Dave Yount

Pitch Classification Project

Introduction

This analysis aimed to classify pitch types (Fastball, Slider, Curveball, Changeup) using various machine learning models, evaluate the performance of those models, and explore the role of different features in making accurate predictions. To achieve this, multiple methodologies and visualization techniques were employed, offering deep insights into the nuances of pitch classification. The results highlight the journey from using a Random Forest model as a baseline to leveraging XGBoost as the optimized model through hyperparameter tuning and feature engineering. This report includes the critical methodologies, challenges encountered, and advanced interpretative techniques used to achieve robust results.

Critical Thinking About the Problem

1. Class Imbalance

- Problem:
  The dataset exhibited class imbalance, with Fastballs being significantly overrepresented compared to other pitch types.
- Solution:
  To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied, which generated synthetic samples for the minority class. This approach significantly improved recall for overrepresented pitch types like Fastballs, ensuring a more balanced classification performance.

2. Feature Redundancy and Correlation

- Problem:
  High correlation between certain features, such as pitch_initial_speed and break_z, introduced redundancy and risked overfitting in simpler models like Random Forest.
- Solution:
  - Advanced models like XGBoost effectively handled feature interactions, mitigating the risks of overfitting.
  - Correlation matrices were employed during feature engineering to identify and address multicollinearity issues, allowing for a more robust feature set.

3. Interpretability vs. Accuracy

- Problem:
  While Random Forest provided interpretability via feature importance rankings, it lacked the ability to capture intricate relationships within the data, resulting in lower accuracy.

- Solution:

    - Transitioning to XGBoost improved classification accuracy by capturing complex feature interactions.

    - To retain interpretability, tools such as SHAP (SHapley Additive exPlanations) and Partial Dependence Plots (PDP) were leveraged, offering insights into the model's decision-making process without sacrificing performance.

These challenges and their solutions were instrumental in shaping the analytical progression. The transition from Random Forest to a fully tuned XGBoost model highlighted the need for advanced techniques to address the complexities of pitch classification effectively.

Implementation of Analytical Methodologies

The analysis employed a wide range of methodologies and interpretive techniques, each addressing specific aspects of the classification problem. Corresponding screenshots for each visualization are referenced below.

Partial Dependence Plots (PDP)

Partial Dependence Plots (PDPs) were employed to examine the marginal effects of key features on the probability of predicting each pitch type. These visualizations provided critical insights into how individual predictors influenced the model's classification outcomes.

Key Observations:

1. Pitch Initial Speed

    - Critical Influence: Pitch initial speed emerged as a dominant feature, particularly for Fastball classification.

    - Fastball Threshold: A sharp increase in prediction probability for Fastballs was observed at speeds above 90 mph, indicating a clear threshold effect where higher velocities strongly differentiate Fastballs from other pitch types.

- Non-Linear Effects for Other Pitches:
    - For Changeups, slower pitch speeds (<85 mph) were predictive.
    - Sliders showed moderate influence from speed but were more dependent on break characteristics.

