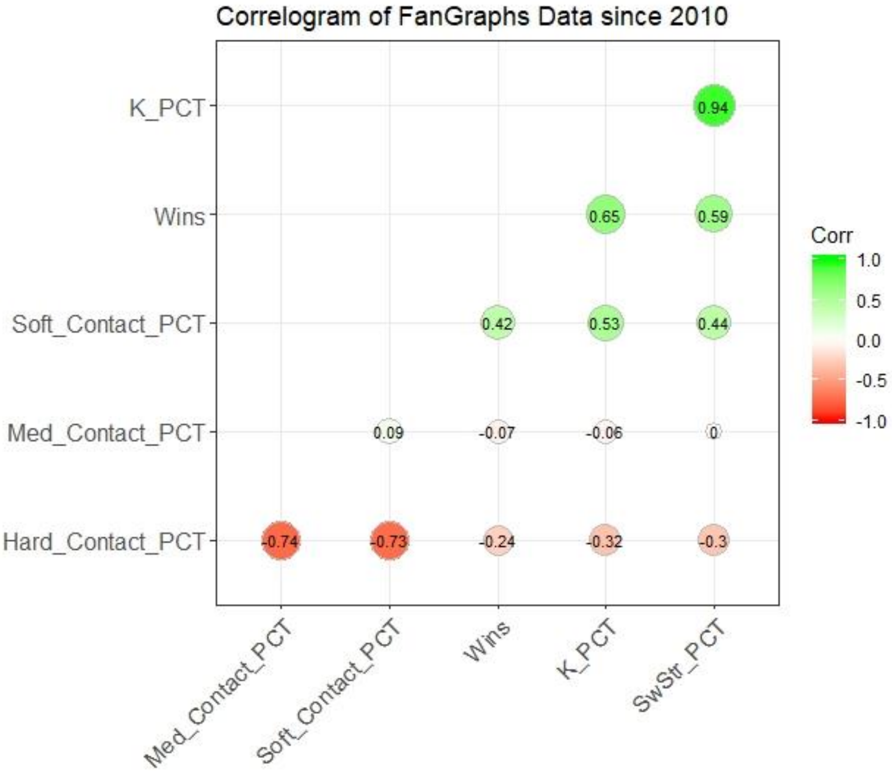
PITCHING LEADERBOARDS

Contents

When given the dataset, I asked, what has a stronger relationship to Wins from pitching staffs, whiffs or type of contact? Whichever had the stronger relationship would be the basis for my leaderboard analysis.

Below is a correlogram of FanGraphs data from 2010 to this past season. The variables include Wins, Strikeout Percentage, Swinging Strike Percentage, Soft Contact Percentage, Medium Contact Percentage, and Hard Contact Percentage. Looking at the circled data, Swinging Strike (i.e. whiffs) Percentage have a stronger relationship to Wins than Soft Contact Percentage. Strikeout Percentage even has a strong relationship to wins.



Now that we see that Swinging Strike Percentage and Strikeout Percentage have a stronger relationship to Wins that Soft Contact, we can now create a model that predicts the probability of a whiff given important pitch characteristics from Statcast data in 2019.

The original dataset of 729,793 observations and 91 variables are reduced to 713,034 observations and 8 variables after removing missing values and matching the columns in the TrackMan data.

Those columns are:

pitch_type – the given pitch type from Statcast

release_speed – which is the pitch velocity (in Miles Per Hour)

p_throws – the throwing side of the pitcher

pfx_x – the horizontal movement in feet from the catcher's perspective

pfx_z – the vertical movement in feet from the catcher's perspective

release_spin_rate – the spin rate of the pitch (in Revolutions per Minute)

whiff – '1' if the pitch ended if a whiff, '0' if the pitch did not

Given the pitch types, we must classify them to match the college TrackMan data.

Four-Seam Fastballs, Cutters, Two-Seam Fastballs, and Sinkers are classified as Fastballs, or FB.

