STRIKEOUT SENSEIS

By Nathan, Yasmine, Andy, and Esmeralda

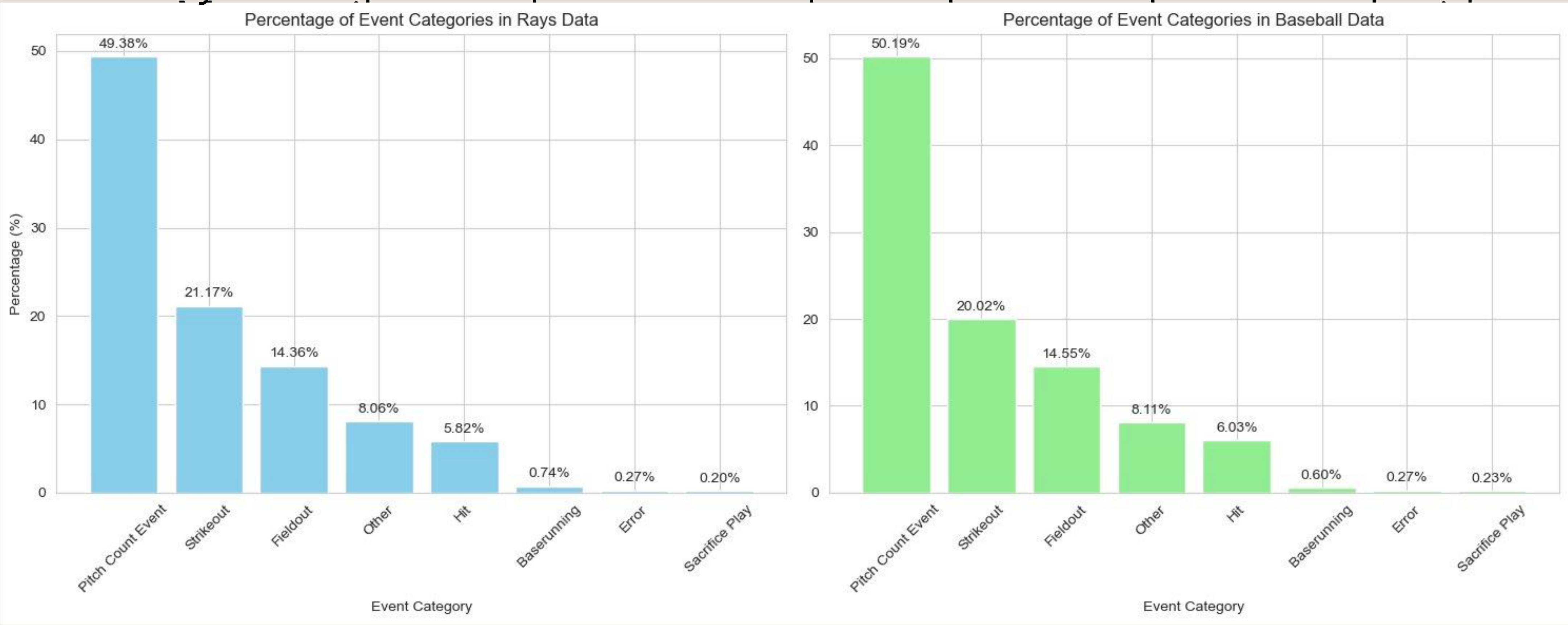
Research Question

HOW DO PITCH CHARACTERISTICS SUCH AS ***RELEASE SPEED***, ***SPIN RATE***, ***HORIZONTAL BREAK***, ***INDUCED VERTICAL BREAK***, *AND* ***PLATE LOCATION*** IMPACT THE LIKELIHOOD OF GENERATING STRIKEOUTS IN 2 STRIKE COUNTS WITH 0 OR 1 BALLS FOR RIGHT-HANDED PITCHERS THROWING FASTBALLS?

