OF Catch Probabilities

https://github.com/KyleBushman/SMT_Challenge2023.git

2023-08-31

Abstract

We have created a statistical model to assign catch probabilities to balls hit into the outfield in baseball. The goal is to obtain accurate catch probabilities that can be used to identify outfielders that outperform their peers. A similar method is already being used, but our method allows for other aspects of playing the outfield to be considered.

Why?

When a ball is played on the infield, a clear majority of those plays result in at least one out. If a ball is misplayed, the damage done is often minimal, usually a runner on first. The outfield is a different story. While it is true that most balls played in the outfield result in an out, the costs of making a mistake are much higher. Many times a run will be given up, but an extra base will almost surely be taken at the minimum. The added risks make having an elite outfield valuable. As players attempt to lower their ground ball rates and home run rates go up, the ball is in the air a lot.

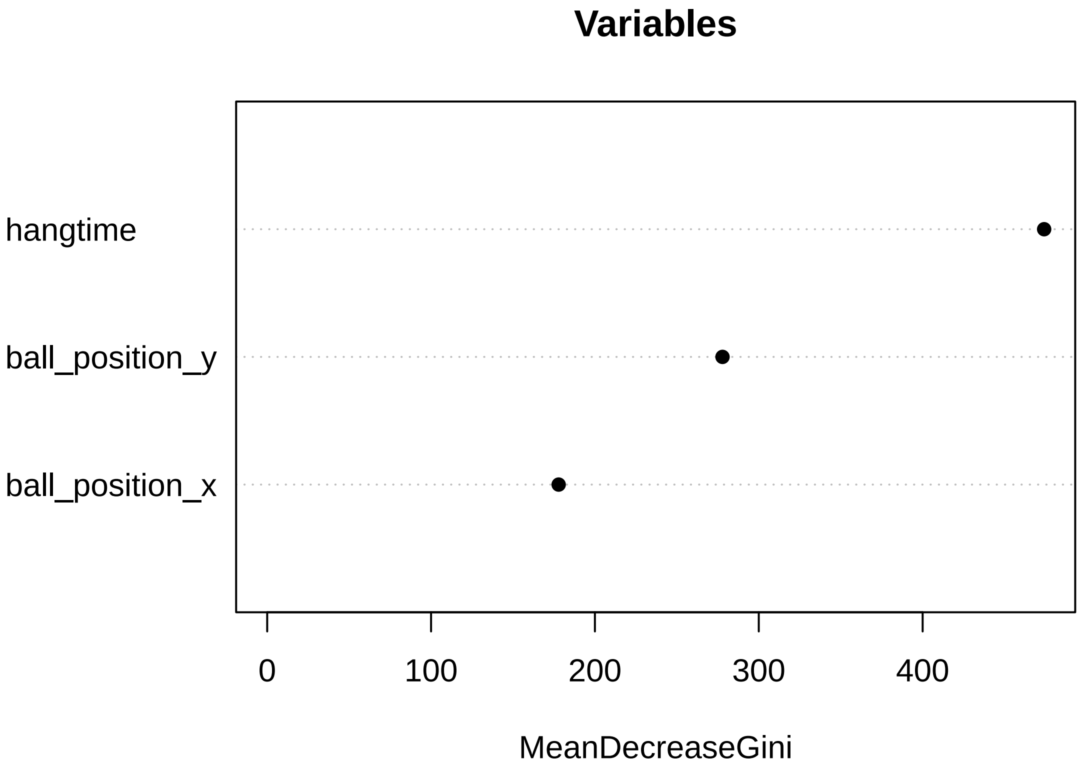
To build the outfield squad that they want, a team needs to have some way to identify who the best outfielders are and how much better they are than other outfielders. One metric used to make this determination is Outs Above Average (OAA). A player who makes catches with a low catch probability will have a high OAA and a player who misses balls with a high catch probability will have a low OAA. For a player's OAA to be calculated, one must first compute a catch probability for the balls played.