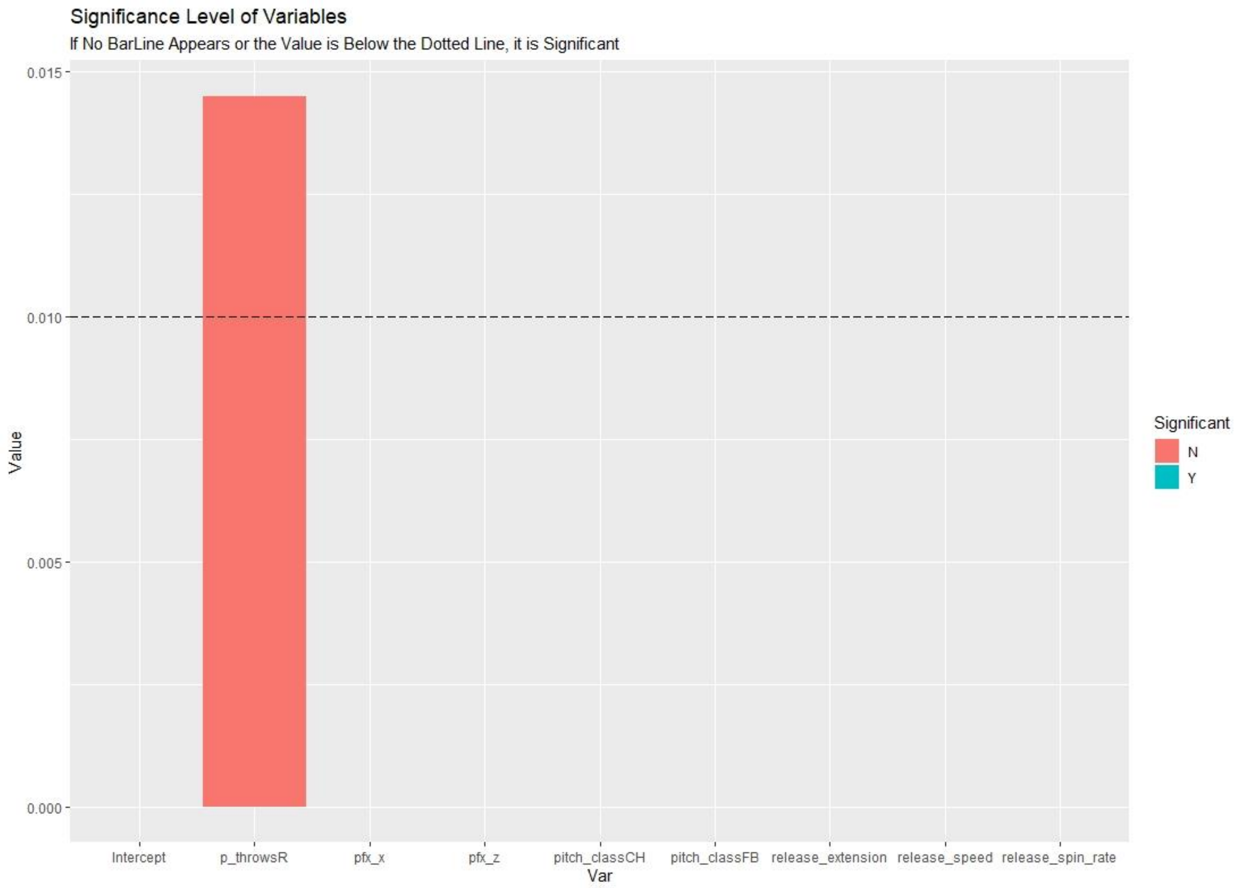
Curveballs, Sliders, Knuckle Curves, Eephuses, and Knuckle Balls, are classified as Breaking Balls, or BRK.

Changeups, Split Fingers, or Forkballs are classified as Changeups, or CH.

Now, we turn to what variables above are crucial in determining a swing and miss.

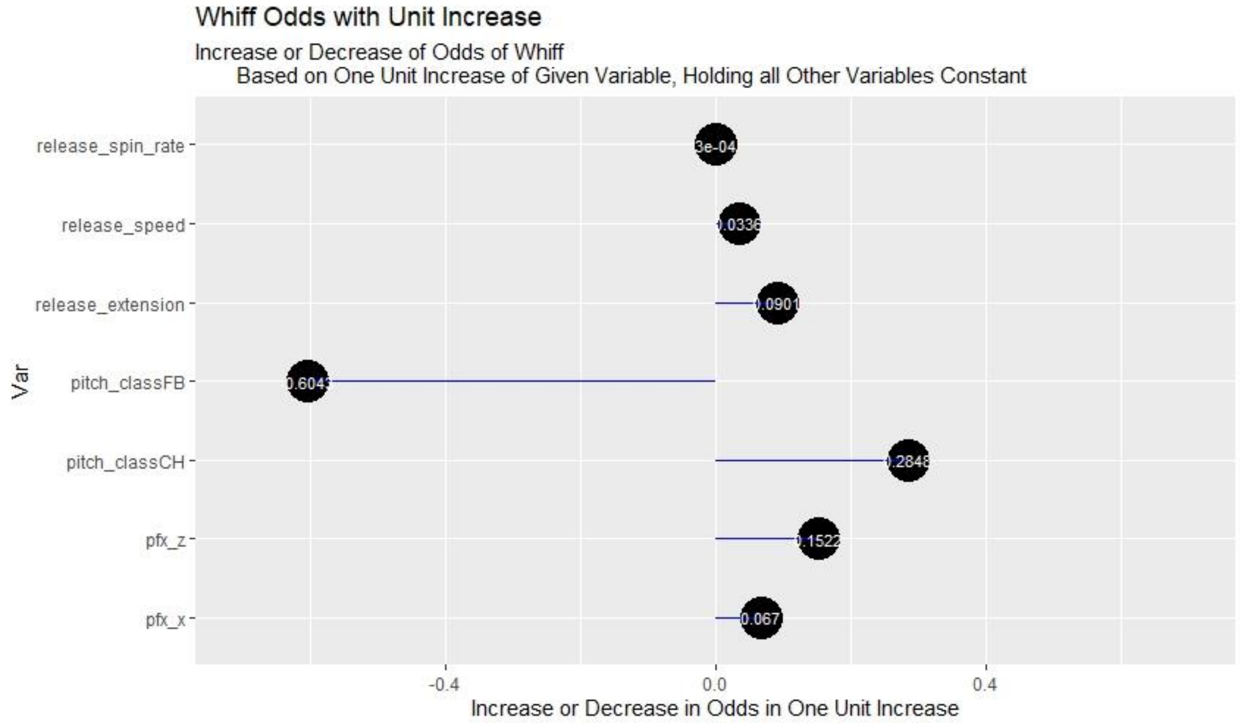
That's where a generalized linear model (glm) comes in to determine what variables carry significance to the model. To achieve a model's performance, the Statcast Data will be split into training data and testing data, as well as splitting the whiffs and no whiffs due to the large proportion of non-whiffs.

