To Win A Cy Young

Jordan Despain and Jack Freeman

04/24/2023

Bob Lyur: The Best Pitching Coach To Ever Live

(Note: Our backstory is fiction and is meant for entertainment)

Abstract

We are interested in taking the stats for pitchers in the MLB from 1956 to 2022, and seeing if we are able to accurately predict the winners of the Cy Young award or at least give a probability of winning based of the stats we have for each pitcher in every year. The method we are taking is to create a fictional story with fictional high school pitchers and using a logistic regression model to predict if these high school pitchers are at a level capable of winning a Cy Young. If not, we want to create a plan to get them to a level where at least some of them have a high probability of winning the award. The model we created has, what we consider, a low error when we used it with a train/test split of the data. We used this model on the real data with MLB pitchers, and then we predicted probabilities on our fictional high school pitchers. Of course, none of them were at a level of a Cy Young award winner. However, we were able to come up with a improvement plan to work on stats over the next four years that get three of the five to a level that we consider "Cy Young ready". We wanted to have fun with this, so we hope the story is enjoyable.

Introduction

A potential tragedy is brewing over at Best Baseball High School, which is the best high school in the world when it comes to the sport of baseball. This high school is known for the fantastic players it produces, who more often than not go on to the major leagues and accomplish great things. The pitching coach for the baseball team of the last 45 years had to retire due to health concerns, and the incoming group of freshmen pitchers need someone new to help guide them to greatness. The team has managed to find an amazing replacement, or so they think. Bob Lyur has been hired to come in and take over the pitching program at Best Baseball High School. The group of incoming freshmen have incredible potential and they need someone who is able to unlock this potential and push them as far as possible in the four years they will be with the program. Just looking at some of their stats coming into the program makes it seem like a doable task for a good coach,

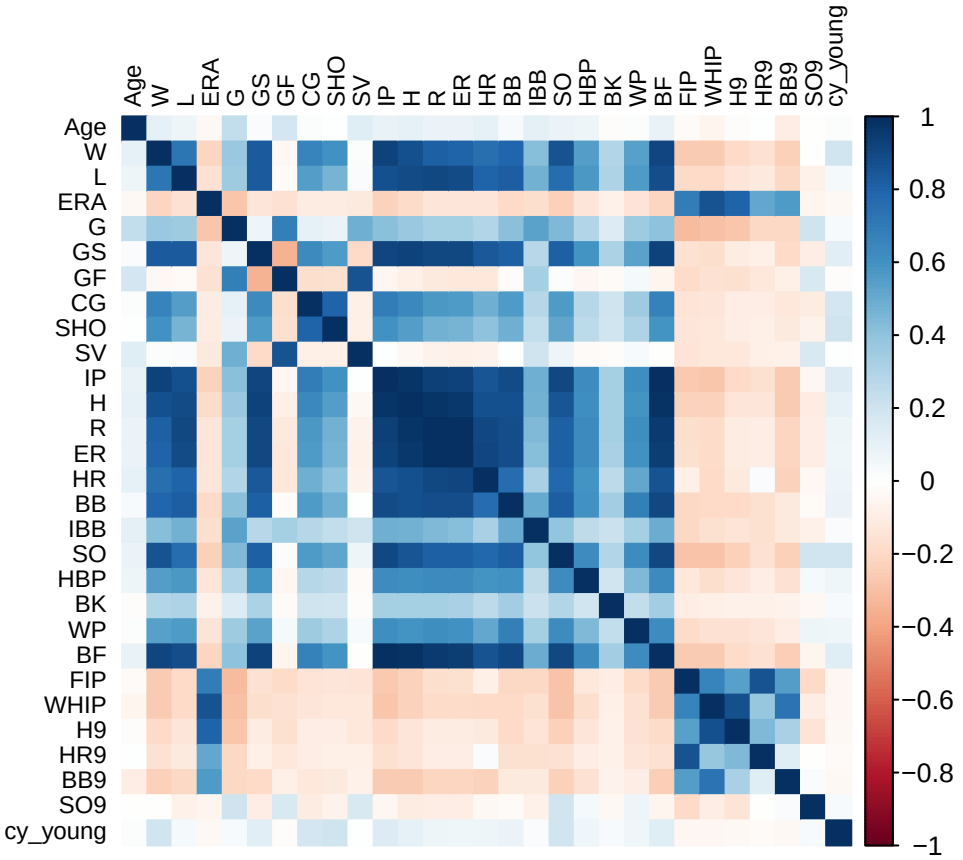
	Name	Age	W	L	ERA	GS	SHO	IP	H	R	ER	HR	BB	SO
1	Elijah Green	14	15	8	3.64	33	0	175.3	197	84	80	25	72	163
2	Druw Jones	15	13	11	4.13	33	0	184.2	221	90	88	29	75	145
3	Termarr Johnson	15	14	10	4.01	32	0	180.7	208	88	81	37	68	132
4	Cam Collier	14	14	9	3.54	32	0	180.6	219	76	75	30	59	170
5	Dylan Lesko	14	17	6	3.11	32	1	191.4	188	70	66	26	57	190