Justification/Goals

The Rays are good in this scenario, but can always be better

Justification (for scenario)



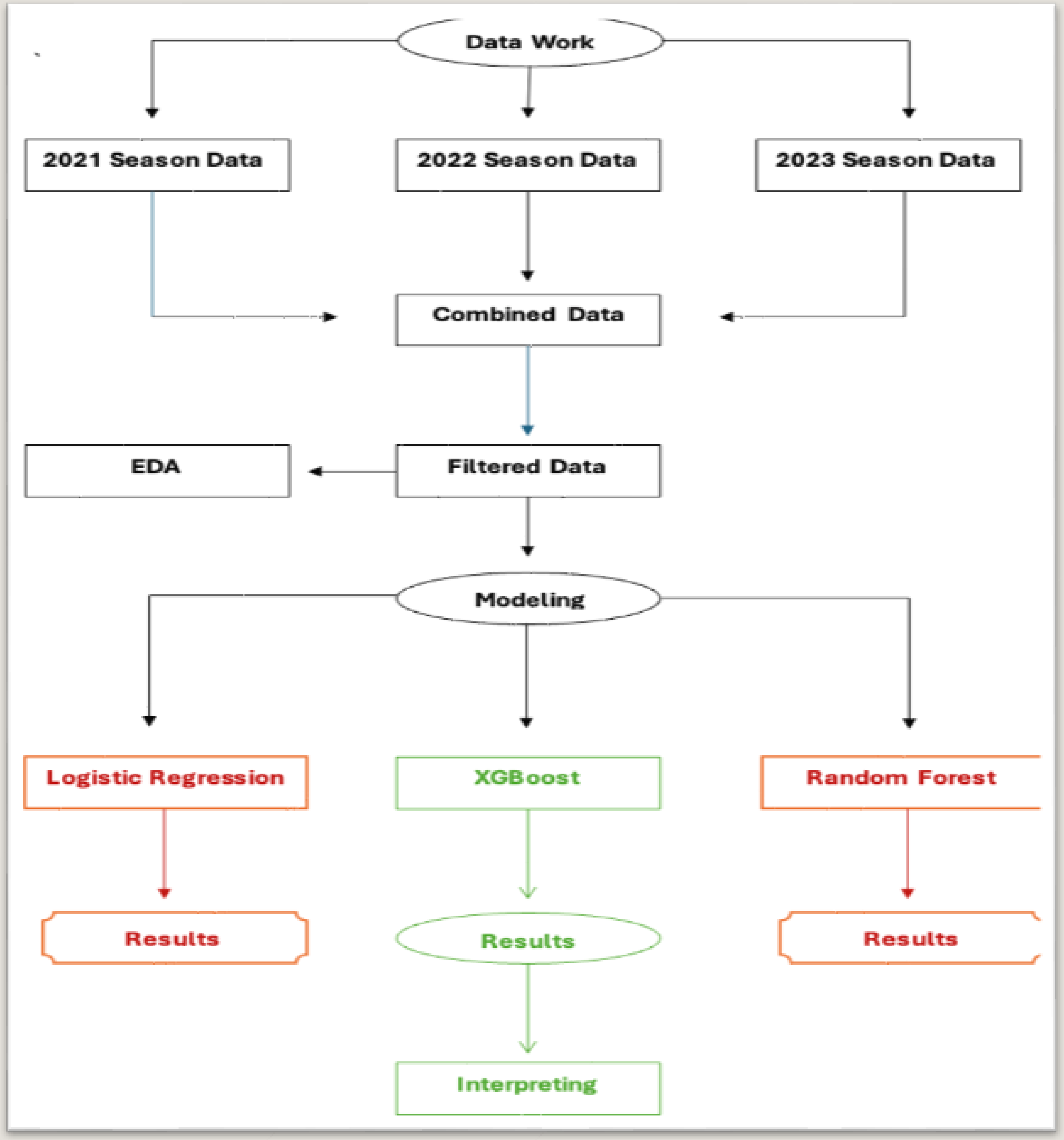
Initial Predictors

- **Release speed:** the miles per hour velocity of the ball as it approaches home plate
- **Spin rate:** revolutions per minute of the ball as it approaches home plate
- **Horizontal break:** horizontal movement of a ball
- **Induced vertical break:** pitcher's contribution over the vertical movement of a ball
- **Plate location side and height:** where the pitch lands in reference to the strike zone parameters