Using a 99 percent confidence interval, we determine what variables mentioned above are statistically significant. For the purpose of the model, p_throwsL (LH pitcher) and pitch_typeBRK (breaking ball) are used as a reference and not included in the output. According to the graph, all variables except p_throws are statistically significant, which we will remove from the model and analyze once again.



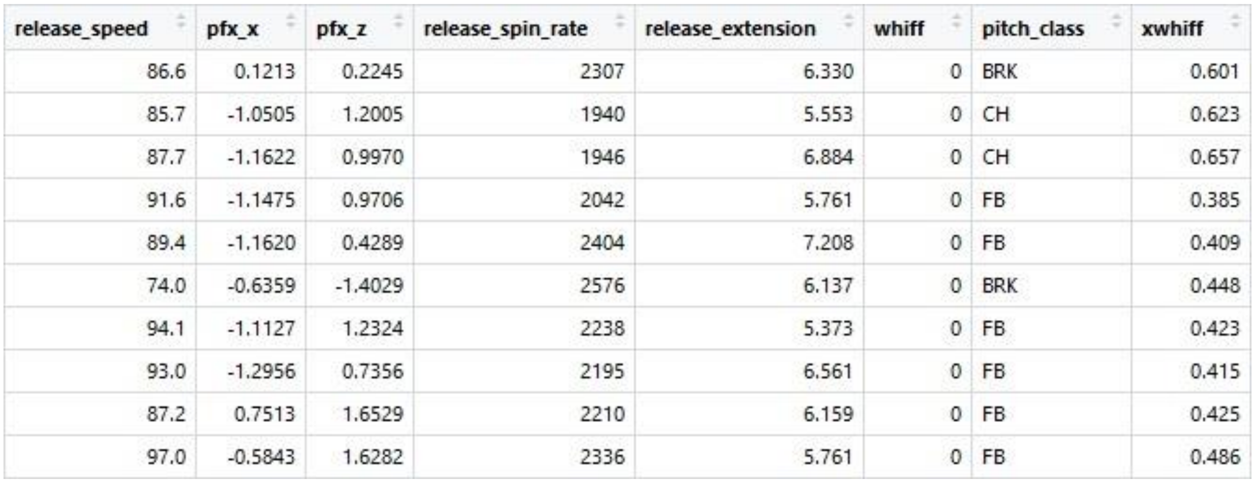
After removing the p_throws column, we run our model again and check for significance, in which all variables this time are significant. Now, we view the odds ratio, which is a way of saying how much does an increase of one unit (based upon variable) increase or decrease the odds of a swing and miss?

The graph below shows the odds in terms of percentages, listed as a decimal. For example, a one MPH increase in pitch velocity, holding all other variables constant, will increase the chance of a swing and miss by 3.36 percent. What stands out is throwing a Fastball. An increase in throwing a fastball, i.e. no spin, no movement, no velocity, with decrease the odds of a whiff by over 60 percent. A 1 rpm increase in spin rate, holding other variables constant, leads to an increase in the chance of a whiff by .0003 percent, which is the same as saying a 100 rpm increase in spin rate increases the chance of a whiff by 3 percent, so on and so forth.



We can now run a model that gives the predicted probability of a whiff given the above pitch characteristics.

Below is a data table showing 10 observations of various pitches and their respective whiff probabilities.