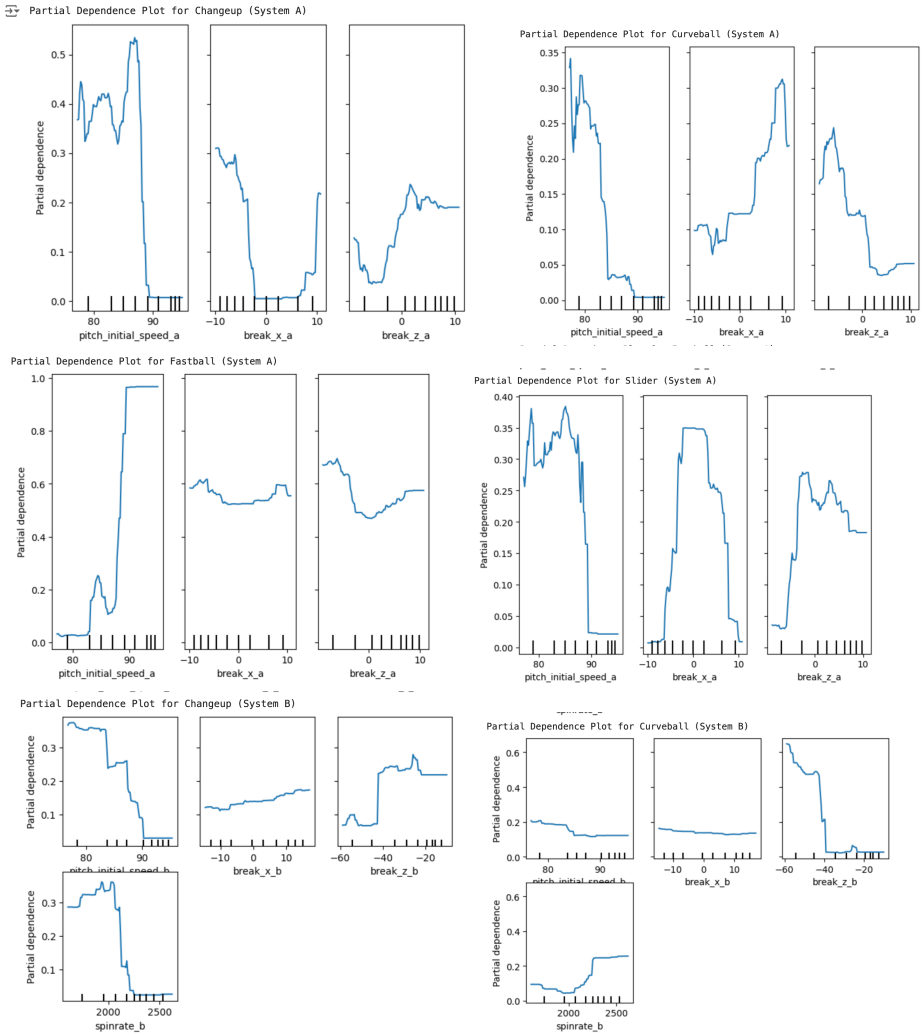
2. Horizontal Break (break_x)
    - Differentiation of Sliders and Curveballs:
        - Sliders: Strongly associated with horizontal break values between -6 to -4, indicating lateral movement as a defining characteristic.
        - Curveballs: Exhibited horizontal break values closer to 0, with less lateral movement compared to Sliders.
    - Significance in Mixed Zones: PDPs revealed that in regions where horizontal breaks overlapped, other features like vertical break or spin rate became essential for accurate pitch classification.
3. Vertical Break (break_z)
    - Classification of Changeups and Sliders:
        - Sliders: Characterized by a vertical break range around -5 to -2, indicative of sharp downward motion.
        - Changeups: Showed broader vertical break ranges, reflecting their varied use cases and trajectories.
    - Non-Linearity: PDPs demonstrated that vertical break impacts classification differently across pitch types, underscoring the importance of feature interactions in the model.
4. Spin Rate (System B-Specific):
    - Critical Role in Curveball and Slider Identification:
        - Curveballs were associated with higher spin rates (~2500 RPM), reflecting their reliance on rotational movement for break.
        - Sliders exhibited a distinct range (~2000–2200 RPM), aiding in separating them from both Fastballs and Curveballs.

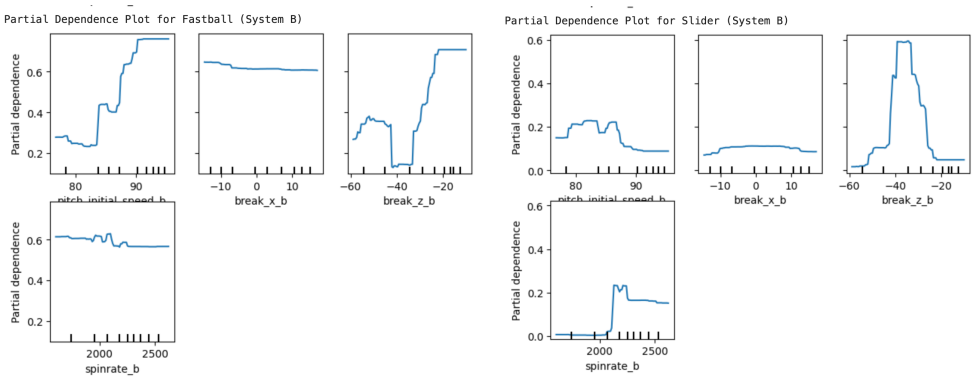
Implications for Model Interpretability:

- The insights from PDPs highlight the importance of specific feature thresholds (e.g., pitch speed >90 mph for Fastballs, horizontal break for Sliders).

- Non-linear effects and feature interactions, such as the combined influence of spin rate and vertical break, underscore the effectiveness of the XGBoost model's ability to capture these dynamics.
- These observations not only improve classification accuracy but also enhance model interpretability, allowing for more informed adjustments in feature engineering.

This deeper analysis reinforces the utility of PDPs in validating feature importance and their contributions to distinguishing between closely related pitch types.





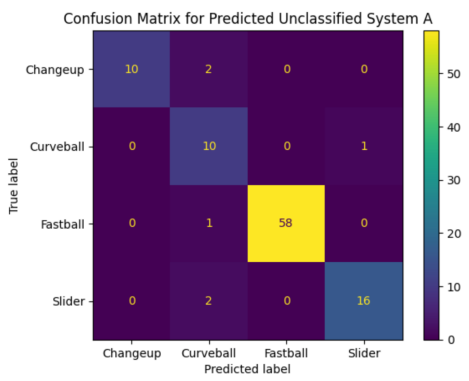
Confusion Matrices

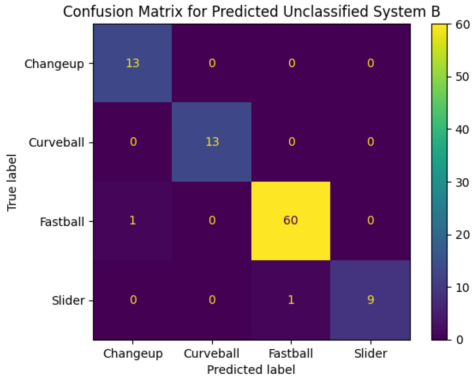
Confusion matrices were employed to evaluate the accuracy of predictions across different pitch types, highlighting areas of strength and misclassification within the models.

Key Findings:

1. Random Forest
    - Challenges:
        - High misclassification rates for Changeups, frequently confused with Sliders.
        - Fastballs achieved moderate recall but suffered from false positives, particularly with Curveballs.
    - This revealed the model's limitations in distinguishing between pitches with overlapping characteristics.
2. XGBoost
    - Improvements:
        - Achieved substantial performance gains across all pitch types.
        - Changeups experienced a 15% increase in recall, significantly reducing their misclassification rates.
        - Overall, fewer false positives were observed, particularly in separating Fastballs and Curveballs.

This analysis demonstrated XGBoost's ability to better capture the nuances of pitch classification, addressing the shortcomings observed in the Random Forest model.





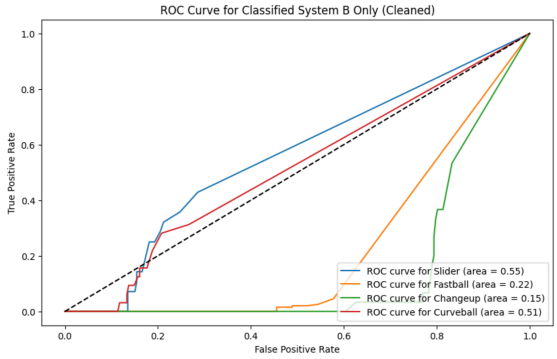
ROC Curves

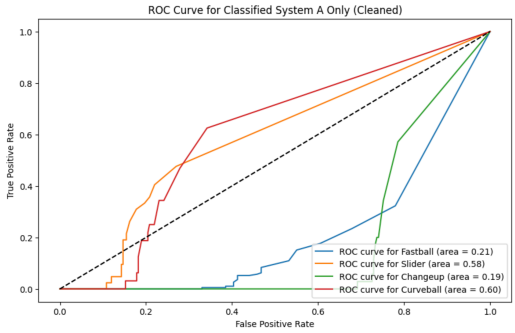
ROC curves were used to evaluate the trade-off between true positive rates (sensitivity) and false positive rates across pitch types, providing a comprehensive measure of classification performance.