Bob Lyur seems like the perfect coach to guide these young pitchers as he has a very impressive track record and resume. The only problem is, he has lied about all his accomplishments and knows very little about what it takes to make these freshmen great pitchers. He has never been a coach for anything as a matter of fact. This is a big, big problem. The school wants things to carry on as they always have with this new coach. Every group of freshmen that have come in past have had at least one pitcher reach a goal of "Cy Young ready" by the time they graduate high school. "Cy Young ready" meaning, their stats are good enough to be a serious candidate for the Cy Young award. This goal has been a possible goal ever since high school baseball switched to 162 game seasons 15 years ago, the same number as the major league. This has made it possible to better compare stats from these high school prospects to those in the majors. Bob Lyur knows absolutely nothing about what it takes to win the Cy Young.

Bob soon realizes the mess he is inevitably going to create because there is no way he can get a single pitcher anywhere near this goal that the school is expecting to be accomplished. Bob Lyur is in panic mode trying to find a way he can make it out of this situation without being ousted as the fraud he is. He has come up with a plan. He knows two great data scientists who love baseball. Data is used to manipulate the sport in many different ways nowadays, so he believes surely it can help him get out of this pickle. Bob Lyur calls up these two data scientists and they agree to help, for a price of course. They believe this is a tall task, but are confident they can come up with a plan.

Methods

The two data scientists are tasked with coming up with a plan for this coach to improve the freshmen pitchers to "Cy Young ready" by the end of their high school career. They first need to collect data and try and figure out which stats are associated with improving the chances of winning a Cy Young award. They know the perfect place to look for this data, baseball-reference.com. The website baseball-reference keeps stats for just about everything in the world of baseball. They were able to get the stats from all pitchers who at least pitched one inning from 1956, when the Cy Young award was first given out, to 2022. The data had to be downloaded by each individual year, so part of their cleaning process was combining each year into one big data set. They also found all the Cy Young award winners throughout history and used this information to create a dummy variable of whether an individual won the Cy Young award for that given year. They believe their data is now ready to be used in a model. Before deciding on a model, the two scientists are curious about how the different variables included in the set are correlated with one another. A simple correlation plot gives them a great idea of each relationship amongst the variables.



The data scientists find a lot of these relationships make perfect sense, like hits and innings pitched having a very strong positive correlation. The more innings you pitch, the more hits you're likely to give up because you are facing more batters. However, they do find some of these very interesting. For example, home runs given up per 9 innings, HR9, having a negative relationship with home runs given up. You would think it would have a strong positive correlation given the number of home runs is in the numerator of the formula for the stat HR9. They believe this may have something to do with if you're giving up more home runs you may be pitching more innings, and innings pitched is the denominator in the formula. Nonetheless, the data scientists must remain focused on their task. So they start focusing on the model the want to build to hopefully save the legacy of Best Baseball High School and the reputation of Bob Lyur.

While considering which model is best fit for this data, the data scientists remember a time they recently used a logistic model. After a little discussion, they believe a logistic model is perfect for their task, because they want to predict *cy_young*, which is a dummy variable that only takes on the two values, 0 and 1. Also, this will allow them to determine a percent chance of winning the Cy Young award, which is exactly what is needed to make a claim that someone is "Cy Young ready". Since they are experts with the programming language R, they will use the glm function for this model. They played around with what variables to include in the model, using the correlation plot and their knowledge of baseball for some reference. The model they came up with is,

$$glm(cy\_young \sim . - Tm - Lg - Year - GF - GS - IBB - BK - BF, data = train, family = "binomial")$$

