The Art of the Swing and Miss: Investigating the Components of Shane McClanahan’s Pitches

Jacob Book

2023-05-04

Abstract

In this project, I aim to build a machine learning model that accurately predicts whether a pitch thrown by Shane McClanahan will result in a swinging strike. I utilize logistic regression, decision tree, and random forest models to classify pitches as either swinging strikes or non-strikes. My dataset consists of data from McClanahan’s pitches over his career. To preprocess my data, I use techniques such as feature scaling and feature engineering. After training and evaluating my models, I aim to interpret their results to identify which of McClanahan’s pitches are most likely to result in a swinging strike. This information can be used by coaches and players to improve their game strategy and decision making. Overall, my project demonstrates the potential of machine learning models to provide insights into the performance of individual players and improve overall team performance.

Introduction

Shane McClanahan, a left-handed pitcher for the Tampa Bay Rays, has been a standout performer in Major League Baseball (MLB) in terms of his ability to generate swing-and-misses. Since the beginning of the 2022 season, McClanahan has maintained a swinging-strike rate that is over 1% better than the next best pitcher in the league (minimum 2500 pitches over that span). The skill of generating swing-and-misses is overwhelmingly considered to be one of the most valuable traits for pitchers. In this project, we aim to build a predictive model using logistic regression, decision trees, random forest algorithms, and boosted trees to accurately predict the likelihood of a swinging strike for any given pitch thrown by Shane McClanahan. We will use a dataset of his pitches over the course of his career to train and validate these models. The data was obtained via the MLB Savant Statcast database – containing over 90 variables that together described every element of the at bat in which the pitch took place. I was able to narrow the dataset to a select few variables that I thought to be most pertinent to the outcome of a given pitch. The ultimate goal is to identify which of his pitches is most effective at generating swing-and-misses and to potentially provide insights that could help him further refine his pitching strategy.

Exploratory Data Analysis

The comprehensive dataset of pitches was trimmed to only include observations in which the batter swung. This dataset contained 2344 swings – 1192 of which were classified as a “swinging strike”. Below can be found several plots containing additional information on Shane McClanahan’s pitches:

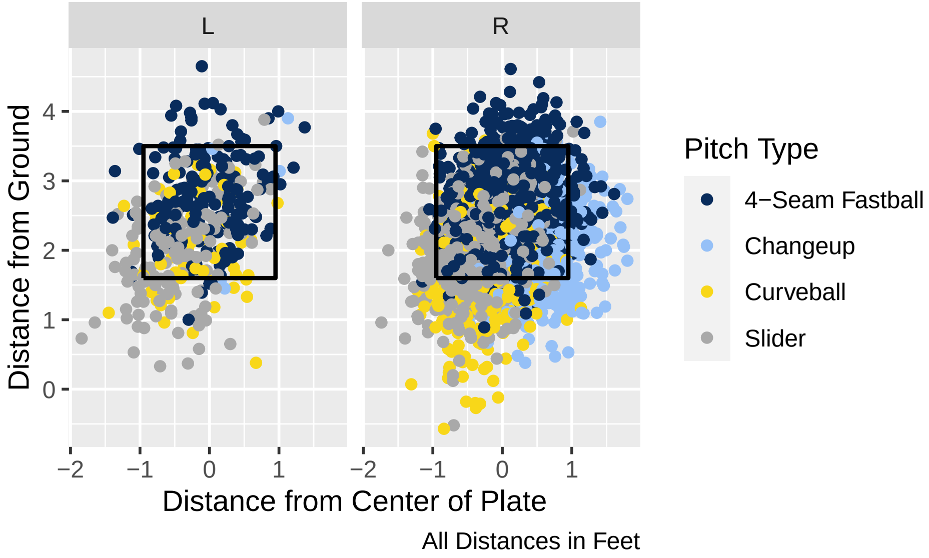


Figure 1: All pitches that generated a ‘swing and miss’ – separated by batter handedness