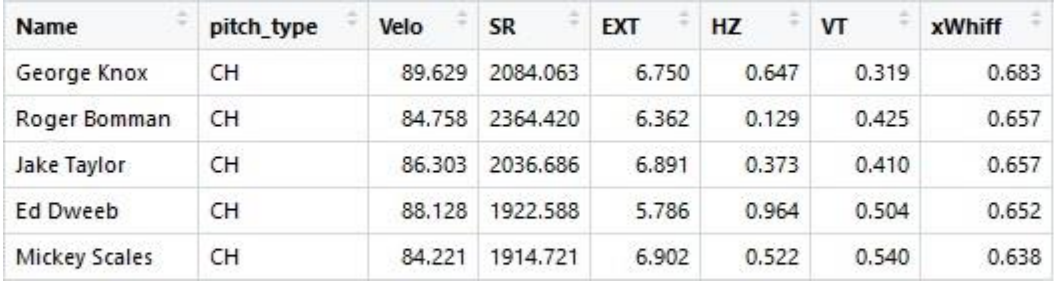


release_speed	pfx_x	pfx_z	release_spin_rate	release_extension	whiff	pitch_class	xwhiff
86.6	0.1213	0.2245	2307	6.330	0	BRK	0.601
85.7	-1.0505	1.2005	1940	5.553	0	CH	0.623
87.7	-1.1622	0.9970	1946	6.884	0	CH	0.657
91.6	-1.1475	0.9706	2042	5.761	0	FB	0.385
89.4	-1.1620	0.4289	2404	7.208	0	FB	0.409
74.0	-0.6359	-1.4029	2576	6.137	0	BRK	0.448
94.1	-1.1127	1.2324	2238	5.373	0	FB	0.423
93.0	-1.2956	0.7356	2195	6.561	0	FB	0.415
87.2	0.7513	1.6529	2210	6.159	0	FB	0.425
97.0	-0.5843	1.6282	2336	5.761	0	FB	0.486

In the last column, ‘xwhiff’, we are given the probability of a whiff with the other characteristics.

Moving to the dataset, we now can predict the probability of a whiff with the pitchers’ data. One variable that was altered was the brk_x and brk_z, horizontal and vertical movement in the TrackMan dataset. The TrackMan movement was calculated in inches, whereas the Baseball Savant dataset was calculated in feet. To correct this, the TrackMan data was recalculated to match Savant’s data.

After cleaning the data like that of Savant’s, we now group pitch types by player name and determine their whiff probability if that pitch’s average rates were thrown. Here are the Top 5 pitchers by pitch type, sorted by whiff probability.