Now that they have decided on a model, they want to use a train/test split on the data to test the accuracy of their model. They create a for loop that splits the data randomly 25 different times. Each time through the loop, the model is used on the training data, predictions are made with the testing data, and the RMSE is then calculated. They then take the mean value of the 25 different RMSE values from their for loop. This is done to get a better idea of the true accuracy of their model. The find an RMSE value of 0.05079, which they are happy with considering their limited time due to the urgency of the issue at hand. This is about a 5 percentage point error on their predictions of winning the Cy Young award. Since they are happy with the accuracy of the model, they now run it on their full data set to understand how each variable effects the probability of winning a Cy Young award. The coefficients from their model are,

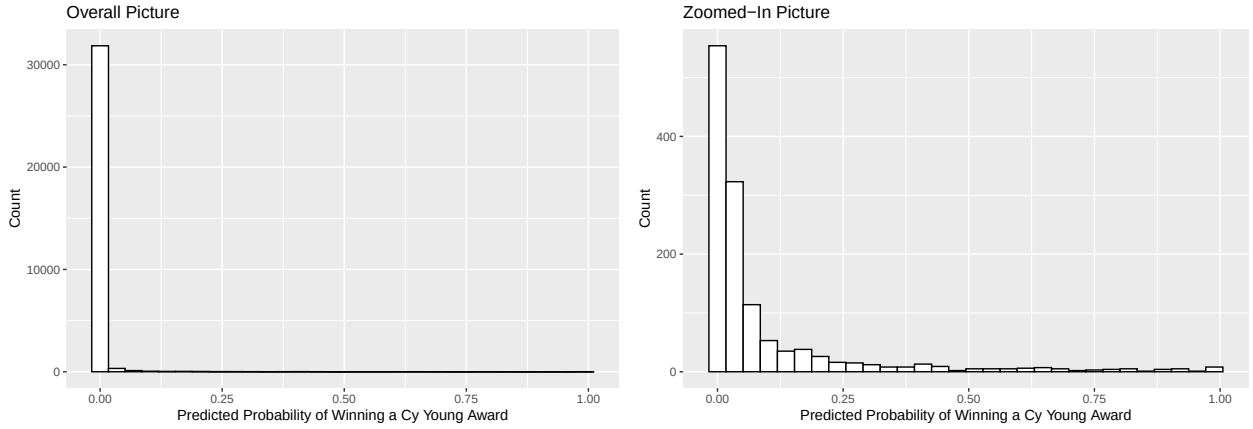
(Intercept)	Age	W	L	ERA
-13.6064393237	0.0249772007	0.4292769201	-0.2216048905	-0.9956380054
G	CG	SHO	SV	IP
-0.0378278377	-0.0871422599	0.0089113949	0.1523740833	0.0345597141
H	R	ER	HR	BB
-0.0051818253	-0.0026021499	-0.0008105601	-0.0821902734	0.0111809163
SO	HBP	WP	FIP	WHIP
0.0047842981	-0.0683079875	-0.0643785450	0.9241862002	-23.2337665831
H9	HR9	BB9	SO9	
2.4966435670	-0.8018882167	1.8299419478	0.3959905863	

Taking the exponent of the coefficient gives the expected increased odds on winning the Cy Young award. For example, an additional win is expected to multiply the odds of winning a Cy Young by $exp(0.4292769201) \approx 1.536$. The data scientists find some of these surprising, like the large positive coefficient on BB9. They want to look at the full summary of the results to better understand what is going on here.

```
Call:
glm(formula = cy_young ~ . - Tm - Lg - Year - GF - GS - IBB -
    BK - BF, family = "binomial", data = model_data %>% select(-c(Name)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0686  -0.0058  -0.0014  -0.0004   3.8411

Coefficients:
               Estimate  Std. Error z value           Pr(>|z|)
(Intercept) -13.6064393   3.9419343  -3.452           0.000557 ***
Age           0.0249772   0.0298179   0.838           0.402223
W             0.4292769   0.0564880   7.599 0.0000000000000297 ***
L            -0.2216049   0.0630387  -3.515           0.000439 ***
ERA          -0.9956380   0.7168723  -1.389           0.164874
G            -0.0378278   0.0315618  -1.199           0.230709
CG           -0.0871423   0.0348371  -2.501           0.012370 *
SHO           0.0089114   0.0836444   0.107           0.915155
SV            0.1523741   0.0402138   3.789           0.000151 ***
IP            0.0345597   0.0161113   2.145           0.031949 *
H            -0.0051818   0.0138006  -0.375           0.707306
R            -0.0026021   0.0323431  -0.080           0.935876
ER           -0.0008106   0.0464917  -0.017           0.986090
HR           -0.0821903   0.0368995  -2.227           0.025920 *
BB            0.0111809   0.0185665   0.602           0.547036
SO            0.0047843   0.0062432   0.766           0.443485
HBP          -0.0683080   0.0392312  -1.741           0.081655 .
WP           -0.0643785   0.0387704  -1.661           0.096812 .
FIP           0.9241862   0.3712234   2.490           0.012790 *
WHIP        -23.2337666  26.6811225  -0.871           0.383867
H9            2.4966436   2.9632394   0.843           0.399487
HR9          -0.8018882   0.9154150  -0.876           0.381039
BB9           1.8299419   2.9527444   0.620           0.535427
SO9           0.3959906   0.1945888   2.035           0.041850 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1629.28  on 32593  degrees of freedom
Residual deviance:  548.52  on 32570  degrees of freedom
AIC: 596.52

Number of Fisher Scoring iterations: 19
```