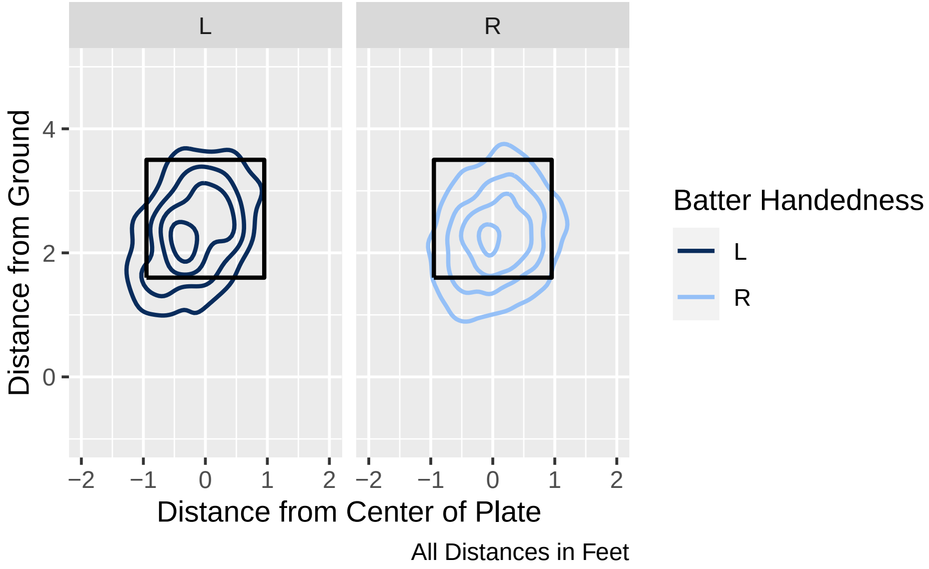


Figure 2: Pitch location tendency – stratified by batter handedness

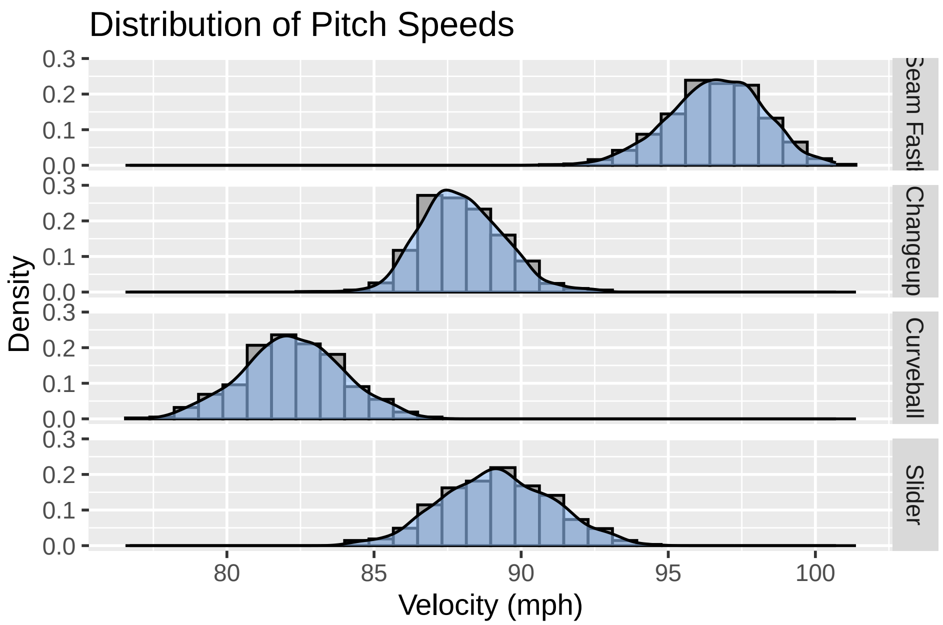


Figure 3: Pitch velocity distribution curves for the fastball, curveball, slider, and changeup

Methods

The data used in this study was collected from Major League Baseball's (MLB) Statcast system. We specifically focused on the pitching performance of Shane McClanahan.The variables used to train the model included horizontal and vertical pitch movement, lateral and vertical position, pitch velocity, spin rate, spin axis, and the zone of the pitch as it crossed the plate.