Our goal is to find catch probabilities for balls hit in the air to the outfield. Statcast uses distance,the straight line path from the fielders position to the point where the ball is caught or lands, and what they call "opportunity time." Opportunity time is the time between the release of the pitch and the catch or bounce of the ball. We want to take a slightly different approach. We want to use hang time instead of opportunity time. That is, we want to use the time from the ball hitting the bat to the play. Instead of distance to travel, we use the x-y coordinates of the catch or ball bounce.

How?

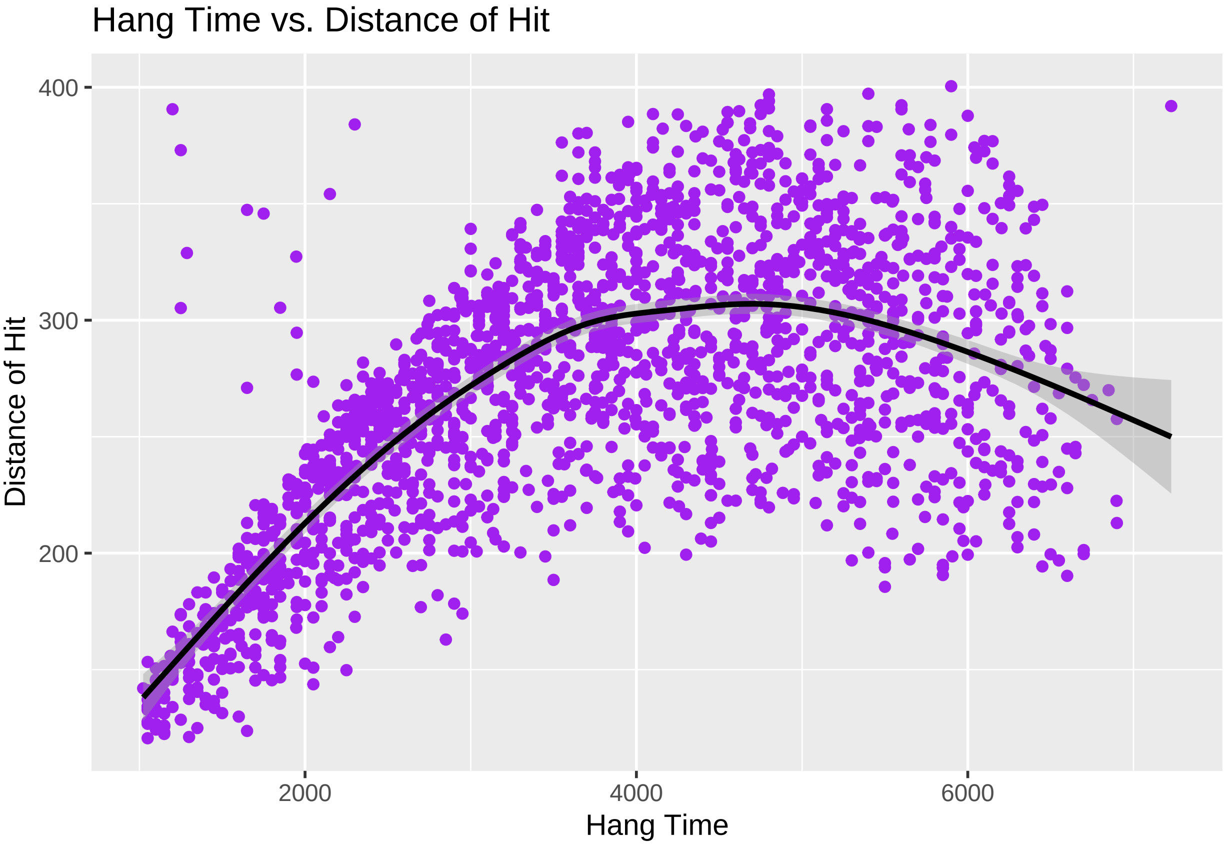
We took three variables: hang time, x-coordinate and y-coordinate. The x-axis is oriented as a straight line from first to third, with zero in the middle. The y-axis is oriented as a straight line from home to second, with zero at home plate. We built a random forest model using these three variables to find catch probabilities.

