and standard deviation of the distribution of xFIP values from the prediction season. By comparing our predictions to actual values in the data, we continued to finetune our submodels and main models. To determine how close our projections were to the actual values, we used the mean absolute difference, which is just the average distance between the actual value and our predicted value. We found that the predictions for the following season registered a mean difference of roughly 0.37 in xFIP, while the year two and year three predictions had values around 0.46 and 0.55, respectively. The mean differences are in the table below. Since xFIP is on the same scale as ERA, this translates to being around a half a run off per nine innings pitched for all three future years of prediction. We thought that this was a reasonable amount of error, and that our model was ready to predict the future.

	Year 1 Prediction	Year 2 Prediction	Year 3 Prediction
Training Error using Model 1	0.3754712	0.4612559	0.556610

For our first attempt at predicting future seasons for all pitchers from 2019, we only used their 2019 statistics to predict their xFIP for the following three seasons. It was fairly simple to source the script we had already created that had the model built and ready to use. We subsetted

our data to 2019 pitchers and generated predictions for the next three seasons. Since the models use and predict standardized xFIP, we needed to decide how to un-standardize xFIP for future seasons, where the distribution of xFIP is unknown. After discussion with Professor Holt, we decided to un-standardize the future xFIP projections in terms of the average distribution of xFIP from the 2017, 2018, and 2019 seasons. By using this procedure, we assumed that the xFIP distribution's mean and standard deviation will be fairly close to what they have been in the past few years. The mean and standard deviation of the xFIP distribution for the past three seasons, along with the projected 2020 distribution, are in the table below. This procedure also makes it easier to compare values in future seasons, as every predicted value is on a level playing field.

Year	Mean xFIP	SD xFIP
2017	4.101	0.611
2018	3.928	0.616
2019	4.391	0.686
2020 (Projected)	4.140	0.638

Our initial predictions looked fairly solid, as good pitchers were projected to remain good and older pitchers were predicted to see a drop in performance. However, we noticed that some results were thrown off by only having one year of data. For instance, Aaron Nola of the Philadelphia Phillies was excellent in the 2018 season (xFIP of 3.21), but had a down year in 2019 (xFIP of 3.82). Since he is a young pitcher, most baseball fans and analysts predict that he will regain his form for 2020 and beyond. However, since our model only factored in the down year for Nola, it predicted that he would remain an average pitcher for the next few seasons. Because of cases like Nola, we decided to incorporate an extra year of data for all pitchers who met the innings threshold in 2018 in order to use more context for the model. For the pitchers that did not reach the 80 inning threshold in 2018, we included only their 2019 season. After adding the extra year of data, the predictions looked much more realistic. We created Models 2 and 3, in addition to our original Model 1, to be used when two or three years of past data were available.

Model 1

$$xFIP = B_0 + B_1(lagxFIP) + B_2(age) + B_3(predictedFBv) + B_4(predictedFBp)$$

$$xFIP = -0.388 + 0.639(lagxFIP) + 0.014(age) + -0.115(predictedFBv) + -0.0025(predictedFBp)$$

Model 2

$$xFIP = B_0 + B_1(lagxFIP) + B_2(age) + B_3(predictedFBv) + B_4(predictedFBp) + B_5(lagxFIP2)$$

$$xFIP = -0.324 + -0.014(lagxFIP) + 0.014(age) + -0.102(predictedFBv) + -0.014(predictedFBp) + 0.484(lagxFIP2)$$

Model 3

$$xFIP = B_0 + B_1(lagxFIP) + B_2(age) + B_3(predictedFBv) + B_4(predictedFBp) + B_5(lagxFIP2) + B_6(lagxFIP3)$$

$$xFIP = -0.267 + 0.465(lagxFIP) + 0.012(age) + -0.109(predictedFBv) + -0.007(predictedFBp) + 0.264(lagxFIP2) + 0.036(lagxFIP3)$$

LagxFIP2 represents standardized xFIP from two seasons prior. *LagxFIP3* represents standardized xFIP from three seasons prior. Further, we found that xFIP from four or more seasons prior had little predictive power for the future. The benefit of specifying three models in this way was so that we could predict xFIP with more prior seasons of data when available. When the last three years of data were available, we used Model 3. When the last two years of data were available, we used Model 2. Finally, when only the last year of data was available, we used Model 1. Thus, whenever possible, we used more xFIP values from past seasons to predict future xFIP.