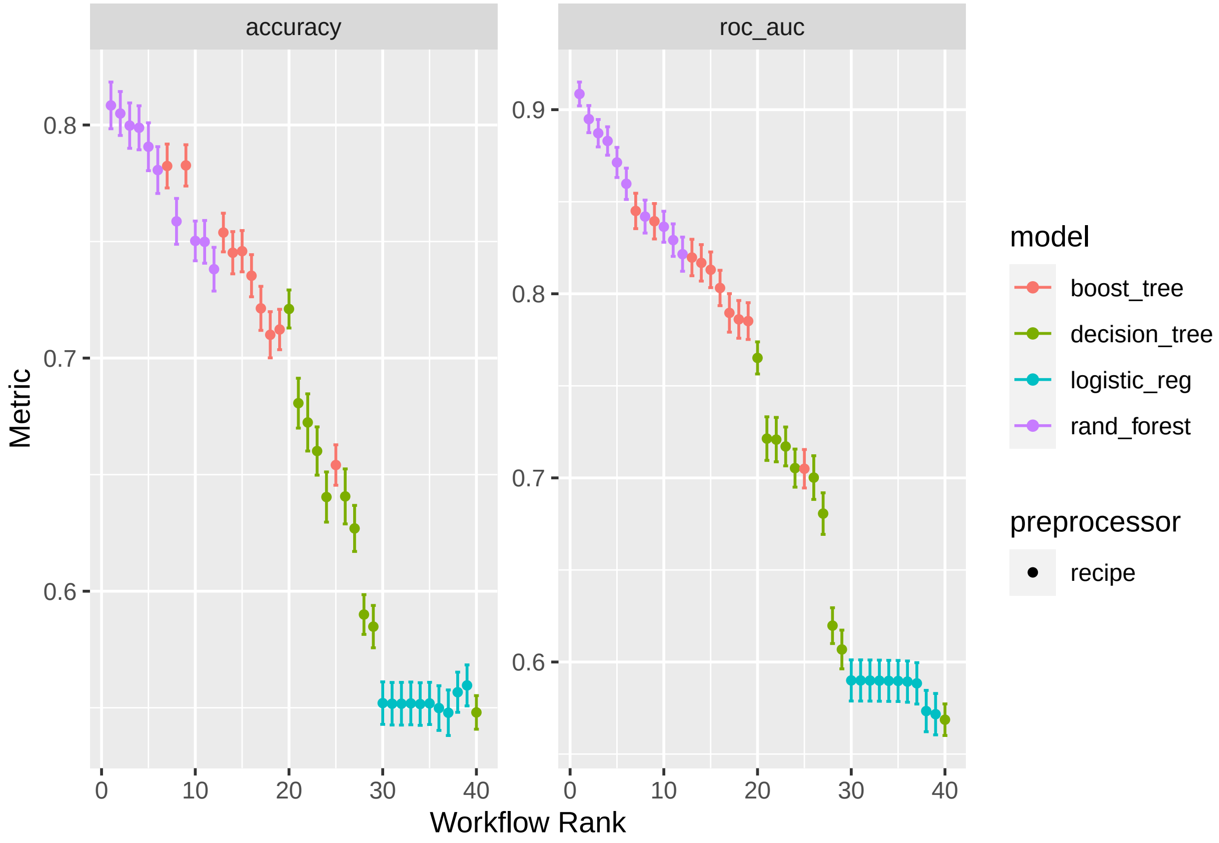
Prior to modeling, the dataset was preprocessed to ensure that it was suitable for analysis. This included removing any missing values, transforming categorical variables into dummy variables, scaling continuous variables to have a mean of 0 and standard deviation of 1, and decorrelating variables. While I preprocessed the data via standard upsampling in order to achieve class balance between "swing-and-miss" and "no-swing-and-miss" outcome variable. However, what produced better results was sampling with replacement based on the class of the outcome variable and the handedness of the batter. That is, I acheived perfect class balance for the swing-and-miss variable for both left handed and right handed batters. This allowed the model to be able to account for left handed batters that rarely face Shane McClanahan.

I developed four different models to predict whether a pitch thrown by Shane McClanahan would result in a swinging strike or not. These models included a logistic regression, a decision tree, a random forest, and a boosted tree model. I tuned penalty and measure for logistic regression, tree depth, cost complexity, and minimum n for the decision tree model, mtry and minimum n for random forest, and for the boosted tree, I tuned mtry, minimum n, and loss reduction. Each model was trained on a random 75% subset of the data and tested on the remaining 25% subset using 10-fold cross-validation, repeated 5 times.

To evaluate the performance of each model, we used several metrics, including accuracy, F-measure, and receiver operating characteristic (ROC) curve analysis. We also examined the feature importance rankings produced by each model to gain insights into which variables were most important in predicting a swinging strike.

Results

After thorough tuning and analysis, the random forest was deemed the best model based on its accuracy and ROC AUC. The best random forest model produced an accuracy of 0.812 and an ROC AUC of 0.909.



Conclusion

Based on the results of the models – the random forest model in particular, I was able to extract variable importance from the model algorithm. While at first, it appears that vertical position as the ball crossed the plate is overwhelmingly the most important factor in whether a swing will result in a strike or not, I realized the model is being swayed by pitches out of the zone. The curveball – notable for its extreme downward movement, often taking the pitch out of the strike zone altogether, was likely convincing the model that a pitch below the zone was a good pitch. While this can certainly be the case, it is intuitively not the most important element. Thus, in an effort to remove pitch location bias, I trained similar random forest model that was trained only on pitches that finished in the strike zone. While this limited model clearly cannot be used to predict pitch results due to a chunk of missing information, it does lend helpful insight into the things that truly influence the result of a pitch.