Below is a plot which tells us which of these variables is the most important.



Clearly, hang time is the most important variable, followed by the y-coordinate then finally the x-coordinate. Next we want to know how each variable relates to the catch probabilities and to each other. The relationships in these plots are not easy to identify from only the data points, so we have added a smoothed line to help visualize the pattern shown by the plots.

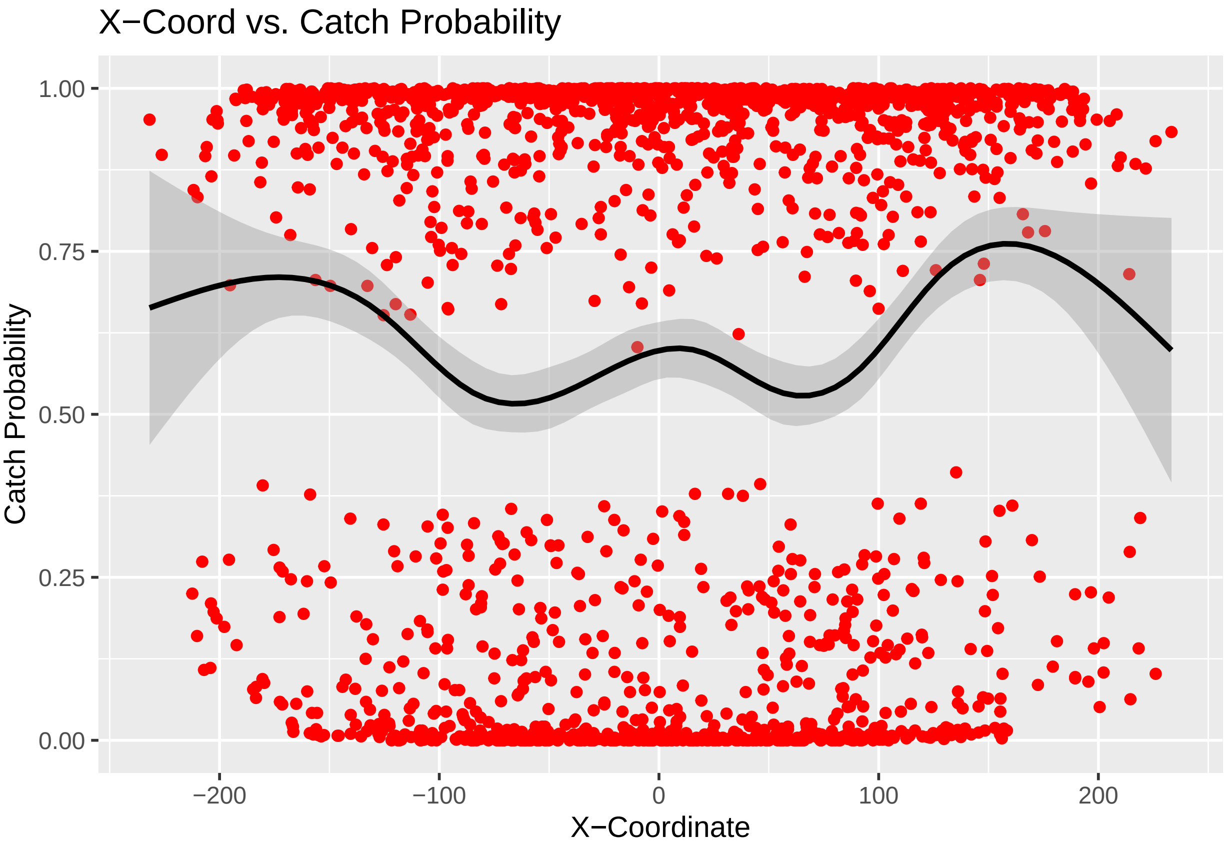
Since the relationship between the x and y coordinates is not interesting, we will instead look at the relationship between the distance the ball travels in the air and hang time.



