Predicting Pitch Mixes for Batters Based on Pitch-By-Pitch Data in Prior Seasons

Methods

Dataset Selection & Pre-processing

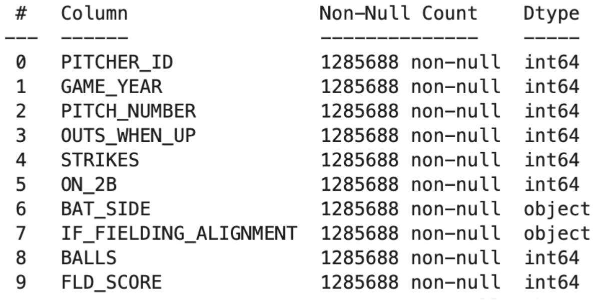
Since, the provided dataset contained numerous variables, it was important to select variables that were correlated with pitch mix. To do this, two methods were used. First, using prior knowledge of baseball statistics, all variables that described the outcome of an at-bat were dropped (ex. HIT_LOCATION, LAUNCH_SPEED, etc.) since they describe events after the pitch was thrown and thus not relevant for predicting the pitch type. Moreover, names of players were also dropped since the BATTER_ID can be used instead to identify the batter. Variables that described batter/pitcher characteristics, game situations (ex. inning, pitch count, pitch number, etc.), and team information were included since these variables all play a part in pitch selection. For example, it is commonly known that a 3-0 count is a fastball count so batters are more likely to see a fastball. Moreover, different teams/pitchers have different strategies that may cause them to throw a certain type of pitch more against a certain batter.

The second method was to select features from performing exploratory data analysis. By creating a correlation matrix, the correlation between each variable and PITCH_TYPE could be visualized, and the variables with the greatest correlation could be selected for modeling. Another approach was to use the sklearn feature selector, which takes advantage of a random forest classifier to determine the most influential features.

In terms of data preprocessing, missing values and mapping the original pitch types into the three categories (see Table 1) had to be handled. Taking a closer look at the dataframe, it was observed that the ON_1B, ON_2B, ON_3B columns had NaN values since at many at-bats, there were no players on the bases. Therefore, NaN values were replaced with 0 and anything else was replaced with 1 for these columns. While some information was lost, specifically the ID of the player on the base, this approach still kept useful information without overwhelming the model with complicated details. Furthermore, columns such as BAT_SIDE, THROW_SIDE, and other binary variables were encoded into 0 and 1. For the infield and outfield alignment variables, since there were a couple thousand missing values, they were imputed by using the most frequent alignment which turned out to be ‘Standard’. This approach was chosen since Standard formation dominated the dataset and the number of missing values were relatively low, so using mode wouldn’t significantly skew the distribution.

Table 1. Raw pitch types in each of the three broader categories

Broader Pitch Categories	Fastballs (FB)	Breaking Balls (BB)	Off-Speed (OS)
Pitch Types in Category	FF, FA, SI, FC, PO	SL, CU, KC, ST, SV, SC	CH, FO, FS, KN, EP



```
 #   Column                 Non-Null Count    Dtype
---  ------                 --------------    -----
 0   PITCHER_ID             1285688 non-null  int64
 1   GAME_YEAR              1285688 non-null  int64
 2   PITCH_NUMBER           1285688 non-null  int64
 3   OUTS_WHEN_UP           1285688 non-null  int64
 4   STRIKES                1285688 non-null  int64
 5   ON_2B                  1285688 non-null  int64
 6   BAT_SIDE               1285688 non-null  object
 7   IF_FIELDING_ALIGNMENT  1285688 non-null  object
 8   BALLS                  1285688 non-null  int64
 9   FLD_SCORE              1285688 non-null  int64
```

Figure 1. Image of Variables Selected for Training after Pre-processing

Modeling

The primary approach for modeling the pitch mix for each player involved building a separate model for each batter using their individual historical data. By tailoring a model to each batter, we can capture the specific patterns and trends in how pitchers have approached them in previous seasons. The trained model was used to simulate over all data, including data that isn't specific to the batter to calculate the pitch mix the batter might face in 2024. The choice to simulate over all data rather than just historical data for each batter was due to three factors.