After analyzing the models as well as the complementary plots, it was found that pitch speed and velocity were very important features. Additionally, it was discovered that vertical position was much more important than previously realized, while lateral and vertical movement may not be as important as originally thought.

The variable importance plots were found to be helpful in understanding the relative importance of the different features in the models, however perhaps they may be even more useful for analyzing individual pitches, as the importance of each feature can vary depending on the pitch type.

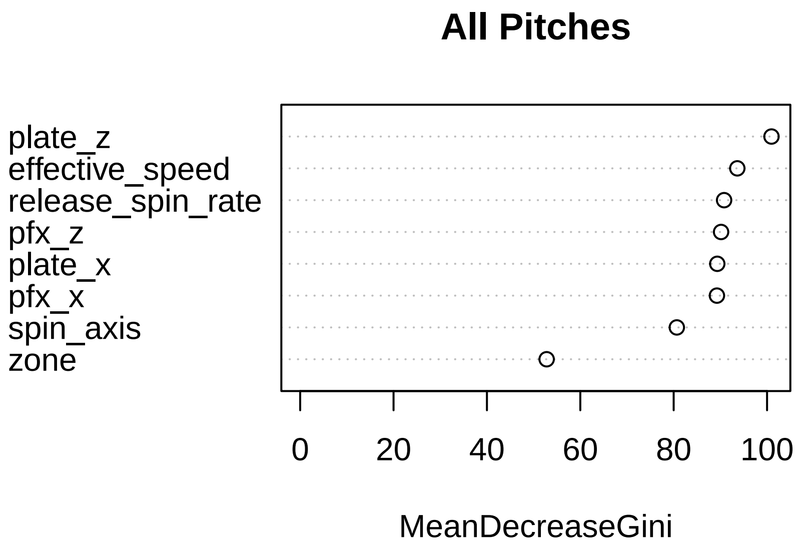


Figure 4: Variable importance for pitch model that contains all swings

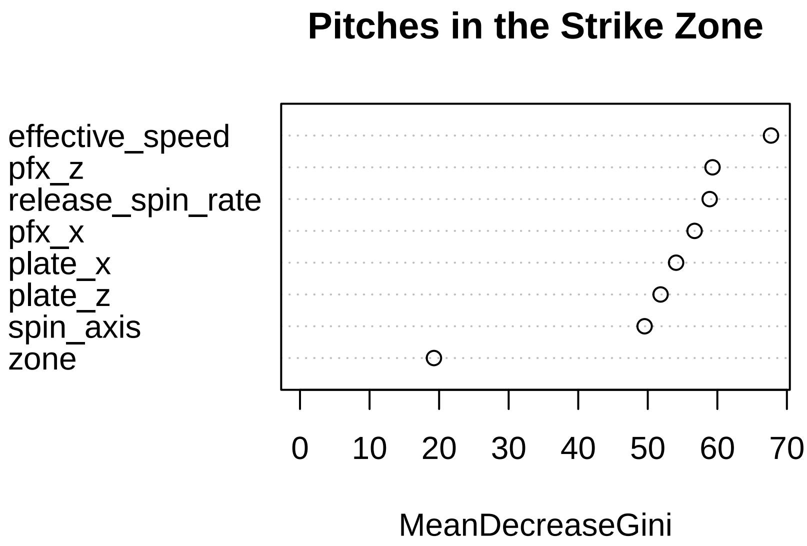


Figure 5: Variable importance for pitch model that contains only swings on pitches in the strike zone

In conclusion, these findings suggest that careful attention to the specific features included in the models can greatly impact their predictive accuracy. The variable importance plots provide a useful tool for gaining insights into the relative importance of different features in the models. Pitch speed and velocity, as well as spin rate, and the resulting movement of the pitch, should be given particular consideration when predicting swinging strikes as well for use in game.

Future Considerations

While the current analysis focused on predicting swinging strikes for Shane McClanahan, there are several areas for further research that could expand upon these findings.

First, it would be interesting to explore how the developed model could be used in real-time analysis during games. This could involve integrating the model with live data streams to provide instant feedback on the likelihood of a swinging strike based on the current pitch and batter information. Additionally, investigating batter data could provide additional insights into swinging strike patterns. By analyzing the tendencies of individual batters, it may be possible to identify factors that increase or decrease the likelihood of a swinging strike. Finally, expanding the analysis to include other pitchers and pitch types would provide a more comprehensive understanding of the factors that contribute to swinging strikes. By examining a wider range of pitchers and pitches, it may be possible to identify common patterns or features that are predictive of swinging strikes across different contexts.