One strong case study that demonstrates the benefit of adding another year of data was a pair of pitchers, Marco Gonzales and Lucas Giolito. Giolito was terrible in 2018 (xFIP of 5.46), but good in 2019 (xFIP of 3.66), and Marco Gonzales was good in 2018 (xFIP of 3.59), but struggled in 2019 (xFIP of 5.11). Giolito was predicted to remain decent the following three seasons despite his poor 2018 numbers, while Gonzales was projected to improve from his bad 2019 but not nearly be as good as his 2018 season.

	Name	Team	2018 xFIP	2019 Age	2019 FBV	2019 FBP	2019 xFIP	2020 xFIP (Predicted)	2021 xFIP (Predicted)	2022 xFIP (Predicted)
98	Lucas Giolito	White Sox	5.46	24	94.3 mph	55 %	3.66	4.13	3.93	4.01
101	Marco Gonzales	Mariners	3.59	27	88.9 mph	39.3 %	5.11	4.5	4.66	4.69

These two pitchers demonstrate how our model accounts for both seasons, but weighs the most recent data far more than two and three seasons ago. Ultimately, we were satisfied with our future projections, and we felt comfortable moving forward with this model.

Analysis

We tried many different combinations of inputs for our full model and submodels, including different age ranges, additional predictors, and including additional seasons from a player's career if they were available. We utilized all of our data where a player had two consecutive seasons of input to build our model and generate training error values, but we also utilized cross-validation where we divided our data into training and test data to measure our confidence in predicting future xFIP values. We experimented with various models and ultimately chose our current model because it generated low test error values, but additionally due to its simplicity. Some models that we tried performed slightly better in terms of test error rates, but we wanted to build a model that was easy to interpret and simple to build. Our final model only requires one year of a pitcher's age, fastball velocity, fastball percentage, and xFIP and can generate predictions for three future seasons of fastball velocity, fastball percentage, and xFIP values. We only use a few inputs for our model and each of our submodels only depend on the pitcher's age. We value our model's simplicity because a larger model would depend on many more submodels, increasing variability and making interpretation much more complicated.

When performing cross validation, we split our data (2002-2019) randomly into two equal portions: a training set and a testing set. Then, we created a model using the training data and generated predictions using this training model on the test data. We generated predictions of standardized xFIP and un-standardized these predicted values using the distribution of xFIP from the year of the prediction. This way, we generated xFIP predictions for each observation in our test set that should predict the actual xFIP values that were recorded during that season. We repeated this process 5000 times, with different training and testing sets and recorded the absolute value error of our predictions against the actual values.

	Year 1 Prediction	Year 2 Prediction	Year 3 Prediction
Train Absolute Mean Error	0.3398959	0.4423486	0.4583183
Test Absolute Mean Error	0.3516178	0.4445261	0.4783301

The table above displays our average residuals from using our training model to predict test data xFIP values. The above errors represent the average distance that the predictions are away from the actual xFIP values for each of the three seasons in the future. This means that on average, we would expect our test predictions for years 2020, 2021, and 2022 to be off by about 0.35, 0.44, and 0.48 from the actual xFIP values respectively. Since adding Models 2 and 3 and more years of data for each prediction, the error rates have reduced notably. We are pleased that

these values are not much higher than our training absolute error, which means that we would expect our model to perform almost as well when predicting future xFIP values than the existing xFIP values.

Model Interpretation

While we have several different models and submodels, we will interpret Model 2, which is the model that accounts for two years of data from the pitcher. Model 2's output table is below.