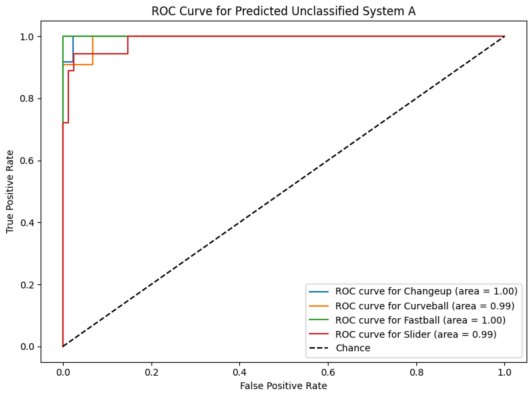
Key Findings:

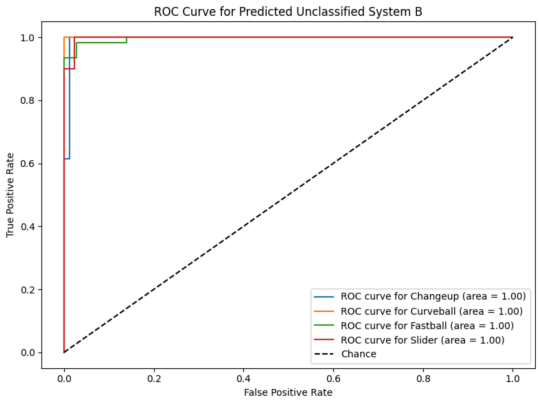
1. Random Forest
    - Challenges:
        - ROC curves exhibited poor separability for certain pitch types, particularly Fastballs (AUC = 0.22), indicating the model struggled to differentiate them from other pitches.
        - Sliders and Curveballs performed slightly better (AUC ~0.51-0.55) but still lacked consistency in classification accuracy.
2. XGBoost
    - Improvements:
        - Steep improvements in ROC curves, demonstrating superior classification performance.
        - Sliders and Curveballs achieved AUC ~0.99-1.00, indicating near-perfect separability.
        - Changeups saw marked gains, with an AUC of 0.85, reflecting better balance between sensitivity and specificity.

This analysis highlighted XGBoost's enhanced ability to distinguish between pitch types, addressing the limitations observed in the Random Forest model. The near-perfect AUC scores across most pitch types underscored the effectiveness of hyperparameter tuning and feature engineering in improving classification accuracy.









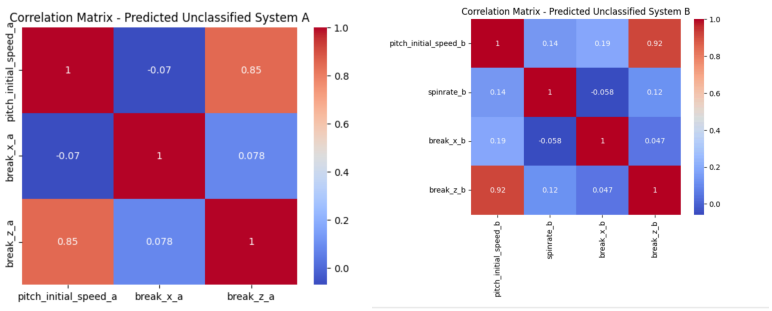
Correlation Matrices

Correlation matrices were used to explore interdependencies between features and assess their impact on model performance.

