Responses to the Mariners 2025 Analytics Intern Problem Set

Austin Craig
craiga1@uw.edu
University of Washington

Problem 1: Predicting Air Out Probability of Batted Balls

Data

It is important that we carefully consider the structure of the provided data before making predictions of the air out probability. There are a couple of key non-linearities which I will discuss before diving into my modelling choices.

First, consider the horizontal exit angle. As shown in Figure 1, we have many observations for balls which are hit straight into foul territory. Naturally, all observed air outs fall within approximately 45 degrees of dead center, give or take a few degrees for balls caught in foul territory or balls which slice fair after leaving the bat. I account for this by including *horz_exit_angle* as a 2nd-degree polynomial in all models.

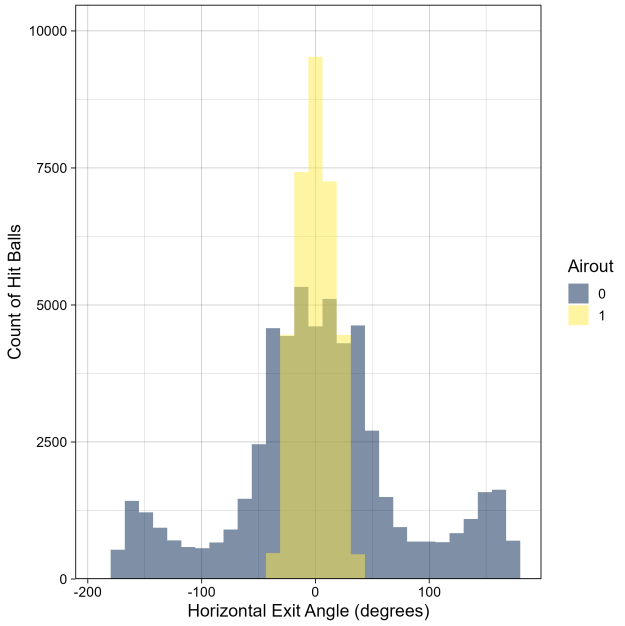

Figure 1: Count of Hit Balls by Horizontal Exit Angle

The key variables used to determine whether a fair ball is catchable by an outfielder are vertical exit angle and exit speed. Importantly, these two variables are more informative when combined. For instance, a ball hit at a relatively high exit angle but low exit speed is much more likely to be a fly out than a ball hit at the same exit angle but with a very high exit speed (which may end up in the bleachers instead).

I adapt the Statcast definition of a **barrel** to account for different types of hits. Specifically, I categorize hit balls into one of six categories based on a combination of exit angle and exit speed (Metzelaar 2020):

1. Poorly hit (weak)
2. Poorly hit (topped)
3. Poorly hit (under)
4. Flare or burner
5. Solid contact
6. Barreled

The distribution of these hit types in the training data is shown in Figure 2. Since the data has been pre-filtered to non-infield plays, we have no observations of poorly hit (weak) balls and only one observation of a poorly hit (topped) ball.

I apply a combination of theory and experimentation (via a 10-fold cross validation procedure) for feature selection. I exclude variables which:

1. Have little theoretical justification for being included in the model, and
2. Add nothing to the predictive power of the model when they are included.

For example, the variable indicating if the hit occurred in the top or bottom of the inning is excluded from most models. There is no clear reason why this variable should affect air out probability, and including does not change the model's performance as measured by accuracy and log-loss score.

Unless mentioned otherwise, the variables included are: temperature, exit speed (and its square), hit spin rate, vertical exit angle, horizontal exit angle (and its square), the interaction between vertical exit angle and hit speed, indicators for if the batter and pitcher are right-handed, a set of indicator variables corresponding to unique venues, and a set of indicator variables corresponding to the hit type variable (described above).

Note that hit spin rate has 1297 missing observations (out of 91,553). In the interest of maximizing the amount of training data available, I impute the missing values at the mean value of 2902.84.

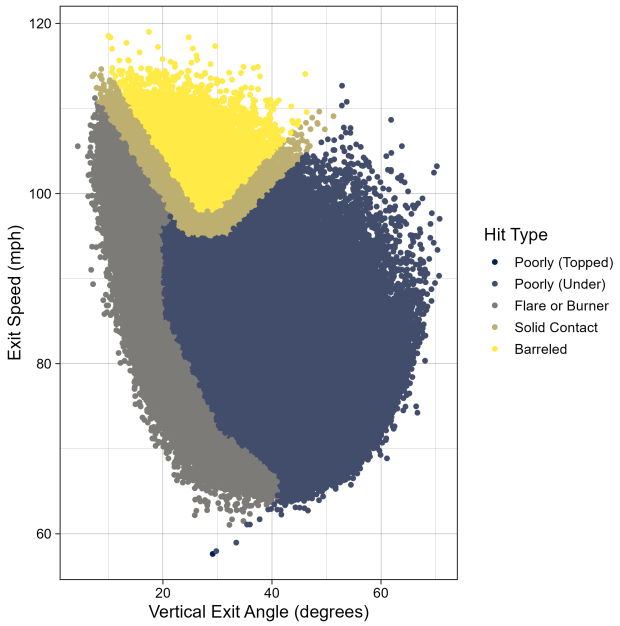

Figure 2: Hit Type by Exit Speed and Vertical Exit Angle

Modeling and Results

I apply three different machine learning procedures to predict the probability of an air out in the test data: logistic regression, logistic regression with principal component analysis (PCA), and a random forest model.

As a baseline, I train a "naive" logistic regression model on a simple set of regressors which does not include squared terms, venue indicators, or hit type indicators. This model performs poorly with a log-loss of 0.616.