We can use the model to predict a pitcher's standardized xFIP in the upcoming year given their xFIP from the previous two years, their age, and their projected fastball velocity and fastball percentage values. The coefficient 0.484 for the lag xFIP means that for every one standard deviation unit increase in lagxFIP (from the previous season), we would expect the pitcher's xFIP for the upcoming season to increase by 0.484 standard units. From un-standardizing our predictions, we know that a standard unit in xFIP is approximately 0.60. This means that if a pitcher's lag xFIP value increased by 0.60, we would expect that pitcher's xFIP value during the upcoming season to increase by about 0.29.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.32454    0.18149  -1.788 0.073996 .
predicted_fbp1 -0.01451    0.02679  -0.541 0.588327
predicted_fbv1 -0.10241    0.03018  -3.394 0.000712 ***
lag_age         0.01407    0.00633   2.223 0.026432 *
lag_xfip        0.48464    0.03038  15.953  < 2e-16 ***
lag_xfip2       0.26639    0.02979   8.941  < 2e-16 ***
---
```

When looking at the coefficient value for lagxFIP2 (standardized xFIP from two seasons ago), we can see the trends that our model is predicated on. The most important predictor of a pitcher's xFIP is his xFIP from the previous season. The coefficient for the standardized xFIP from two years in the past is 0.266. This value is much smaller than the coefficient for the pitcher's standardized xFIP from the previous season (0.484), implying that if a pitcher's xFIP two seasons ago was 0.60 xFIP units (one standard unit) higher, we would expect that pitcher's xFIP in the upcoming season to increase by 0.16. This value is significantly less than the 0.29 increase in xFIP for each standard unit increase in lag xFIP. We can also look at the coefficient in Model 3, which includes three years of lag xFIP values. In this model, the coefficients for lagxFIP1 and lagxFIP2 are similar to Model 2, with values 0.465 and 0.264, respectively. The coefficient for lagxFIP3 is just 0.036, which is much smaller than either of the coefficients for a pitcher's lag xFIP value from the past two seasons. Once again, this coefficient value reinforces

the idea that a pitcher's xFIP is most likely to be influenced by a pitcher's immediate lag xFIP values from the past two years, with a smaller influence from their xFIP from three seasons ago.

When observing the remaining predictors in the model, age, fastball velocity, and fastball percentage remained significant, but have much smaller coefficients than the xFIP predictors. It should be noted that age was the only predictor that we did not standardize in our model. Age's coefficient of 0.012 means that as a player gets older by one year we would expect his upcoming season's xFIP to increase by 0.012 standard units. While this value is extremely small and has a small effect on the xFIP prediction value, it should be noted that age was also used as an important variable in our submodels for fastball velocity and fastball percentage. Additionally, the coefficient for fastball velocity was -0.109, which means for every one standard unit increase in fastball velocity, we would expect a pitcher's upcoming season's xFIP to decrease by 0.109 standard units of xFIP. A standard unit of fastball velocity is about 2.2 miles per hour, which means that for every 1 mile per hour that a pitcher increases his fastball velocity, we would expect his un-standardized xFIP prediction to decrease by approximately 0.03 xFIP units. This also means that given two pitchers of the same age, lag xFIP values, and fastball percentages, we would expect a pitcher whose average fastball velocity is 100 mph to have an xFIP that is 0.30 xFIP units lower than a pitcher whose average fastball velocity is 90 mph.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.317636   0.248970   1.276   0.2022
age_rangeprime            0.278367   0.295261   0.943   0.3459
age_rangeyoung            0.507589   0.523250   0.970   0.3322
lag_age                  -0.016490   0.007157  -2.304   0.0213 *
lag_fbv                   0.934858   0.010182  91.812   <2e-16 ***
age_rangeprime:lag_age   -0.009149   0.009195  -0.995   0.3199
age_rangeyoung:lag_age   -0.020353   0.021226  -0.959   0.3378
---
```