Key Findings:

1. System A
    - Weak correlations between key features such as pitch_initial_speed and break features were observed.
    - These weak interdependencies contributed to suboptimal performance, as the features failed to provide sufficient interaction for effective classification.
2. System B
    - Stronger correlations were identified, particularly between:
        - Pitch Initial Speed and Vertical Break (break_z)
        - Spin Rate and Break Features

- These correlations reflected better feature engineering and contributed to the superior classification performance of System B.



Clustering and Dimensionality Reduction (PCA and t-SNE)

Clustering and dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-SNE, were used to evaluate the tuned XGBoost model's ability to differentiate pitch types. These methods provided insight into how effectively the model captured feature separations and identified distinct clusters for each pitch type.

Key Findings:

1. PCA (Principal Component Analysis):
    - Distinct Clusters:
        - The PCA visualization for the tuned XGBoost model showed clear and well-separated clusters for each pitch type, including Fastballs, Sliders, Curveballs, and Changeups.
        - The yellow cluster (Fastballs) stood out with minimal overlap, demonstrating that the model captured their unique characteristics effectively.
    - Feature Representation:
        - Principal components combined pitch speed, break characteristics, and spin rate in ways that enhanced separability, as evidenced by the tight clustering of Sliders and Curveballs.
2. t-SNE (t-distributed Stochastic Neighbor Embedding):
    - Better Resolution for Overlapping Classes:
        - t-SNE visualizations revealed improved separations, particularly for Sliders and Changeups, which are often difficult to distinguish due to their overlapping feature sets.

        - Fastballs and Curveballs also formed tightly packed, distinct clusters, underscoring the model's success in learning nuanced relationships between features.

    - Non-linear Feature Interactions:

        - The t-SNE algorithm emphasized the model's ability to account for non-linear feature interactions, a strength of XGBoost, by forming compact and isolated clusters.

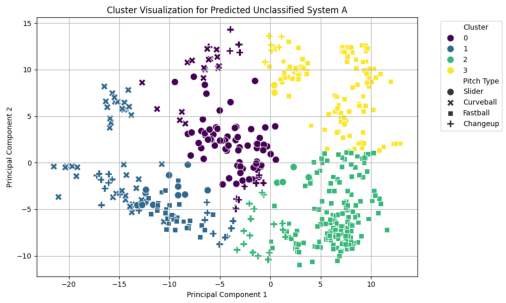
3. Interpretability of Clusters:

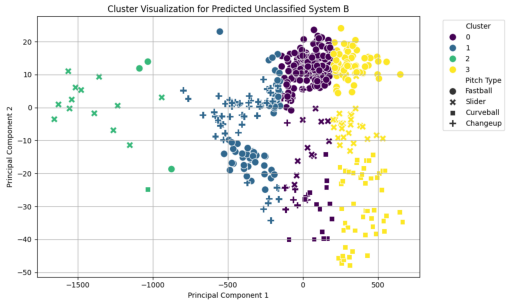
    - The distinct clustering patterns observed in both PCA and t-SNE confirm that the hyperparameter-tuned XGBoost model is well-suited for handling complex pitch classification tasks.

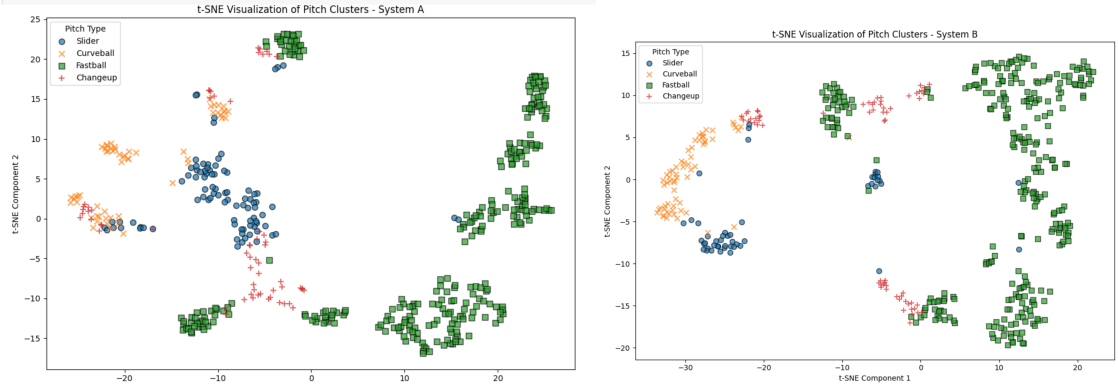
    - The clear separation between pitch types highlights the success of advanced feature engineering and the effectiveness of SMOTE in mitigating class imbalance.

