Introduction:

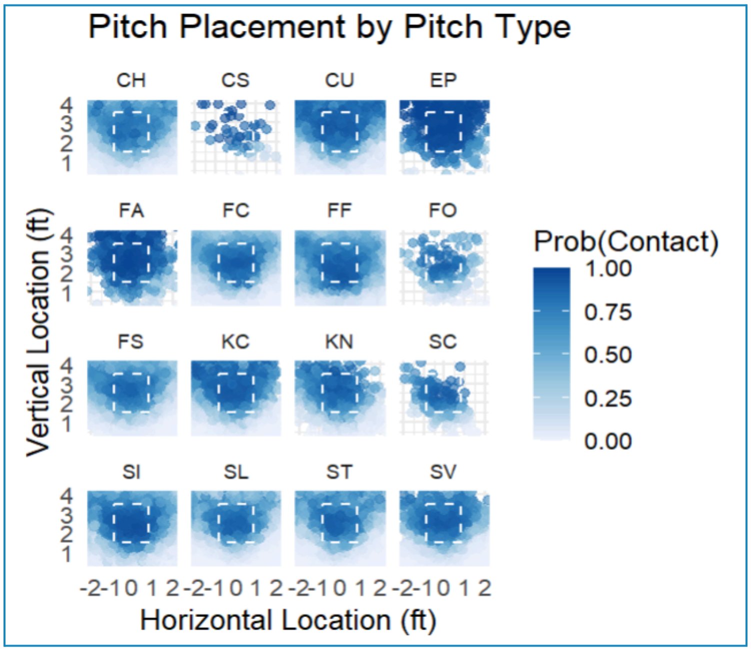
The purpose of this report is to analyze pitches from the ongoing 2025 MLB season with the intention of predicting contact probability given that a player swings at the pitch. We want to identify potential predictors for high contact probability, and inversely predictors for low contact probability. Low contact means a high whiff rate, and if leveraged appropriately in high-swing-probability situations, could help MLB pitchers gain an advantage, and can help identify pitchers who are above or below average at outperforming contact likelihood on their pitches.

Part 1: Modeling

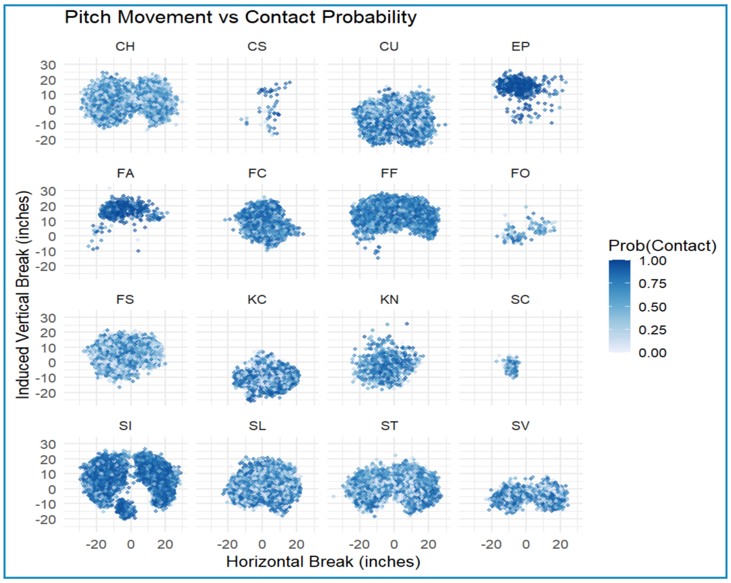
In the base model, we used current ball-strike count, pitch location, and pitch speed as predictors of contact probability. In my improved model, I chose to add in additional variables as predictors. I first chose horizontal break, with the expectation that more movement reduces contact probability. I also chose induced vertical break for the same reason, and chose to use induced as well as total vertical break because that better shows the pitcher's effect on the ball as opposed to gravity's, so I felt it would be important to show that as a predictor. I then added pitcher extension, with the expectation that the later a pitcher releases the ball, the less time a batter has to diagnose the pitch type, thus resulting in lower contact rates. Lastly, I used pitch type, with the assumption that certain pitches are just naturally more prone to whiffs than others, such as splitters and sliders as opposed to fastballs. Additionally, when later using a gradient boosting model, it will be able to pick up on the nuances of different pitch types and what locations are best for them.