We can see that hang time and distance generally increase together, until a point where the relationship begins to become less predictable.

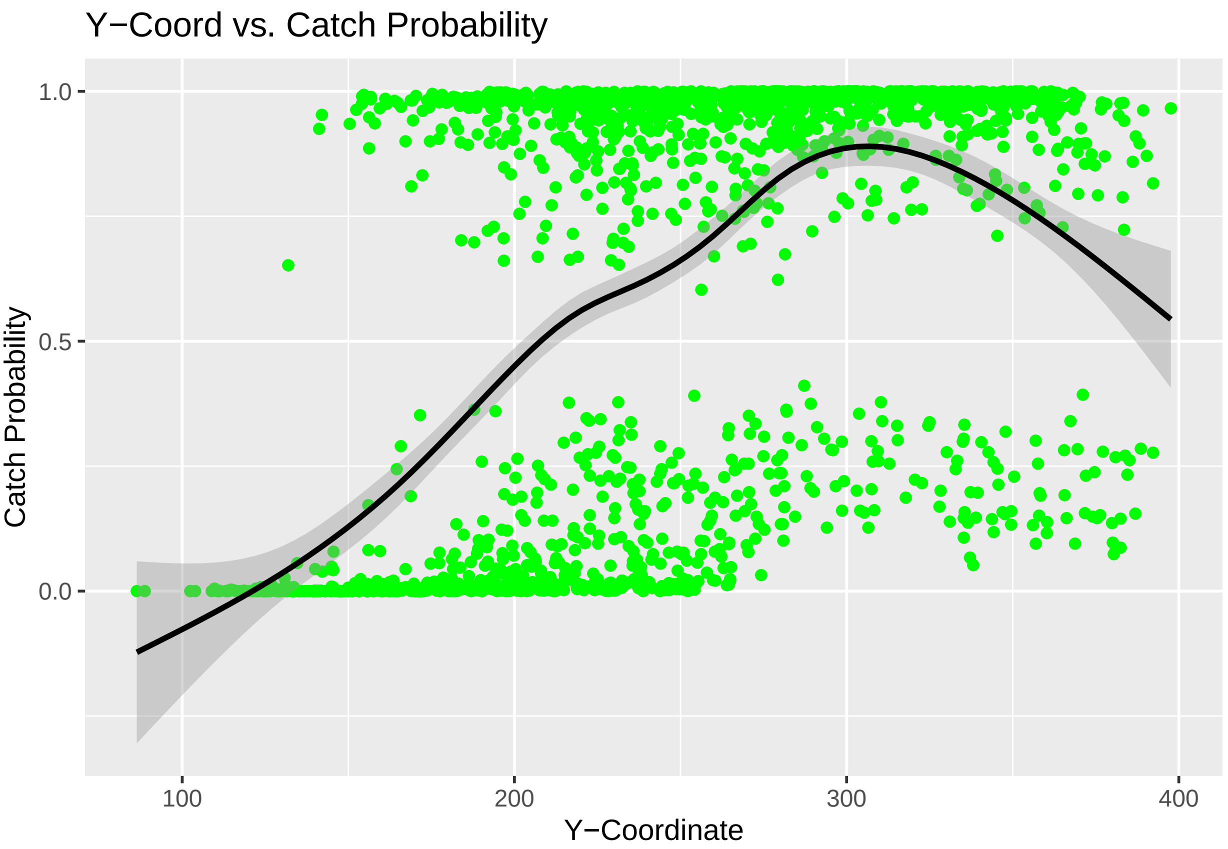
Now that we know how the predictive variables relate to each other, we want to know how they affect catch probability individually. The following plots show these relationships.

First, let's look at the x-coordinate vs. out probability plot.