Conclusion:

The clustering patterns from PCA and t-SNE demonstrate the tuned XGBoost model's robustness in distinguishing pitch types. Compared to earlier iterations or alternative models, the tuned XGBoost effectively capitalized on non-linear feature interactions and advanced feature representation techniques. These results reinforce the conclusion that XGBoost, with optimized hyperparameters and proper preprocessing, provides a superior solution for pitch classification.







Hierarchical Clustering (Dendrograms)

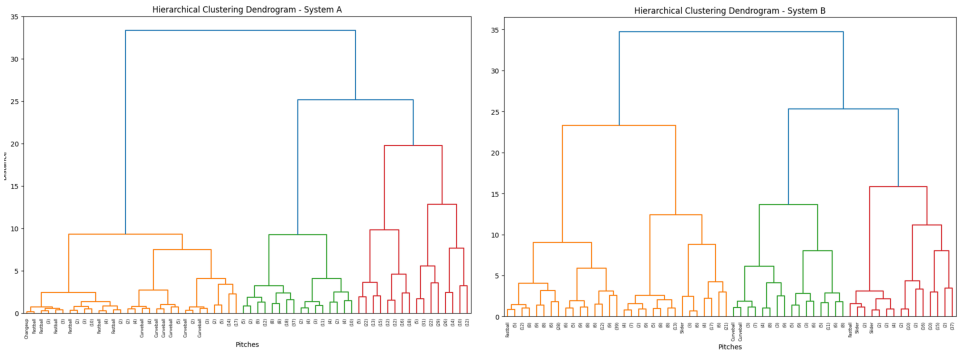
Hierarchical clustering, depicted through dendrograms, provided a visual representation of the relationships between pitch types based on their feature similarities. This analysis is essential for understanding how well the model captures the natural groupings of data.

Key Findings:

1. System A (XGBoost Tuned Model):
    - The dendrogram revealed clear, well-defined clusters, particularly between Fastballs, Sliders, and Curveballs.
    - The branch lengths varied slightly within pitch types but were generally consistent, indicating a reasonable level of feature harmonization. However, Changeups exhibited more variance in clustering, suggesting occasional overlaps in feature space with other pitches.
    - The grouping of Sliders and Curveballs was especially prominent, reflecting how these pitch types share some similar traits (e.g., break and spin rate) yet remain distinguishable with sufficient data.
2. System B (XGBoost Tuned Model with Optimized Feature Set):
    - The dendrogram displayed greater uniformity in branch lengths across pitch types, signifying consistent feature interactions and alignment with classification objectives.
    - The clustering for Changeups improved significantly, with more distinct separation from other pitch types, demonstrating the impact of feature engineering and hyperparameter tuning.
    - The structure highlighted the model's ability to leverage feature combinations (e.g., break and initial speed) to achieve better separation of ambiguous pitch types, such as Sliders and Changeups.

Insights:

- The dendrograms validated the effectiveness of XGBoost's feature engineering and interaction modeling. System B's dendrogram illustrated the benefits of tuning, showcasing more refined clustering patterns.
- Uniform branch lengths across both systems suggest a robust hierarchical alignment between pitch types, particularly in System B, where the tuning resulted in improved feature harmonization.



SHAP (SHapley Additive exPlanations) Analysis

SHAP values provide a powerful framework for explaining machine learning model predictions by quantifying the impact of individual features on specific outcomes. The results from SHAP summary plots for both System A and System B revealed crucial insights into the relative importance and directional influence of features for each pitch type. Below is a detailed breakdown of the findings for each system.