Just as they suspect, the coefficients that surprised them are not statistically significant. They now want to move their focus towards the high school pitchers and continue to work towards completing the task they were hired to do.

Satisfied with their model, they use it to predict the probability of winning the Cy Young award on the full data set of all the pitchers that pitched at least one inning from 1956 to 2022. They really like what they see with their predictions. The average predicted probability is 0.0038044, which is a good sign considering only a very tiny number of pitchers in this data have won the award. They look at the distribution of the predicted values to get a better understanding of what they are working with. They look at the overall picture and a more zoomed in picture that discludes probabilities below 0.005. The reason for this is there are so many really small values, it makes it impossible to see whats going on with any predicted values above 0.05.



They notice there are very few predicted probabilities that are more than 0.1. They filter the data to Cy Young winners with predicted probabilities of at least 0.1, and find 93 of the 124 Cy Young winners have a predicted probability of at least this value. They also find only 178 of the 32470 non-winners have a predicted probability above 0.1. To see if this is a potential issue, they look further into these pitchers and found almost all of them received votes for the award in the respective year. Also, the pitchers that did not receive votes played on multiple teams that year, most resulting in splitting their time between the two leagues. This makes sense as to why they did not receive votes, because for most of the Cy Young award's history, the award is given to someone in both leagues. If your stats are split across the two leagues, you won't look as impressive when looking at the stats for either individual league. So, they really like 0.1 as a probability goal to set for the high school pitchers. They believe a predicted probability of at least 0.1 is a good value to say that pitcher is "Cy Young ready".

Results

Now, they want to use the high schooler's stats with the model to get an idea of where they currently are in regards to being "Cy Young ready". The create predicted probabilities for each of the five pitchers.

```
             Name predicted_probs
1    Elijah Green   0.00013021341
2      Druw Jones   0.00001362983
3 Termarr Johnson   0.00001166585
4     Cam Collier   0.00012104394
5     Dylan Lesko   0.00357896355
```

They find there's definitely some work to be done to get them to the goal of a predicted probability of 0.1. Using the model's results and the correlation plot from before, they work to put together a four-year plan that will get at least the majority of the pitchers past the threshold.