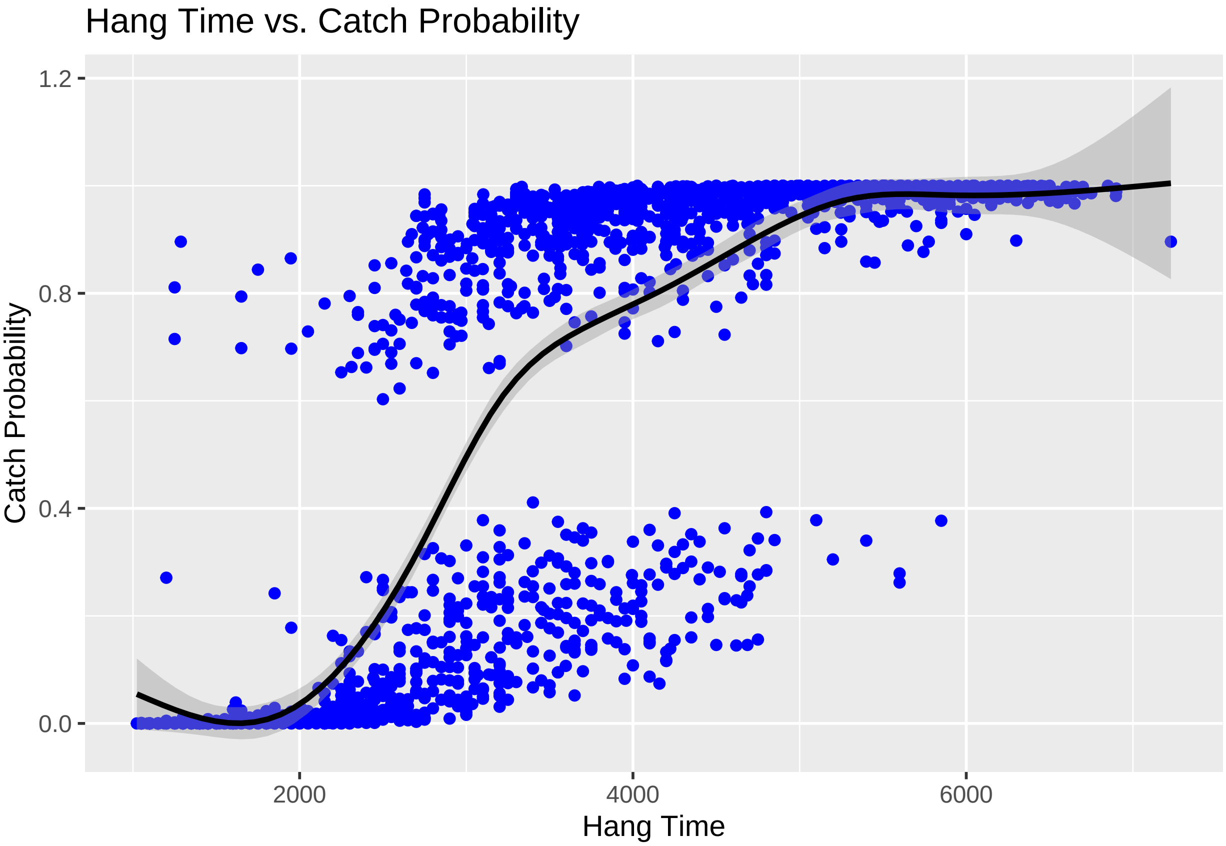
Given that x is zero in the middle of the field, this plot is telling us that catch probability is high to straightaway center, lowers on each side and then increases again toward right and left field, before decreasing again in the corners. This makes intuitive sense, because if the ball is hit directly at the fielders, they will probably catch it.

Next is y-coordinate vs. out probability.



This plot clearly shows an increase in catch probability as the y-coordinate grows, until about 300 feet, then a decreasing effect takes over.