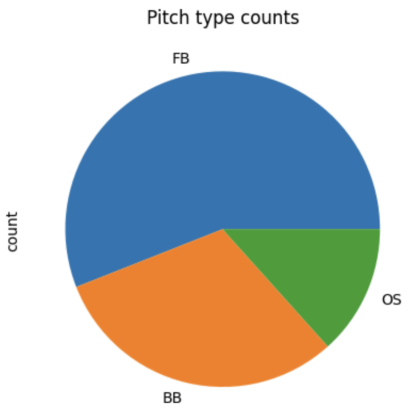
First the variables included in the model weren't team or batter-specific. For example, if BATTER_ID or HOME_TEAM were included, then using all the data wouldn't work since different batters and teams would be included to predict the pitch mix of the batter the model was trained on.

More importantly, since we don't know the game situations, pitchers, etc for the 2024 season, by simulating over all the data, it exposes the model to a variety of new pitchers and situations beyond past experiences, enabling the model to possibly reflect emerging trends in 2024 that weren't seen in previous years.

For experimentation, three models were chosen for testing: Random Forest, Gradient Boosting, and Multinomial Logistic Regression. While the accuracies of the models were about the same, each model has its own strengths and weaknesses. The main benefit for logistic regression is its interpretability. Since the logistic regression offers coefficients that indicate the influence of each predictor on the log-odds of the outcome classes, it is valuable for understanding how different factors affect pitch selection, which can be communicated effectively to stakeholders. However, the logistic regression assumes a linear relationship between variables and the log-odds of the outcome which may incorrectly capture the complexity of pitch selection. Random Forest and Gradient Boosting offer the ability to capture nonlinear complex patterns but are less interpretable and computationally intensive.

Other models that were considered include more complex machine learning techniques such as Neural Networks, and time series analysis; however, all of these models are considered 'black boxes' making it difficult to explain how the input variables affect the output. Moreover, with their high complexity and computational demand, there is a higher risk of overfitting, especially due to the limited dataset provided (a couple thousand data points for each batter over 3 years). Furthermore, many time series models require a consistent time interval and require features other than the provided year (2024) and BATTER_ID to predict pitch mixes for future years.

Discussion & Limitations



Given that the dataset was limited to 3 years between 2021-23, with only 8 months of data per year, it would be hard to find a pattern given 24 data points. It would be difficult to utilize time series analysis since many models such as ARIMA require more seasons/data points. Incorporating more extended historical data could improve the model's ability to capture seasonal variations and long-term trends.

Moreover, when performing EDA, it could be seen that FB data was more than 50% of the entire dataset which resulted in model bias towards the majority class. Although techniques like undersampling or class weighting were considered, further exploration of advanced imbalance handling methods (e.g., SMOTE, ADASYN) could enhance model performance for minority classes. Moreover, accuracy is not always a good metric, especially with imbalanced datasets. While the model may achieve high accuracy by predicting FB most of the time, this overlooks poor performance on minority classes (Breaking Balls (BB) and Off-Speed (OS)). Thus, metrics like precision, recall, F1-score, and ROC-AUC were used to provide a more balanced view of performance.

Furthermore, while feature selection was used to select the most influential features, there may be additional features that influence pitch selection that were not included. For instance, the combination of batter-handedness and pitcher-handedness could be an important interaction, but without explicit interaction terms, it may not be adequately modeled.

Additionally, baseball strategies evolve over time, with teams and pitchers continuously adapting and changing. Without information about which pitchers are still in the league, injuries, changes in technique or team strategy, the model may not account for these variations leading to less accurate predictions for future years.