When looking at our submodel that predicts fastball velocity above, we were able to generate a model that accounts for about 88% of the variation in fastball velocity (adjusted r^2 of 0.8846). As should be expected, the most valuable predictor is the same pitcher's fastball velocity from a season ago. We included interaction terms between a player's age and their age group. We designated three age range classifications as pitchers 24 or younger were classified as *young*, pitchers between 25 and 31 were in their *prime*, and pitchers who are 32 or older were classified as *old.* Essentially this model looks at a pitcher's fastball velocity from the previous season and their age and adjusts the pitcher's predicted fastball velocity accordingly. From looking at our output and results of this submodel, *young* pitchers tend to maintain their fastball velocity from the previous season while pitchers in their *prime* have their fastball velocity decrease slightly. *Old* pitchers have their predicted fastball velocity decrease by large amounts, accelerating as pitchers approach 40 years old. The main takeaway from this submodel is that as

pitchers age, their fastball velocities decrease and the yearly drop in fastball velocity accelerates as pitchers get older.

The submodel for fastball percentage generated similar results, where a pitcher's fastball percentage is predominantly determined by their fastball percentage from the previous season. Younger pitchers tend to maintain their fastball percentage, but as pitchers grow older their fastball percentage decreases steadily. This relationship makes sense because as a pitcher ages, his fastball velocity decreases, forcing the pitcher to rely more on their off-speed pitches.

Conclusions

Through analyzing our model's predictions, we have been able to draw a number of conclusions about starting pitchers in baseball. An important inference we were able to draw is that the typical baseball pitcher's career follows an arc in which their performance steadily improves in their early years in the league, plateaus in their prime, around the 27 to 31 age range, and quickly declines in their later years, often even "falling off a cliff". Our model also provided a good, real example of the fact that while outliers certainly do exist, regression to the mean is much more prominent. Essentially, we should not expect pitchers to remain excellent for extended periods of time because it is really difficult to do. However, this also means that pitchers who can consistently maintain excellence should be perceived as extremely valuable.

In our analysis, we also discovered that one's performance from the past year is overwhelmingly more important than prior performance and any other potential predictor. This makes sense for most players, but it somewhat debunks the idea that a player just had a "down year" and will return to their previous form, as their down year seems to actually be more indicative of subsequent performance. Looking at our predictions for Gerrit Cole, who signed a nine-year $324 million contract this past offseason and served as part of the inspiration for our project, provides an interesting example of our model in action. Our model predicts Cole to continue his dominance and be the best pitcher in baseball for each of the next three years, even in 2022 when he will turn 32. This means that our model projects Cole to have the league's lowest xFIP value during each of the next three seasons. Though it is almost inevitable that Cole will fall off before the end of his contract, when he will be 38, even three years of Cy Young caliber performance could justify such a massive deal, especially if it pushes the Yankees over the hump to win a World Series.

Another conclusion we were able to draw is that fastball velocity shares a very strong relationship with performance. Fastball velocity was the most important predictor of performance behind the xFIP values included in the model, and our submodel defined a strong negative linear effect of age on the change in fastball velocity year to year. It is no surprise that as a player's physical skills wane, his average fastball velocity decreases. It is also clear in our analysis that harder throwing pitchers generally have better xFIP values. That said, it is interesting to see that as a pitcher ages, his velocity decreases, and xFIP tends to decline

alongside a declining fastball velocity. It is very possible that performance is influenced by a chain of cause and effect in this way. Even if that is not the case, our analysis shows that it is very important for teams to evaluate how the speed of a pitcher's fastball is changing, because it could very well have some indication for how they will perform through the future.

There were many variables we initially hypothesized would be important predictors in our models. This was not the case for most of them, however, and their unimportance provides for some additional conclusions. There are many relevant performance measures -- strikeout rate, ground ball percentage, fly ball percentage, WHIP, etc. -- each indicative of value in a certain area of the game. The fact that these variables are not important predictors of xFIP establishes confidence in xFIP as a powerful metric that is both descriptive of past performance and predictive of future value. We believe that because lag xFIP was included in our model and that xFIP mathematically includes outcomes a pitcher can control (strikeouts, walks, fly balls), predictors like strikeout rate and walk rate did not add much value to the model because it is redundant information that is already captured by xFIP.