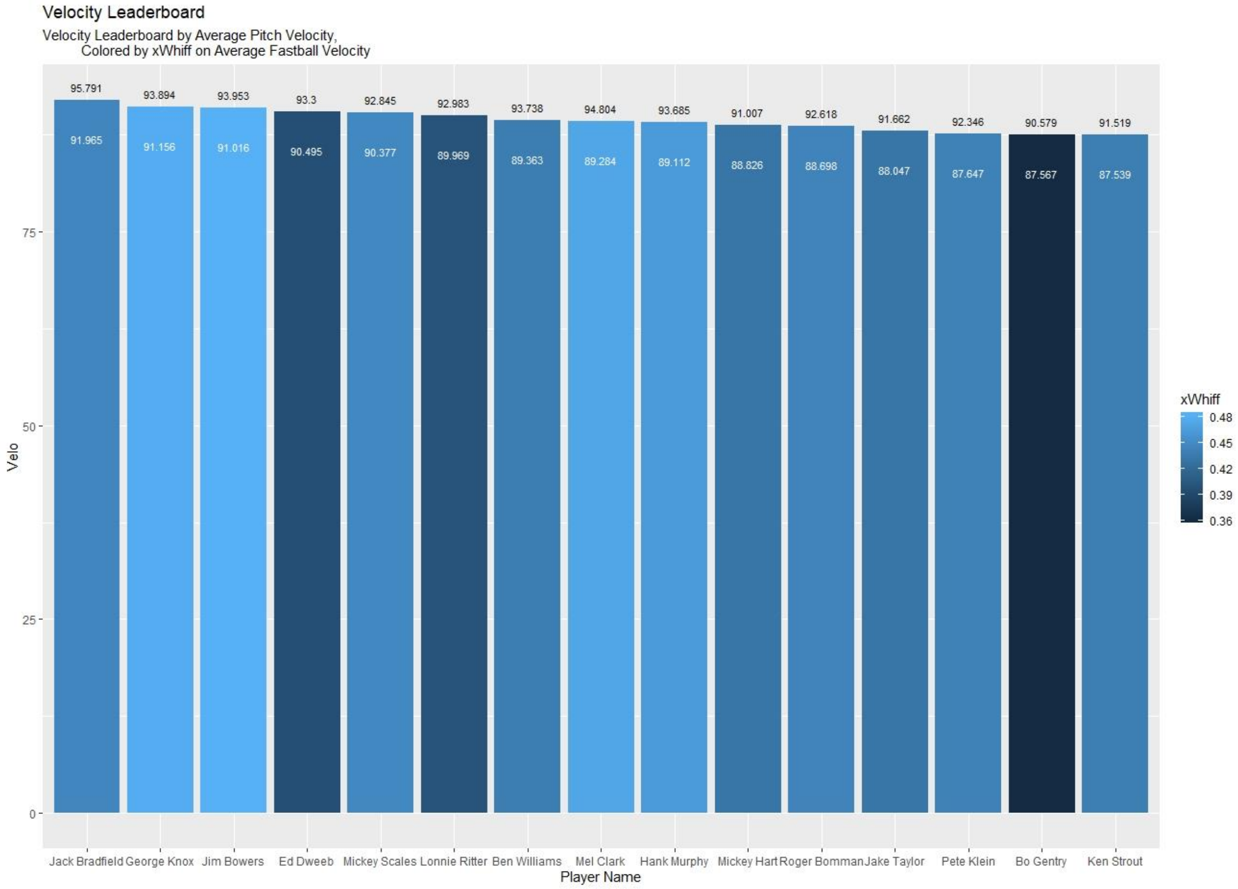


Name	pitch_type	Velo	SR	EXT	HZ	VT	xWhiff
George Knox	CH	89.629	2084.063	6.750	0.647	0.319	0.683
Roger Bomman	CH	84.758	2364.420	6.362	0.129	0.425	0.657
Jake Taylor	CH	86.303	2036.686	6.891	0.373	0.410	0.657
Ed Dweeb	CH	88.128	1922.588	5.786	0.964	0.504	0.652
Mickey Scales	CH	84.221	1914.721	6.902	0.522	0.540	0.638

Changeups play a huge role in Whiff Probability, as high-velocity, high-spin, closer extension to home plate, and a positive horizontal and vertical movement garners a large whiff probability.

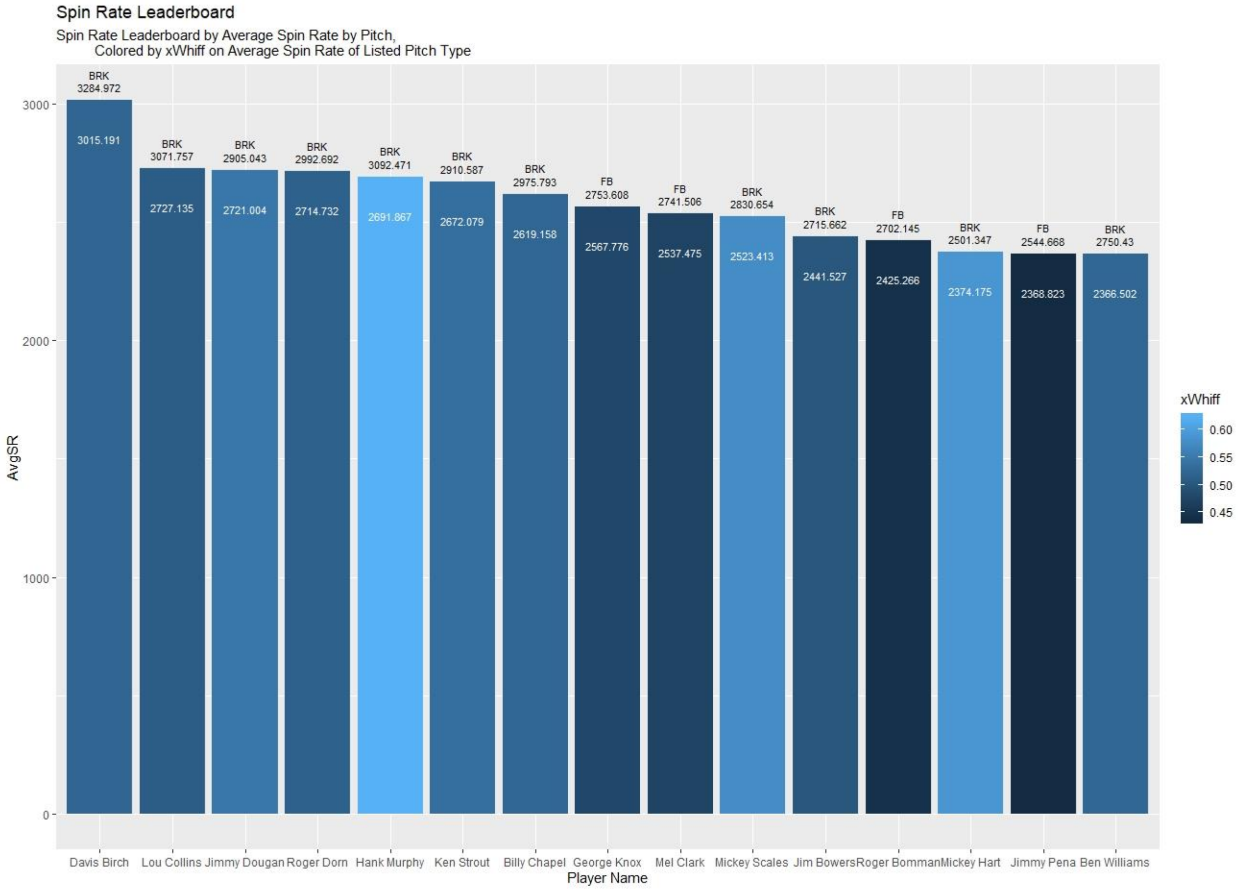
Now, given the model and its characteristics, I present the leaderboard for each of velocity, spin rate, extension and overall with the whiff probability being the driving factor.

