**Technical Report: Batter Pitch Mix**

*Written By Alex Kelleher*

**Problem Analysis**

The attached model is designed to predict the pitch mix a batter will face in the upcoming year by analyzing the previous years’ data. The pitch mix is broken into three categories: **offspeed**, **breaking**, and **fastball** pitches. The goal of the output is to give insight into the pitches a batter is likely to receive.

**Domain Knowledge**

The data set provided consists of 1,286,181 pitches that took place across the 2021-2023 seasons. Each pitch is described by 56 variables. Before beginning the analytical feature selection, certain variables were eliminated because their value does not impact model accuracy (date, time of day, top/bottom inning etc.) Based on domain knowledge, the below assumptions were made to narrow down the field of descriptors.

**Assumptions:**

* At bats will be normally distributed over the course of the game and season.

For example: a batter will not have one at bat in the first inning and then have the following four occur in the ninth inning.

* We do not know what pitchers the team will face next year.

A pitcher’s arsenal greatly decides what a batter will face in a given game; however, the pitchers a batter will face cannot be predicted prior to the season.

* We do not know the score or how many men are on base for each at bat in the upcoming season.

Without knowing the score, no variable regarding win expectancy can be used.

* Variables about pitch location cannot predict the future pitch mix.
* The quality of a batter’s balls in play are best summarized by two variables.

ESTIMATED\_WOBA\_USING\_SPEEDANGLE is used to judge the quality of balls in play; the variable DELTA\_RUN\_EXP, accounts for change in game states on each pitch, both in play and not in play.

Two new variables were created prior to beginning the preprocessing steps:

BALLS\_STRIKES: Instead of having two separate variables, balls and strikes, these were combined to better capture the game state. The relevance of strikes in a count is dependent on the number of balls and vice versa.

Class: The classification of the pitch. Either a FASTBALL, BREAKING, or OFFSPEED pitch. **This is what we are trying to predict.** Baseball Savant’s definition of each classification is how the PITCH\_TYPE was translated to Class.

The initial data set included the categorization ‘Other’. All of these pitches were removed from the dataset because they accounted for only 2.53% of the data and at such a low frequency, the value of these pitches in terms of preparation and evaluation is near zero.

**Model selection**

A random forest model was used for three primary reasons:

1. Ability to handle multi-output regression
2. Feature importance scores
3. Flexibility with data types

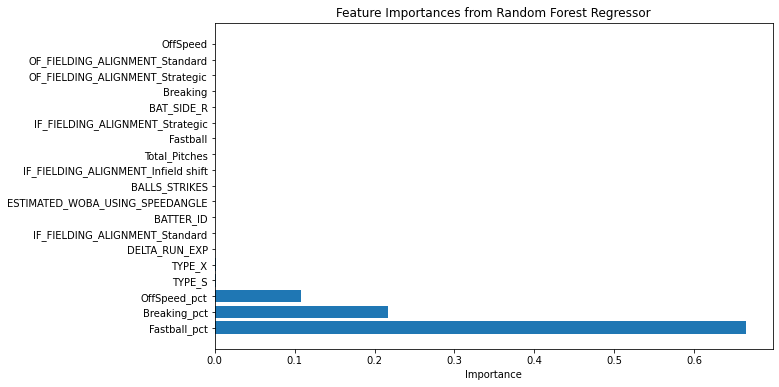
The rate a batter receives Fastballs, Breaking balls, and Offspeed pitches is the desired output. Feature importance scores are an accurate way to evaluate each variable’s impact on the model. The evaluation tools used to assess model quality are MSE, RMSE, *R²*, MAE, and average cross validation score.

The full model was made by breaking down class predictions by count, and making a prediction for each batter’s 12 potential counts faced. After checking for variables that were independent of BATTER\_ID the only variables with a strong correlation to pitch mix were balls and strikes. BALLS\_STRIKES had an *R²* value of 0.6309 and no other variable had an *R²* above 0.400. The final model for the submission file, averaged out these class predictions based on frequency of each count.

Batter performance was expected to be a significant factor in pitch mix received. This was factored into initial Models A and B and was shown to not have a strong impact on pitch mix. Model A and Model B, broke down DELTA\_RUN\_EXP and ESTIMATED\_WOBA\_USING\_SPEEDANGLE by pitch type and based on the results of Model A and B, and the low factor importance of the models’ specific variables these variables were excluded from the final submission.

| Model Evaluation | | | |
| --- | --- | --- | --- |
| Model | **Model A**  DELTA\_RUN\_EXP  (By Pitch Type) | **Model B**  ESTIMATED\_WOBA\_USING\_SPEEDANGLE  (By Pitch Type) | **Model Submitted**  No Performance Pitch Breakdown |
| MSE | 0.0022 | 0.0021 | **0.0002** |
| RMSE | 0.0469 | 0.0457 | **0.015** |
| R² Score | 0.7969 | 0.8047 | **0.9693** |
| MAE | 0.0161 | 0.0157 | **0.0017** |
| Average cross-validation score | 0.8381 | 0.8448 | **0.9704** |

| Feature Importance | | | |
| --- | --- | --- | --- |
| Model | **Model A**  DELTA\_RUN\_EXP  (By Pitch Type) | **Model B**  ESTIMATED\_WOBA\_USING\_SPEEDANGLE  (By Pitch Type) | **Model Submitted**  No Performance Pitch Breakdown |
| Fastball\_pct | 0.5906 | 0.5911 | **0.6649** |
| Breaking\_pct | 0.1654 | 0.1657 | **0.2172** |
| OffSpeed\_pct | 0.1014 | 0.1033 | **0.1084** |
| Highest Model Specific Factor | D\_R\_E\_FB  0.0084 | EST\_WOBA\_U\_S\_FB  0.0074 |  |



**Conclusions**

The feature importance analysis table below indicates that Fastball\_pct of the previous year was the most influential predictor of the next year’s pitch mix. This is in line with the other strong indicators Breaking\_pct and OffSpeed\_pct. The strong *R²* value of 0.9693 indicates 96.93% of the variance in the next year's pitch mix distribution is accounted for in the model.

Surprisingly, models factoring in batter performance on each pitch type with DELTA\_RUN\_EXP and ESTIMATED\_WOBA\_USING\_SPEEDANGLE, performed much worse than models that excluded these variables. It was expected that if a batter improved their performance against OffSpeed pitches, the rate they would receive them would be significantly impacted. The models contradict this assumption.

The previous years’ batters pitch mix and pitch count were by-far the most significant variables within the model. The pitfalls of the model include the length of time it takes to accurately predict a rookie’s’ pitch mix’s. The model was only trained on the previous three seasons; longer league-wide trends could change over a longer period of time. This model should be updated each year to monitor potential trend changes. The more actionable version of the model will not factor in IF or OF alignment because of their low significance.