The logistic regression model using the full set of predictors performs significantly better, with an accuracy score of 0.868 and a log-loss of 0.308. Using principal component analysis before training the model does not improve predictive power.

I trained the random forest model on 128 trees. This model performs best, with an accuracy score of 0.892 and a log-loss of 0.248. This is the model I use to make predictions on the test data and to generate a fielding score for use in question 2.

Extensions

There are several ways I would expand upon the analysis presented here given more time and resources. First, I would be interested in incorporating more detailed location and weather data. Temperature is a good start, but I would like to tie each venue_id and game date to local conditions including humidity, wind direction, wind speed, and an indicator for if the venue is indoors.

Second, I would like to investigate hit spin rate in more detail. It is difficult to know how hit spin rate will affect the distance a ball travels, and thus the likelihood of an air out, without also knowing if it is backspin, topspin, or side spin.

Finally, incorporating data on player positioning could produce interesting results. This would likely also interact with the handedness of the batter and pitcher, increasing the amount of useful variation I could incorporate into the model.

Problem 2: Report on Player 15411's Outfield Defense

The following evaluation is based on 375 hits for which player 15411 (henceforth, "the player") was identified as the primary fielder. The player's only recorded position for these 375 hits is center field.

The player began the 2023 season at level B and was transferred to level A about a month later (on May 5, 2023). The player was responsible for 237 air outs over the course of the season.

Bottom Line Up Front

The player is a strong outfield defender capable of taking away big hits from the opposition. The player made the adjustment to level A smoothly and maintained a high level of play throughout the season.

Overall Performance

The player's outfield performance during the 2023 season was in the 83rd percentile among qualified outfielders (at least 10 air outs during the season) in terms of mean air outs above expected (mAOAE). This statistic represents how much more likely a player is to catch a given ball than the average player. The player had a mAOAE of 0.026, so they were about 2.6 percentage points more likely to catch a given ball than the average player.

That said, this number conceals significant variation in the player's performance based on the type of ball which was hit into center field. As shown in Figure 3, the player was significantly above average at catching hits classified as a barrel or solid contact, and below average at catching hits classified as a flair or burner.

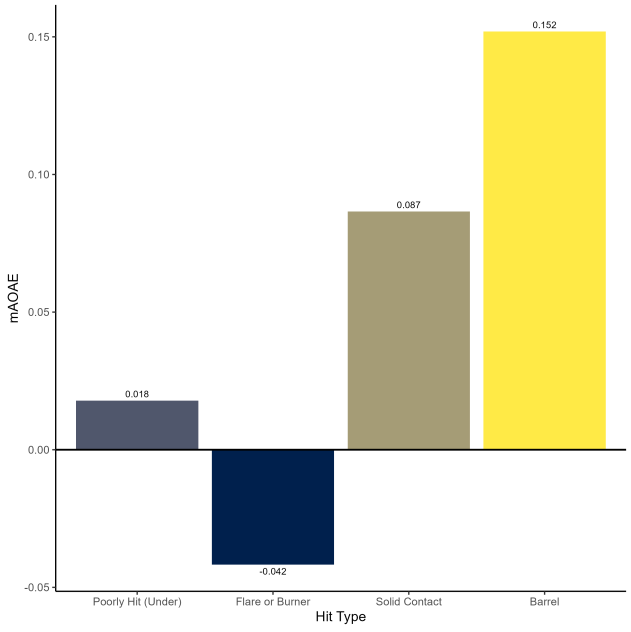

Figure 3: Player 15411's Fielding Performance Against Hit Type

This has significant implications for fielding performance. More than half of barrels result in a home run, and nearly 80 percent result in a base hit (Metzelaar 2020). Solid hits fall just short of being classified as a barrel, but they are excellent at generating extra-base hits. In short, this player is capable of taking away the most valuable hits at a high rate.

Adjustment to Level A

The player's best defensive month of the season was the first month, which the player spent most of at level B. However, a battery of statistical tests do not show any evidence that the player's performance significantly declined after being transferred to level A. This can be seen in Figure 4, where the vertical line represents the day on which the player was transferred to level A.

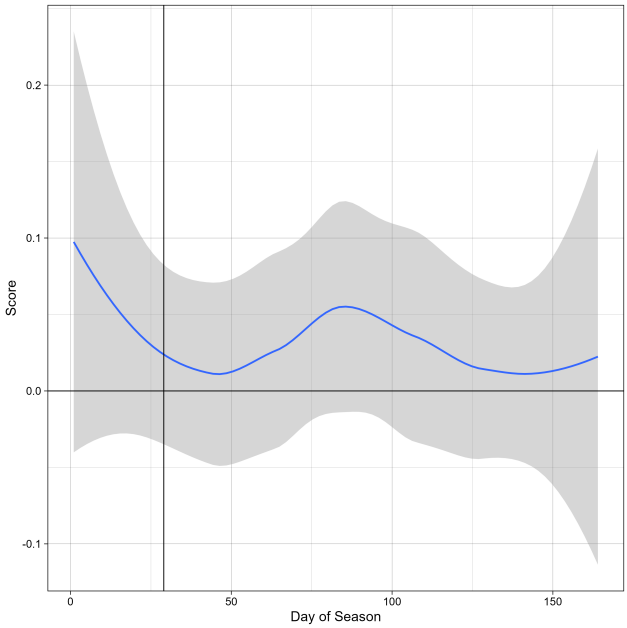

Figure 4: Player 15411's Fielding Score over Time

Problem 3: A Recent Mistake

Some text here

References

Metzelaar, J. 2020. Beyond the barrel: An introduction to ideal contact rate. https://pitcherlist.com/beyond-the-barrel-an-introduction-to-ideal-contact-rate/. Retrieved on Oct 1, 2024.