I first fit a logistic regression model using these new features, as well as fitting an XGBoost model. The results of these models indicate that the features used in the original are very likely to be the most important features available. By a pretty large margin, pitch location (especially height) was the most significant feature. Vertical break was the next most important feature, with horizontal break not too far behind. Ball-Strike count was next, with release speed not too far behind. In last was extension and pitch type. The only pitch type that was more significant than the other features used was if a pitch was a sinker, likely due to it having the lowest whiff rate in the league.

With location and movement being the two most important features, I wanted to visualize these based on each pitch. Seen on the next page is a visual depicting pitch location and contact probability for each of 16 different pitch types. Seen below is horizontal break and induced vertical break as well as contact probability. The first



visual helps to demonstrate how for different pitch types, certain locations are more or less prone to whiffs. The same goes for the movement of a baseball, with certain breaks being more or less favorable. For instance, four-seam fastballs are one of the most hittable pitches, being the second highest contact rate (**81.44%**) among common pitches (n > 500 thrown) and are one of the few pitches that see high contact rates below the zone, as a result of its more predictable movement. Looking at movement, it's still evident that certain pitches are more hittable than others, and that certain locations for a given pitch are more or less favorable than others. I chose to map induced vertical break instead of overall vertical break to account for the pitcher's role in the movement.



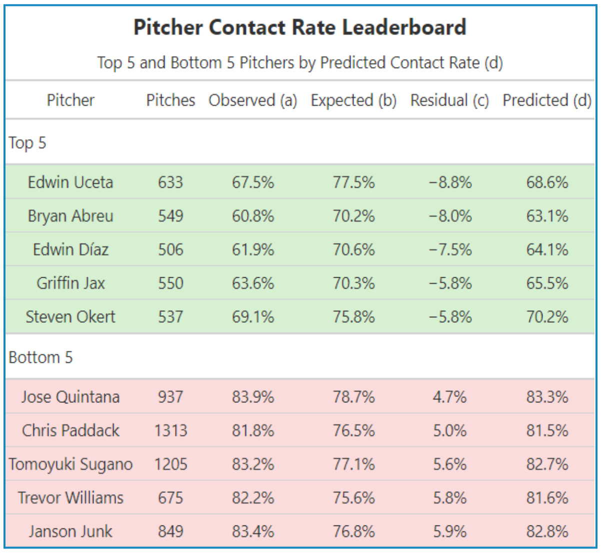
For instance, fastballs that stay up, or "rise" in the zone without much horizontal movement are pretty easy to hit compared to the few that drop. Inversely, for pitches like a curveball or knuckle-curve, the ones that drop too much are more hittable than the ones that don't see too much of an induced vertical break.

While the model is capable of predicting contact probability from a limited number of features, it does have its limits. For instance, certain pitchers are skilled at making pitches break super late in their approach, making them much more prone to whiffs. This is not able to be captured by the model, because it only has beginning and end, as opposed to frame by frame data. Additionally, it does not account for greater context, such as score, sequencing, tells, etc., causing potentially lower or higher predictions than what may actually occur.

Part 2: Validation

Now, I aim to determine how effective these models are, and to see if my implementation is better than the baseline one that was applied at first. Using an 80/20 split to perform out-of-sample evaluation, I calculated model effectiveness using negative log loss. With this, I got the result that I should have, which was that the original model performed well, but both of the other models I made performed at least a bit better. The improved GLM reduced log loss relative to the baseline GLM (**.449** vs. **.455**), indicating a modest gain from incorporating movement and pitch-type features. The XGBoost model achieved the lowest log loss (**.436**), suggesting that nonlinear interactions such as pitch-type-specific location effects provide additional predictive value beyond what a generalized linear model can capture. My takeaway from this is that the original model already captured many of the most important features, but using additional, more specific ones does allow for improved performance in capturing pitched ball contact probabilities.