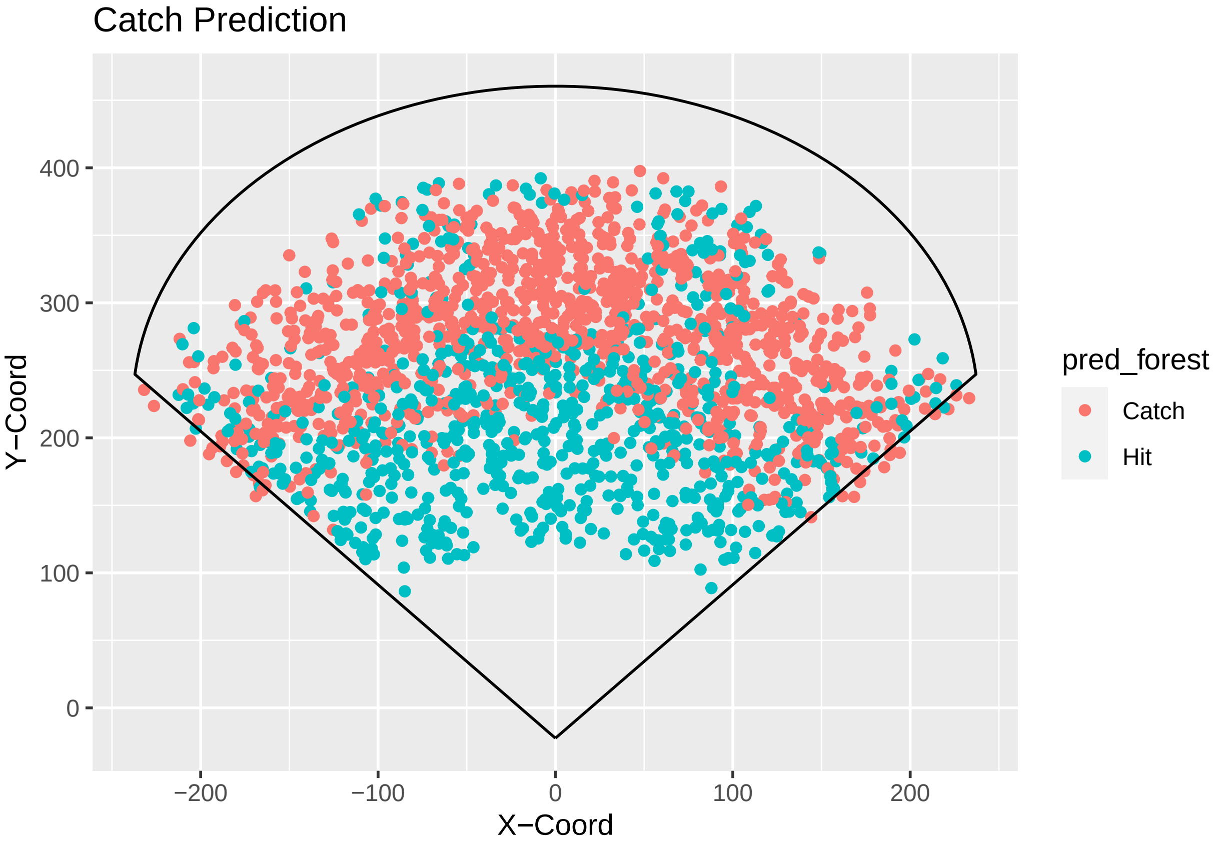
Finally, the most important relationship, hang time vs. catch probability.



In line with intuition, the plot shows that increased hang time always increases catch probability, first very quickly, then at a decreasing rate.