Target Variable

- **Strikeout** (1); classified as "strikeout" and "strikeout double play"
- **No strikeout** (0); classified as every other outcome, from the "eventtype" column

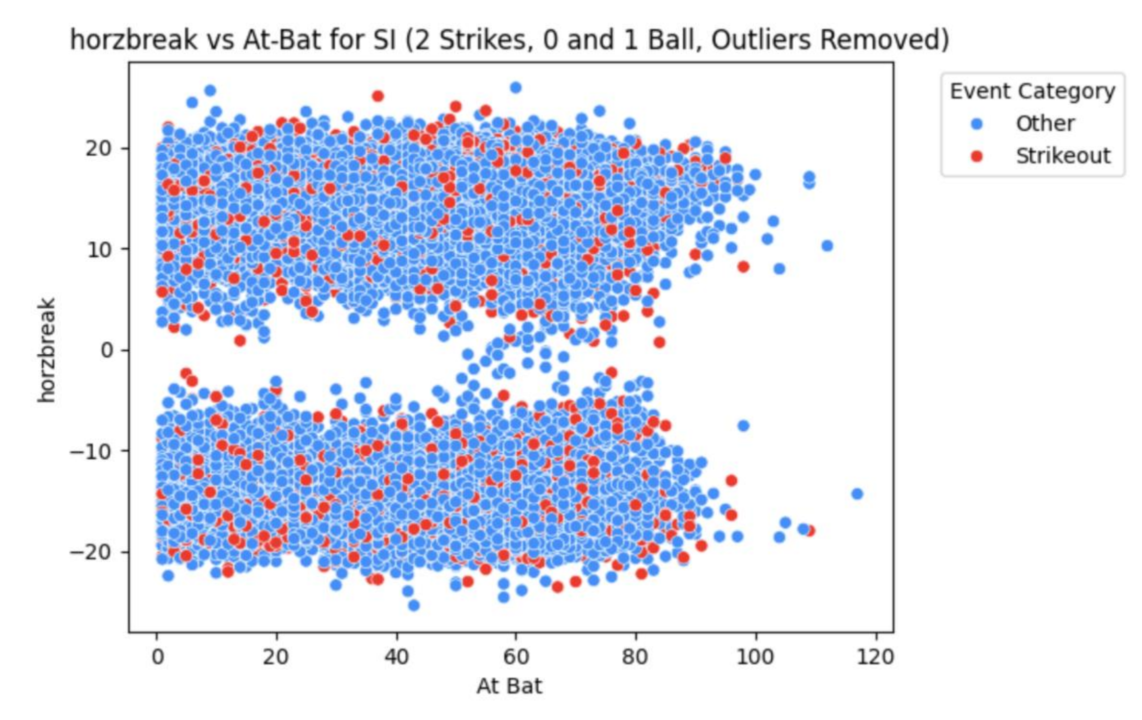
Process

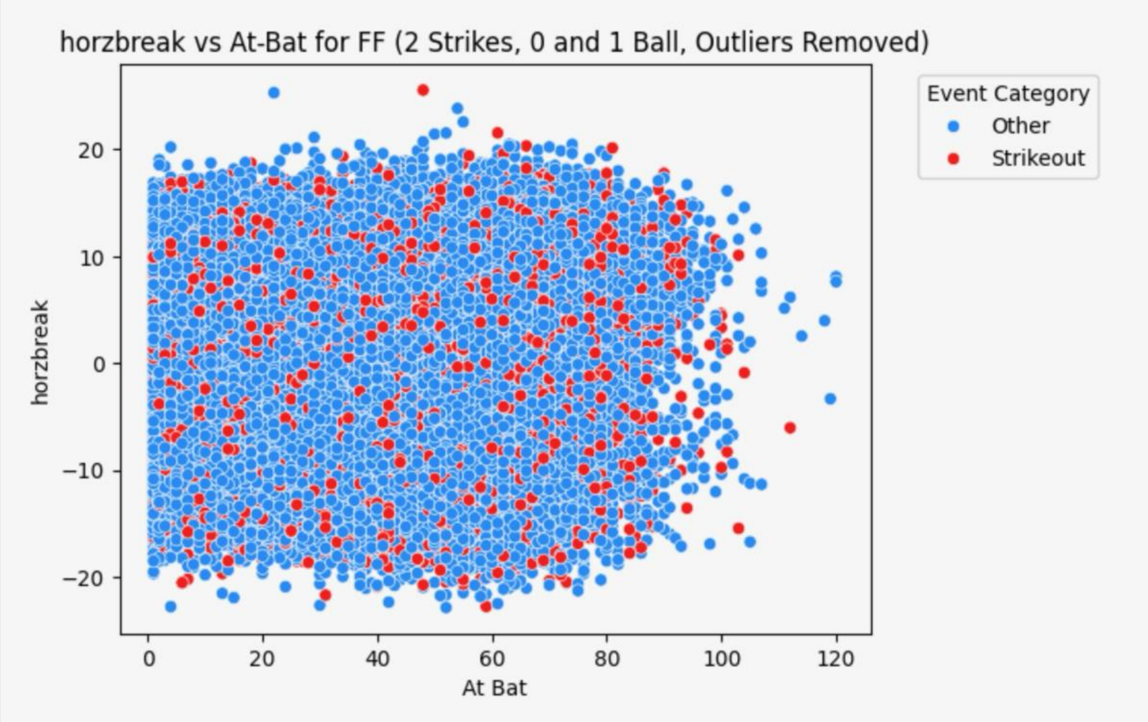


Data Work
2021 Season Data
2022 Season Data
2023 Season Data
Combined Data
EDA
Filtered Data
Modeling
Logistic Regression
XGBoost
Random Forest
Results
Results
Results
Interpreting