Part 3: Player Evaluation

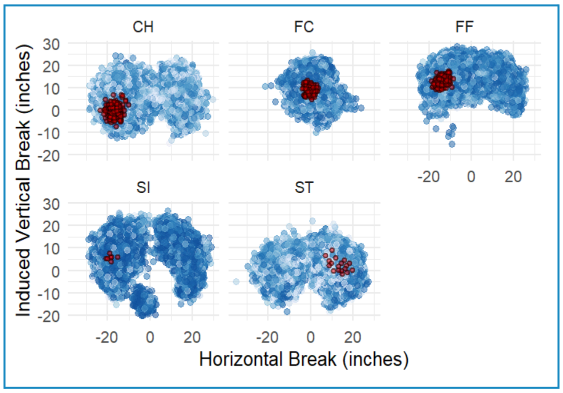


Pitcher Contact Rate Leaderboard

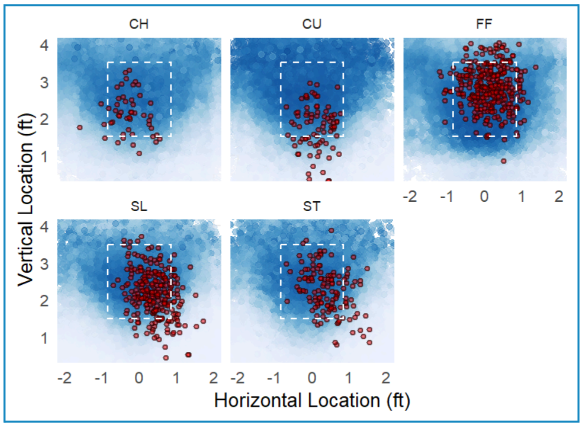
Top 5 and Bottom 5 Pitchers by Predicted Contact Rate (d)

Pitcher	Pitches	Observed (a)	Expected (b)	Residual (c)	Predicted (d)
Top 5					
Edwin Uceta	633	67.5%	77.5%	−8.8%	68.6%
Bryan Abreu	549	60.8%	70.2%	−8.0%	63.1%
Edwin Díaz	506	61.9%	70.6%	−7.5%	64.1%
Griffin Jax	550	63.6%	70.3%	−5.8%	65.5%
Steven Okert	537	69.1%	75.8%	−5.8%	70.2%
Bottom 5					
Jose Quintana	937	83.9%	78.7%	4.7%	83.3%
Chris Paddack	1313	81.8%	76.5%	5.0%	81.5%
Tomoyuki Sugano	1205	83.2%	77.1%	5.6%	82.7%
Trevor Williams	675	82.2%	75.6%	5.8%	81.6%
Janson Junk	849	83.4%	76.8%	5.9%	82.8%

Finally, we can use these predictions to evaluate pitcher performance as a metric of their ability to over/underperform their expected contact rates. To do this, I calculated each pitcher's observed contact rate (a), expected contact rate (b), and residual contact rate (c), performed regression to the mean on (b) and (c), and summed them together to find their true predicted contact rate (d). From this, I could create a leaderboard for which pitchers have the most or least skill for throwing pitches that are predicted to be whiffs (n > 500 to qualify). At the top of this leaderboard is Edwin Uceta, with a residual contact rate of **-8.8%** and at the bottom is Janson Junk, with a residual contact rate of **+5.9%**. Edwin Uceta generates exceptional whiffs due to a combination of mechanical deception and pitch quality. His unusually low arm slot and nearly identical release points for fastballs and changeups make it difficult for hitters to pick up the pitch type. His elite fastball, including a two-seam sinker that mimics a rising four-seam, and a changeup with sharp downward movement, create very effective velocity and movement separation. His addition of a cutter adds lateral variation to his mix, further keeping hitters off balance and contributing to consistently high



swings-and-misses. As seen to the left, his fastball and changeup get very similar horizontal movement, but the vertical break is incredibly similar yet offset, meaning that when paired together, these pitches might look similar out of his hand but break differently. Similarly, his cutter breaks just to the right of his fastball, while maintaining a pretty similar lift. His pitches complement each other incredibly well, helping to explain why his residual is so low.



On the other hand, Janson Junk struggles with generating whiffs because his fastball and off-speed pitches lack elite velocity and movement. His “pitching to contact” approach allows for overall effectiveness, yet it produces a higher observed contact rate than expected, resulting in a positive residual. In line with this, he also happened to set the MLB record for walk-rate, with a mere 2.9% walk rate this season. He “pounds the zone”, and trusts his command to induce weak contact. Because he rarely pitches outside of the zone, batters are better able to make contact because there is much less variation in where Junk’s pitches end up. While he posted a respectable 4.17 ERA this season, this is likely an unsustainable style of play.

Conclusion:

By adding more features, we were better able to predict contact probability for pitched balls by considering pitch location, movement, speed, context (ball-strike count), extension, and pitch type. A model which could consider more detailed stats such as when the break occurs, or the sequencing prior to the pitch could better predict contact likelihood. Pitchers such as Uceta who overperform their predicted contact rate likely do so as a result of their sequencing or mechanics, thus being able to deceive batters. On the other hand, some pitchers such as Junk might prefer to stick to the strike zone, resulting in high expected contact rates and even higher actual contact rates. Without much variation, batters can pick up on pitcher tendencies, resulting in high residuals.