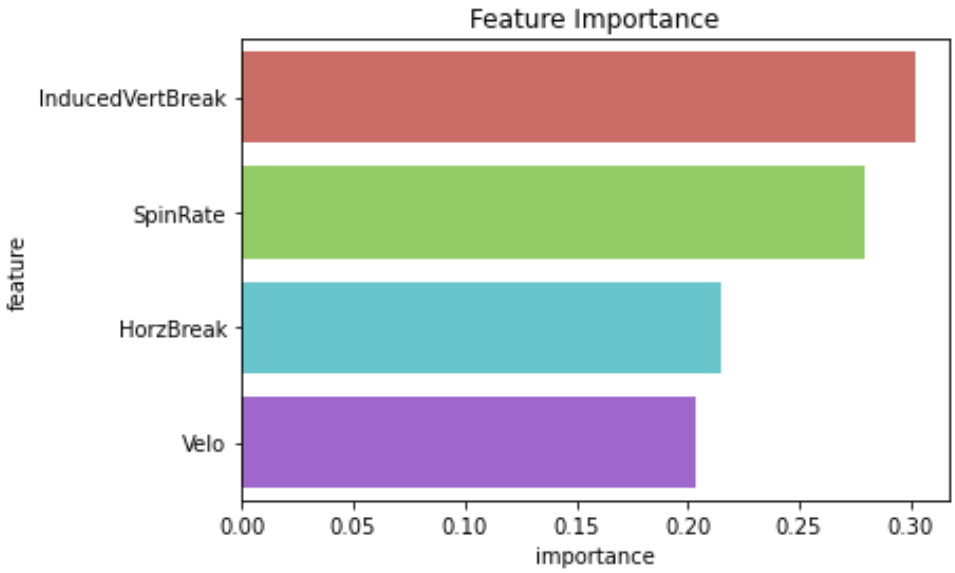
Fastball Factor Analysis

1. Predict the chance of a pitch being put in play. Please use this model to predict the chance of each pitch in the "deploy.csv" file being put in play and return a CSV with your predictions.
    - See `deploy_pred.csv` for hard prediction
    - See `deploy_proba.csv` for probablity prediction
2. In one paragraph, please explain your process and reasoning for any decisions you made in Question 1.
    - ETL
        - **KNN Impute**: Since the `SpinRate` feature contains nearly 20 rows with missing values, we needed to impute this data. The KNN imputer ensures that the most contextually appropriate average value is selected based on KNN clustering.
        - **SMOTE balance**: The distribution of data points for in-play versus not-in-play is approximately 3:7. This imbalance causes models to predominantly predict not-in-play outcomes, targeting an accuracy of 70% (highly biased towards the majority class). Therefore, I oversampled the dataset using the SMOTE method to increase the in-play data points, while preserving the feature patterns.
    - Metric Selection
        - **F1 score**: This metric strikes a balance between precision and recall, particularly important when dealing with imbalanced datasets where performance for the minority class is crucial.
    - Model Attempts
        - **Logistic Regression**: This is the foundational model, serving as the baseline for this prediction task.
        - **Gradient Boosting**: A renowned ensemble learning method that capitalizes on weak learners to enhance prediction accuracy via boosting techniques.
            - Other Tree-Based Models: RF, DT
        - **Neural Network**: Utilizes layers of nodes to discern non-linear boundaries.
        - Etc.
    - Best Model
        - **Gradient Boosting**: F1 = 0.42; Accuracy = 0.55

3. In one or two sentences, please describe to the pitcher how these 4 variables affect the batter's ability to put the ball in play. You can also include one plot or table to show to the pitcher if you think it would help.



- The batter's ability to put the ball in play is most influenced by the pitch's vertical break, followed closely by its spin rate, horizontal break, and velocity, with each factor altering the ball's trajectory and the batter's predictability in making contact.

4. In one or two sentences, please describe what you would see as the next steps with your model and/or results if you were in the analyst role and had another week to work on the question posed by the pitcher.

    - Given another week in the analyst role, I would employ a grid search for hyperparameter tuning across all models to optimize their performances and ensure a comprehensive comparison. Furthermore, I'd enhance the visualization aspect, providing intuitive graphics to help the pitcher more clearly grasp and interpret the results.