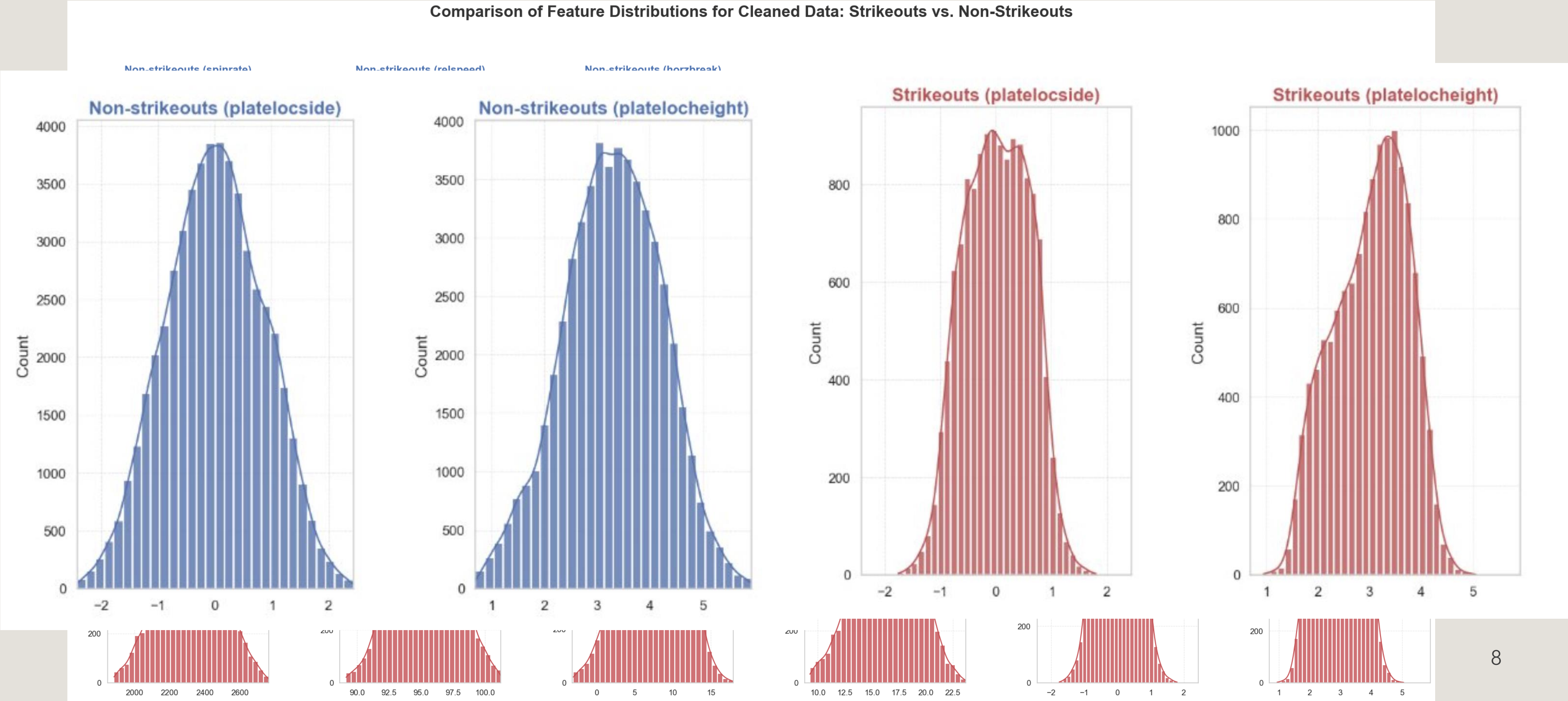
Data Exploration

- We centered the model on fastballs because it is the most common pitch type and allows us to get specific contributions to this pitch type.
- Focusing on right-handed pitchers to eliminate the impact of confounding variability.





Data Exploration Continued



Data Cleaning

- Kept the pitch features and the event type column
- Filtered out pitch outs and balls in the dirt, as these types of throws can be outliers in our data.
- Focused on right-handed pitchers to eliminate the impact of confounding variability
- Focused on pitchers with > 100 appearances in our scenario
- Removed outliers
- This leaves us a dataset with 47,828 rows

eventtype	spinrate	relspeed	horzbreak	inducedvertbreak	platelocside	platelocheight
double	2154.819580	93.462410	13.426973	11.059589	-0.826248	2.734800
ball	2176.248779	94.833099	3.791252	10.746025	0.064407	3.774802
swinging_strike	2037.569946	93.123344	12.063048	9.522058	-0.639474	2.198678
foul	1999.427734	93.897316	3.704415	11.381925	0.084060	3.053724
walk	2491.611084	95.446953	3.417144	18.279982	0.663778	1.232030

Feature Engineering

Relspeed_diff:

- This is the difference between a pitcher's average release speed and that observation's release speed.
- This variable aims to capture relationships that are not immediately available with the original *relspeed* variable.
- Top feature in terms of predicting

$$relspeed_{average} - relspeed_{observed}$$

Relspeed_inducedvertbreak:

- This was simply an interaction term created by multiplying release speed and induced vertical break together.