Velocity Leaderboard:



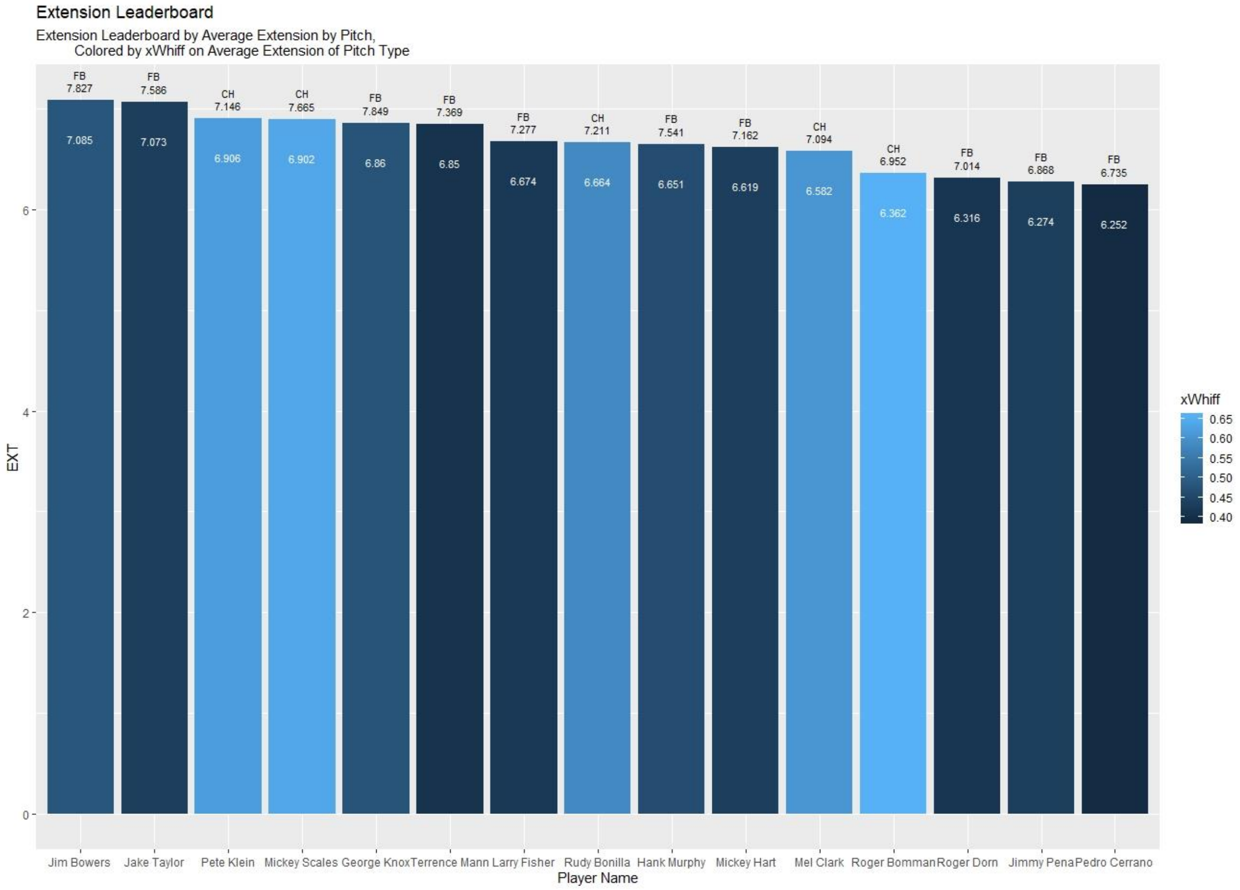
This graph shows the top 15 pitchers in average velocity (the white number inside the bar), colored by whiff probability of the average pitch velocity of all pitches. The number above the bar is the average Fastball Velocity of the pitcher. For the most part, the top pitchers who throw harder across all pitches have a greater chance of generating a whiff than those who don’t throw as hard, given averages of other features like spin rate and movement. Although Jack Bradford may throw the hardest, George Knox and Jim Bowers can generate more whiffs.