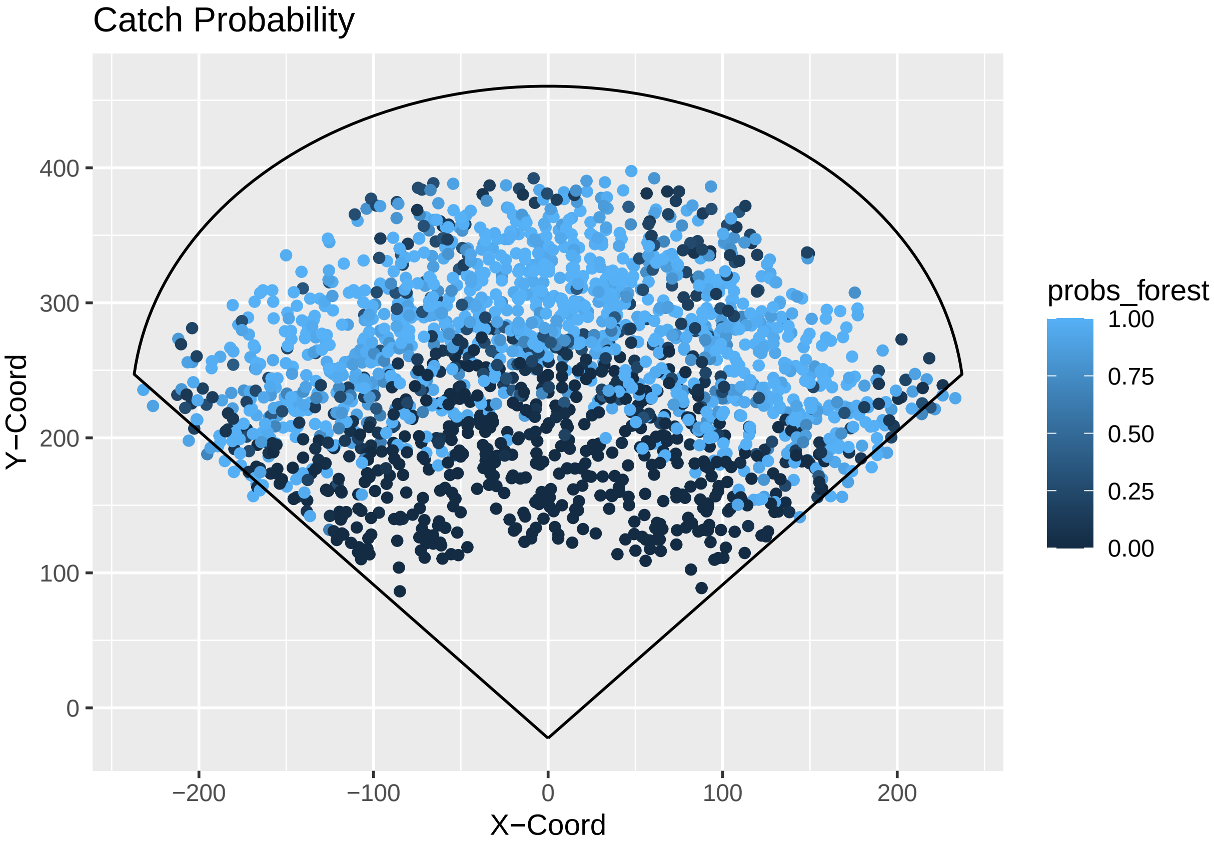
While the relationship between catch probability and each variable individually is important to investigate, reality always includes all three of these variables and all three of the variables affect each other, as shown by the distance vs. hang time plot. Thus, it is vital to see how all of the variables relate to the outcome together.

First we will look at a simple spray chart with a prediction about whether or not the ball is caught, no catch probability or hang time included.