- We were interested in this result, as we had previously found a relatively strong linear correlation of 0.75 between induced vertical break and release speed when we filtered the data by only outs.
- Was not able to predict.

$$relspeed \cdot inducedvertbreak$$

XGBoost Model

HOW THE MODEL WORKS

- XGBoost, or eXtreme Gradient Boosting, is a machine learning algorithm that builds an ensemble of decision trees to make predictions
- It’s called "boosting" because it builds trees one by one, with each new tree focusing on fixing the errors made by the previous ones

WHY IT'S USEFUL FOR OUR DATA

- Since errors are often higher for the minority class in our imbalanced data, new trees are likely to focus on those harder-to-predict instances
- This process of iteration makes XGBoost good at picking up on patterns in the minority class (strikeouts)

XGBOOST EVALUATION METRICS

- **Precision**: Out of all the strikeouts predicted, what percentage are truly strikeouts
  - 62% of the strikeouts the model predicts are correct
  - **Precision Equation** = TP / (TP + **FP**)

- **Recall**: Out of the total strikeouts, what percentage are predicted as a strikeout
  - The model successfully identifies 86% of all true strikeout
  - **Recall Equation** = TP / (TP + **FN**)

- **ROC AUC**: How good the model is at distinguishing between strikeouts and no strikeouts

Metric	Value
Precision	0.62
Recall	0.86
ROC AUC	0.73

SHAP Values

Purpose of SHAP

- Helps understand how each feature increases or decreases the likelihood of a strikeout.
- Provides both the magnitude of feature importance and the specific influence of each feature on outcomes.