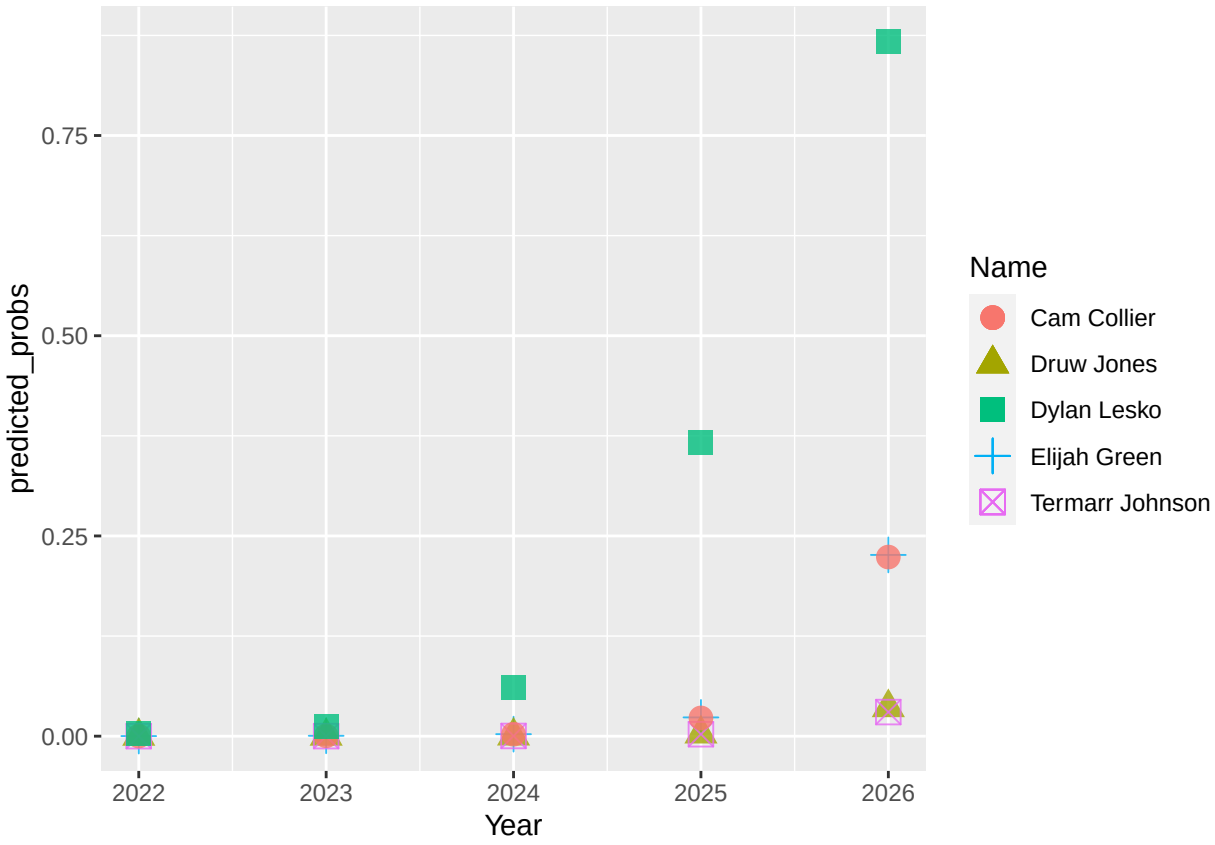
The Plan

Stat	Year 1	Year 2	Year 3	Year 4
Wins	+1	+1	+2	+2
Losses	-1	-1	-1	-1
ERA	-5%	-10%	-15%	-20%
Hits	-5%	-5%	-5%	-10%
Runs	-5%	-5%	-5%	-10%
Earned Runs	-5%	-10%	-10%	-10%
Home Runs	-5%	-10%	-10%	-10%
Strike Outs	+5%	+10%	+10%	+10%
Walks	-5%	-10%	-5%	-10%
Hit Batters	-1	-1	-1	-2
Wild Pitches	-1	-1	-2	-2

If Bob Lyur manages to stick to this plan and somehow improve each pitcher's stats accordingly, the data scientists are interested if this will be enough to get some of these pitchers above the goal they set. To see this, they rerun to model to see the predicted probabilities of the pitchers' chances of winning the Cy Young award. They get results with the pitchers' stats if Bob Lyur is able to successfully get each of them through the plan every year and make the set improvements.

```
             Name predicted_probs
1    Elijah Green      0.22630696
2      Druw Jones      0.03570142
3 Termarr Johnson      0.03008447
4     Cam Collier      0.22366311
5     Dylan Lesko      0.86758620
```

It's a success! The data scientists are happy with the results and eager to get the plan to Bob Lyur so he can begin putting it into action. Along with the plan the data scientists also put together a plot so display the progress each pitcher makes over the four years.



Their job is done, and all that's left is to hope Bob Lyur can pull this off the save the reputation of the school!

Conclusion

We believe a logistic model is appropriate for this setting as we feel very satisfied with the results we got. There are definitely a lot more advanced statistics out there that could probably increase the accuracy even further, but we think with the time we have this is at least a good starting point. A main lesson to take away from this report is that logistic models can be very useful when trying to predict the probability of success or an event occurring. The correlation plot was a great way to understand how the variables tend to interact with each other. We hope this was as interesting to read as it was to create.