Spin Rate Leaderboard:



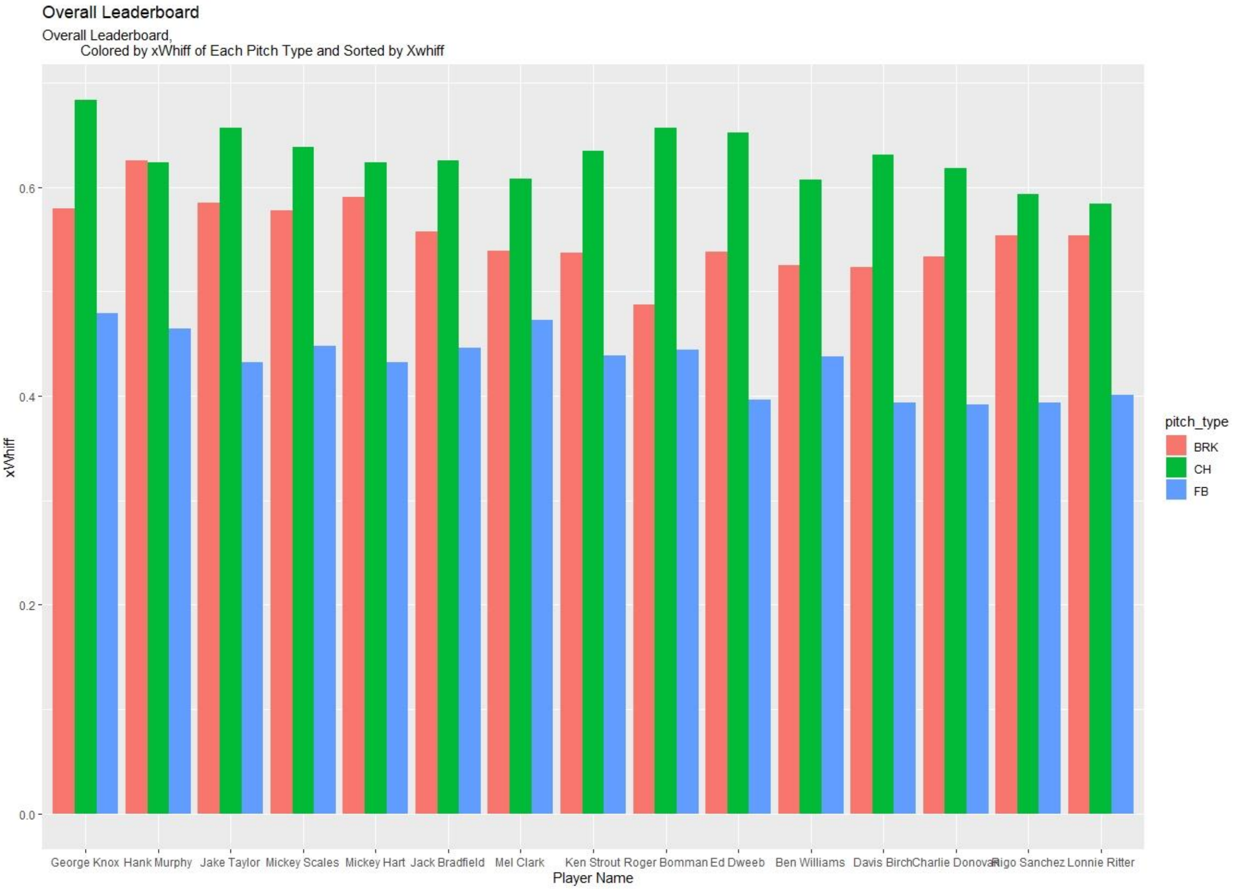
The Spin Rate Leaderboard displays the average spin rate of the top 15 pitchers' highest spin pitches inside the bar. Each pitcher in the leaderboard has their highest average spin pitch shown, as well as the max spin rate of that pitch above the bar. Per the graph, Spin Rate is varied in xWhiffs given average characteristics of other features as mentioned previously above. Davis Birch has the highest spin rate on his breaking ball, but not the highest xWhiff.

Extension Leaderboard:

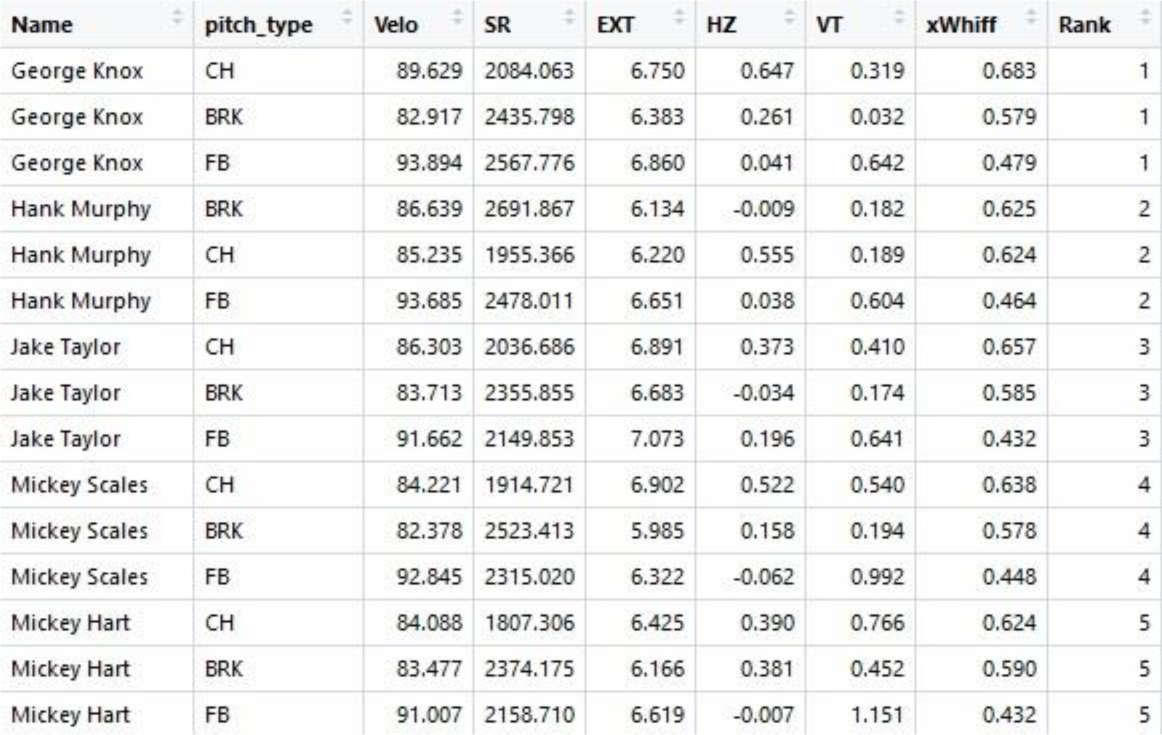


The Extension Leaderboard, like the Spin Rate Leaderboard, displays the average extension of the top 15 pitchers' greatest extension pitches inside the bar. Each pitcher in the leaderboard has their highest extension pitch shown, as well as the max extension of that pitch above the bar. As for extension on whiff probability, given average characteristics of other features, it is once again varied, however changeups have a higher whiff probability than a fastball. The consensus for fastballs is that, as a pitcher increases its extension, so does the whiff probability.

Overall Leaderboard:



The overall leaderboard displays the Top 15 pitchers, sorted by average whiff probability of all pitches and each pitch’s whiff probability by color. Since changeups have such an importance in whiff probability from my model, they have the highest whiff probability. Look at Hank Murphy, however. His breaking ball has a higher whiff probability than his other pitches, making it a valuable pitch and a 2nd place ranking on my leaderboard. Below is a table that displays the top 5 pitchers above and lists each of their average Velo, Spin Rate, Extension, Horizontal, and Vertical Movement by each pitch type, as well as the predicted whiff probability.



Name	pitch_type	Velo	SR	EXT	HZ	VT	xWhiff	Rank
George Knox	CH	89.629	2084.063	6.750	0.647	0.319	0.683	1
George Knox	BRK	82.917	2435.798	6.383	0.261	0.032	0.579	1
George Knox	FB	93.894	2567.776	6.860	0.041	0.642	0.479	1
Hank Murphy	BRK	86.639	2691.867	6.134	-0.009	0.182	0.625	2
Hank Murphy	CH	85.235	1955.366	6.220	0.555	0.189	0.624	2
Hank Murphy	FB	93.685	2478.011	6.651	0.038	0.604	0.464	2
Jake Taylor	CH	86.303	2036.686	6.891	0.373	0.410	0.657	3
Jake Taylor	BRK	83.713	2355.855	6.683	-0.034	0.174	0.585	3
Jake Taylor	FB	91.662	2149.853	7.073	0.196	0.641	0.432	3
Mickey Scales	CH	84.221	1914.721	6.902	0.522	0.540	0.638	4
Mickey Scales	BRK	82.378	2523.413	5.985	0.158	0.194	0.578	4
Mickey Scales	FB	92.845	2315.020	6.322	-0.062	0.992	0.448	4
Mickey Hart	CH	84.088	1807.306	6.425	0.390	0.766	0.624	5
Mickey Hart	BRK	83.477	2374.175	6.166	0.381	0.452	0.590	5
Mickey Hart	FB	91.007	2158.710	6.619	-0.007	1.151	0.432	5

There is a strong relationship from Whiff Rate to Wins. Using Statcast Data from 2019, the model predicts the probability of a whiff, by given characteristics like Spin Rate, Pitch Type, Velocity, Extension, Horizontal Movement, and Vertical Movement. Off-speed pitches are valuable in generating swings and misses (like changeups), but not without the increase in the other characteristics as well.

The Top 5 pitchers in the upcoming draft based upon this information are George Knox, Hank Murphy, Jake Taylor, Mickey Scales, and Mickey Hart.