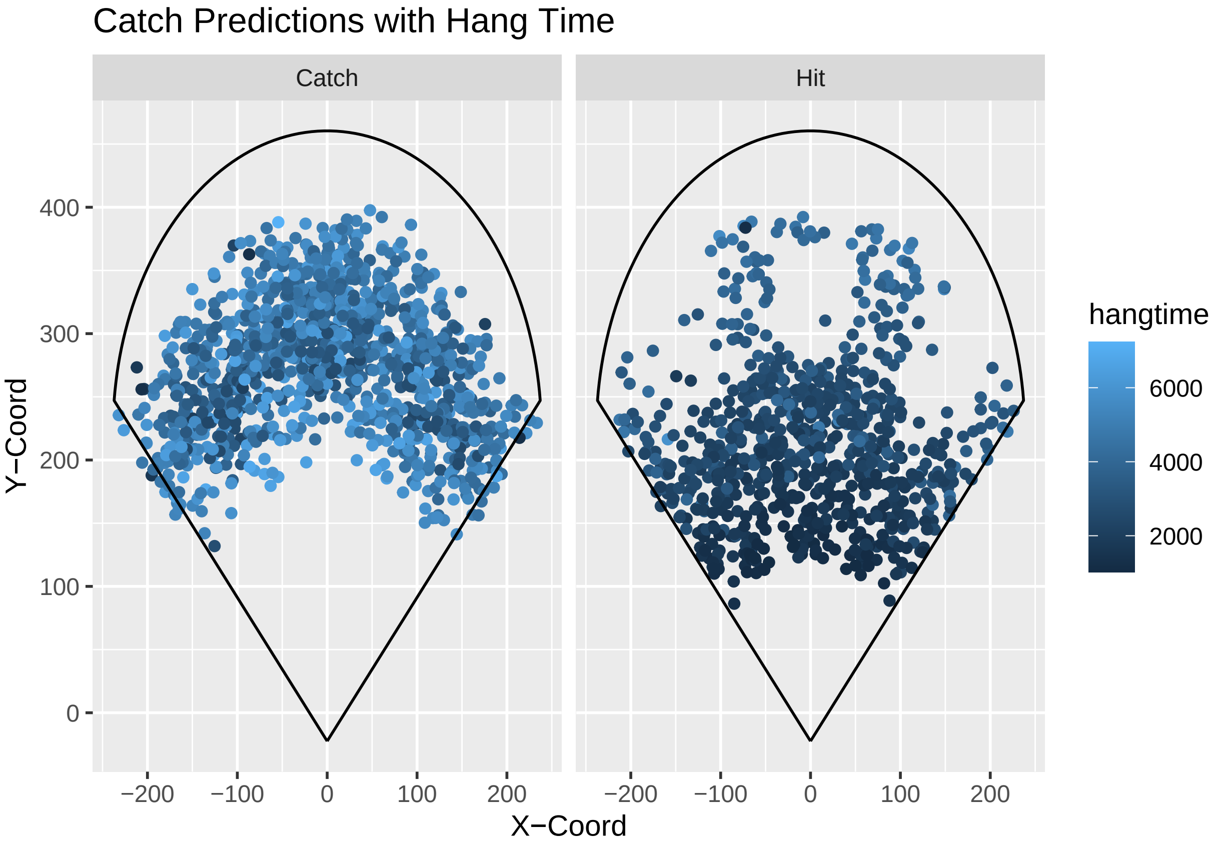
There is a distinct line where predicted hits turn into predicted catches. You can also see a hint of predicted hits in the gaps and corners. Just as the previous plot, this makes sense. Hits generally occur short, in the gaps or in the corners.

Now let's look at a probability plot. This plot still doesn't include hang time, but this time we are looking at catch probabilities instead of just the predictions.



Darker points are less likely to be caught, while lighter points are more likely to be caught. The conclusion is the same as before, balls in the gaps and corners are not caught, as well as short hits.

Lastly, we will look at a plot including hang time. To make it easier to see, I split the plot into predicted hits and predicted catches. Catches on the left and hits on the right. This time, darker points have less hang time than lighter points.



This model is useful for finding the probability of a catch being made given the location and hang time of the ball. One vital way that our model is distinct from the statcast model, is that our model may account for the direction that a player must run to make a catch. If you use only the straight line distance from the fielder's starting position to the point of the catch, you will be ignoring which way they have to run to make that catch. Because it much easier to run forward to make a play than it is to turn around and run toward the wall while looking over your shoulder, the direction ran becomes much more important than it seems.

Next?

While our model can somewhat account for direction ran, it does have its shortcomings. We are essentially neglecting the starting position of the fielder. We assumed that the starting position would be relatively uniform across teams and plays. This is not necessarily the case and the catch probabilities may not be reflective of the actual difficulty of the play, if the player is positioned away from any usual spot.

This model could be improved by including the initial position of the fielder, in x-y-coordinate form. This would help account for that issue as well as improve the ability of the model to determine the direction ran by the fielder. This would be a good addition for the future.