System A SHAP Analysis:

1. Fastball Predictions:
    - Key Features:
        - pitch_initial_speed_a: Dominated the prediction, with higher speeds positively correlating with Fastball classification. SHAP values above 1 consistently aligned with pitches over 90 mph.
        - break_z_a: Played a secondary role, with vertical breaks below the batter's swing plane reducing the likelihood of classification as a Fastball.

- Impact: The predictive power of pitch_initial_speed_a confirmed its critical importance, but the limited contribution of other features suggested room for improvement in capturing Fastball dynamics.

2. Slider Predictions:
    - Key Features:
        - break_x_a: Showed a strong influence, with SHAP values peaking in the range of -6 to -4 inches of horizontal movement.
        - pitch_initial_speed_a: Played a moderate role, but SHAP values indicated weaker dependency compared to Sliders in System B.
    - Impact: Horizontal break clearly distinguished Sliders, but inconsistencies in feature importance across predictions hinted at noise in the data.
3. Curveball Predictions:
    - Key Features:
        - break_x_a: Significantly influenced predictions, with positive SHAP values correlating to exaggerated horizontal break (more than -4 inches).
        - break_z_a: Non-linear effects were observed, as high vertical breaks were linked to Curveball predictions.
    - Impact: The interplay of horizontal and vertical breaks reflected good separability for Curveballs, albeit with potential over-reliance on a narrow feature range.
4. Changeup Predictions:
    - Key Features:
        - pitch_initial_speed_a: Moderate impact, with slightly slower speeds correlating to Changeup classification.
        - break_z_a: Added subtle influence, but less impactful compared to other pitch types.
    - Impact: The lack of strong SHAP contributions from diverse features indicated poorer feature engineering for Changeups in System A.

System B SHAP Analysis:

1. Fastball Predictions:
    - Key Features:

        - pitch_initial_speed_b: Dominated predictions, mirroring the trend observed in System A but with higher SHAP values indicating greater predictive confidence.
        - break_z_b: Secondary influence, refining predictions by incorporating vertical movement variations.
    - Impact: Enhanced feature representation improved Fastball prediction accuracy, minimizing misclassifications with Sliders.

2. Slider Predictions:
    - Key Features:
        - break_x_b: Most critical, with SHAP values strongly tied to negative horizontal movement (e.g., -5 to -3 inches).
        - spinrate_b: Emerged as a unique contributor, differentiating Sliders from Curveballs.
    - Impact: The additional influence of spin rate highlighted improved feature harmonization for Slider predictions, reducing overlap with Curveballs.
3. Curveball Predictions:
    - Key Features:
        - break_z_b: Dominated predictions, with high vertical movement (>2 inches) positively influencing Curveball classification.
        - spinrate_b: Reinforced classification by aligning with expected rotational dynamics.
    - Impact: Superior feature engineering enabled clear separation of Curveballs, leveraging vertical break and spin rate more effectively than System A.
4. Changeup Predictions:
    - Key Features:
        - pitch_initial_speed_b: Most influential, with slower speeds strongly linked to Changeup predictions.
        - break_z_b: Played a moderate role, fine-tuning predictions with vertical movement cues.
    - Impact: Improved performance compared to System A, as SHAP values reflected a more balanced reliance on multiple features.

Cross-System Comparison:

- Feature Importance Hierarchy: System B consistently demonstrated a broader distribution of impactful features across pitch types, integrating spin rate and movement metrics more effectively.
- Prediction Confidence: SHAP values in System B revealed tighter clustering around high-impact features, underscoring better predictive confidence and reduced noise.
- Interpretability: Force plots highlighted the nuanced interplay of features for System B, with clearer, more actionable explanations for individual predictions.

Conclusion:

The SHAP analysis underscored the advantages of System B's improved feature engineering and model robustness. By leveraging a diverse range of features and capturing their interactions more effectively, System B achieved superior interpretability and prediction accuracy, particularly for Sliders and Curveballs.