**Technical Report: Pitch Type Prediction Model**

**Summary**This report presents the development and evaluation of a pitch-type prediction model for the Cincinnati Reds, aimed at enhancing player strategy and decision-making. The model utilizes historical game data to predict the likelihood of a batter encountering various pitch types: Fastball (FB), Breaking Ball (BB), and Off-Speed (OS). By employing a Random Forest algorithm, the model achieves a Mean Squared Error (MSE) of 0.20203 for Fastball predictions and provides actionable insights for the coaching staff.

**Introduction**

**Purpose of the Report**The purpose of this report is to summarize the data used, the modeling approach taken, and the results of the pitch-type prediction model developed for the Cincinnati Reds.

**Context**Understanding pitch types is crucial for player performance, as it influences batting strategies and outcomes. This model aims to provide predictive insights that can aid in tactical planning during games.

**Dataset Description**  
The dataset comprises historical game data from [insert date range], containing [insert number of records] records with various features related to player performance, pitch types, and game situations.

**Features Included**The model uses the following key features:

* **BATTER\_ID**: Unique identifier for each batter.
* **PLAYER\_NAME**: Name of the player.
* **BAT\_SIDE**: The batting side of the player (left/right).
* **THROW\_SIDE**: The throwing side of the pitcher.
* **INNING**: Current inning in the game.
* **OUTS\_WHEN\_UP**: Number of outs when the batter is at the plate.
* **BALLS**: Number of balls in the current count.
* **STRIKES**: Number of strikes in the current count.
* **PLATE\_X**: Horizontal position of the pitch.
* **PLATE\_Z**: Vertical position of the pitch.
* **PITCH\_NUMBER**: Count of pitches in the at-bat.
* **GAME\_YEAR**: Year of the game.

**Preprocessing Steps**

* **Handling Missing Values**:
  + Dropped columns with over 60% missing data.
  + Imputed missing values in critical columns (e.g., PLATE\_X, PLATE\_Z) with their median values.
* **Encoding Categorical Variables**:
  + Categorical features such as BAT\_SIDE and THROW\_SIDE were label encoded to facilitate model training.

**Modeling Approach  
Model Selection**A Random Forest algorithm was selected due to its robustness in handling nonlinear relationships and ability to capture feature interactions. Based on the complexity of the data, this model was deemed appropriate for predicting pitch types.

**Feature Engineering**Several new features were engineered to enhance the model's predictive power:

* **Cumulative Pitch Count**: Tracks the number of pitches faced by each player, contributing to game context.
* **Inning Pressure**: Quantifies the pressure based on the inning and score difference, reflecting the game's stakes.
* **Pitcher-Batter Matchup**: Encodes the matchup between the batter and pitcher, which is critical for understanding pitch selection.

**Training Process**The data was split into training (80%) and testing (20%) sets. The model was trained using the training data, and various metrics were employed to evaluate its performance.

**Model Evaluation  
Metrics Used**The model's performance was evaluated using the following metrics:

* **Mean Squared Error (MSE)**: A measure of the average squared difference between predicted and actual values.
* **R-squared (R²)**: Indicates the proportion of variance in the dependent variable predictable from the independent variables.

**Results**The model achieved the following evaluation results:

* **MSE for Fastball prediction**: 0.20203
* **MSE for Breaking Ball prediction**: 0.17925
* **MSE for Off-Speed prediction**: 0.09318

**Feature Importance**The importance of various features was assessed to identify the most influential factors in predicting pitch types. The following features were found to be the most significant:

1. **PLATE\_X**
2. **PLATE\_Z**
3. **BALLS**
4. **STRIKES**
5. **INNING**

Visualizations of feature importance are included   
1. Feature Importance for Fastball Prediction  
A graph with blue bars

Description automatically generated

2. Feature Importance for Breaking ball Prediction

A graph with blue bars

Description automatically generated

**3.** Feature Importance for Off-Speed Prediction

**A graph with blue bars

Description automatically generated**

**Limitations  
Data Limitations**

* The historical data may not capture all variables influencing pitch types, such as recent changes in player performance or pitcher strategies.
* Missing values were handled but may still affect model reliability.

**Model Limitations**

* The model may overfit the training data, leading to less accurate predictions on unseen data.
* Predictions are probabilistic and do not guarantee specific outcomes in games.

**Generalization**The model's predictions are based on historical data, and future game conditions (e.g., player injuries, and changes in coaching strategies) may impact its accuracy.

**Average Pitch type predictions : Below is an illustration of Average Pitch Type predictions  
A graph of different colored rectangles

Description automatically generated**

**Recommendations**

* Utilize model predictions to inform strategic decisions regarding player matchups.
* Consider additional data collection on recent player performance for more accurate predictions.
* Regularly update the model with new data to enhance its predictive capabilities.

**Conclusion**The pitch type prediction model provides valuable insights that can significantly impact game strategy for the Cincinnati Reds. By leveraging historical data and advanced modeling techniques, the team can make informed decisions to improve performance on the field.