From completing this project, we learned about the general trends in baseball over the past 20 years. As we expected, the strikeout and home run rates have grown rapidly, which has transformed the game of baseball into a completely new game today. Many players and pitchers who would have been valuable in the early 2000s would not have a job today. For example, a fly ball pitcher is a pitcher who relies on the batters to hit the ball in the air for outs, and the league consisted of a larger proportion of fly ball pitchers fifteen years ago. However, with home run rates surging, it is dangerous to be a fly ball pitcher in today's game. Currently, pitchers with high ground ball rates are a valuable asset in Major League Baseball because a pitcher with a 60% or 65% ground ball does an excellent job at keeping the ball out of the air. This means a smaller proportion of fly balls and lessens the risk of home runs.

Additionally, we also learned that predicting baseball is a very complicated concept. While we are able to estimate xFIP values with reasonable accuracy, there would need to be many more model adjustments to allow a baseball general manager to make decisions about player personnel. Teams have to assess the health of the athletes, the potential variability of their performance, and their importance to the franchise as a teammate and role model. If a pitcher is expected to have an extremely low xFIP but only stays healthy for half of the season, how should that pitcher compare to an average pitcher who has stayed healthy his entire career? General managers have to be able to assess the talent that is available in the league at each position, and the replacement level value of pitchers in each team's farm system or in free agency.

Next Steps

If we were to build upon our analysis further, there would be certain directions in which we would want to move forward. The baseball world is currently on hold due to the Coronavirus pandemic, but once baseball returns, it would be very valuable to evaluate our predictions for

2020 against the real season. We would have the opportunity to adjust our model appropriately. The game of baseball is constantly changing, and we expect our model to continuously evolve so that it can be as accurate as possible.

Our main inspiration for this project was the increasing size and length of contracts given to starting pitchers. Initially, we hoped to build predictive models of performance in order to infer a player's fair contract value. Players are offered contracts that guarantee a certain amount of money over many years, and it is important for teams to evaluate the expected performance of a pitcher throughout their entire contract. While a stellar pitcher like Gerrit Cole might be worth 36 million dollars (or even more) this year, his level of performance may steeply decline throughout his nine year contract. He will likely be worth well below 36 million dollars when he is 38 years old. It is important to weigh a player's expected future performance against the contract they are offered. Ultimately, as a next-step, we would want to evaluate the fairness of contracts based on the predictions our model makes. This would add a lot more nuance to the project because contracts are not just based on player performance. Other aspects such as past injuries, "star" value, and the signing team's championship hopes are surely large factors. That said, in order to accurately assess contract value, we would have to gather additional data and include these factors in our models.

Our App

After generating our predictions and tinkering with our model, the final step of our project was to find a valuable way to display our findings. We wanted something that was easy to use but also far beyond simply publishing our final dataframe online. Ultimately, we hoped our final deliverable would mimic a baseball card, where the user would be able to view basic information about each player, including our future predictions. Although we collectively had little experience with it, we decided to try to build a Shiny application, as we could develop and run the application in R where the remainder of our code was. Fortunately, it was relatively easy to get our dataframe of predictions to display in the app, as well as menus which allow the user to subset to a specific player or team. After this basic functionality, we began to brainstorm and implement many other features that boosted the usability and value of the app. We were able to scrape biographical information and a picture of every player from FanGraphs and mlb.com, and easily put these objects onto the player pages of the app. Additionally, we displayed graphs that showed past and predicted xFIP values to the player and team pages. One of the most interesting features that we added to the app was an xFIP calculator, where the user can plug in values and our model will give predictions for future xFIP based on the inputted variables. Ultimately, developing our Shiny app was a very smooth process. The seamless integration that it provides with R enabled us to easily upload our dataframes and create interesting graphs without having to build a website from the ground up. We are all extremely pleased with how the app turned out

from an aesthetic and usability standpoint, and think that people with limited or advanced baseball knowledge can gain valuable information from using it.