Key Insights

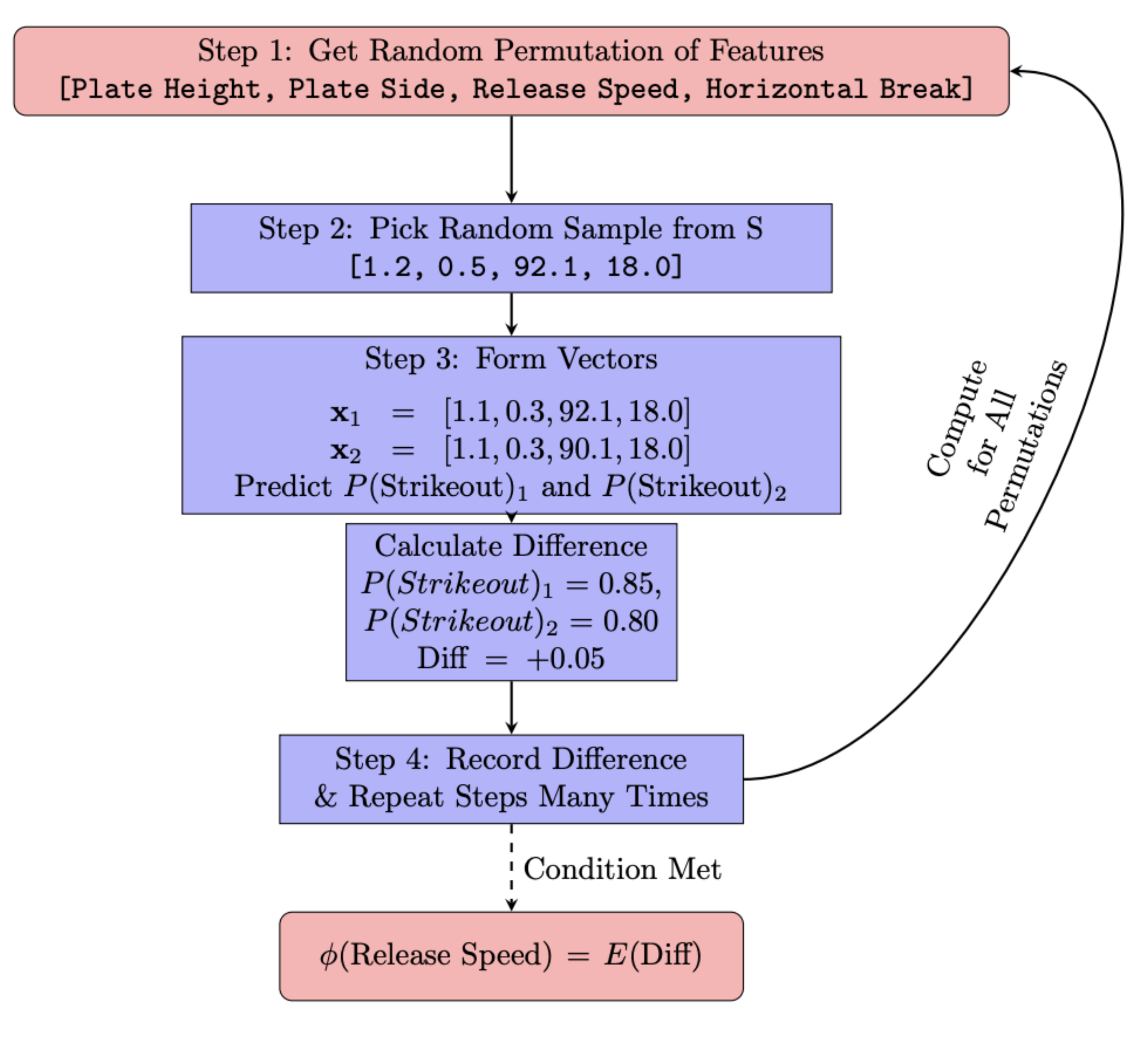
- Guides us in determining which characteristics contribute most to the success of selected pitchers.

Features' Equation:

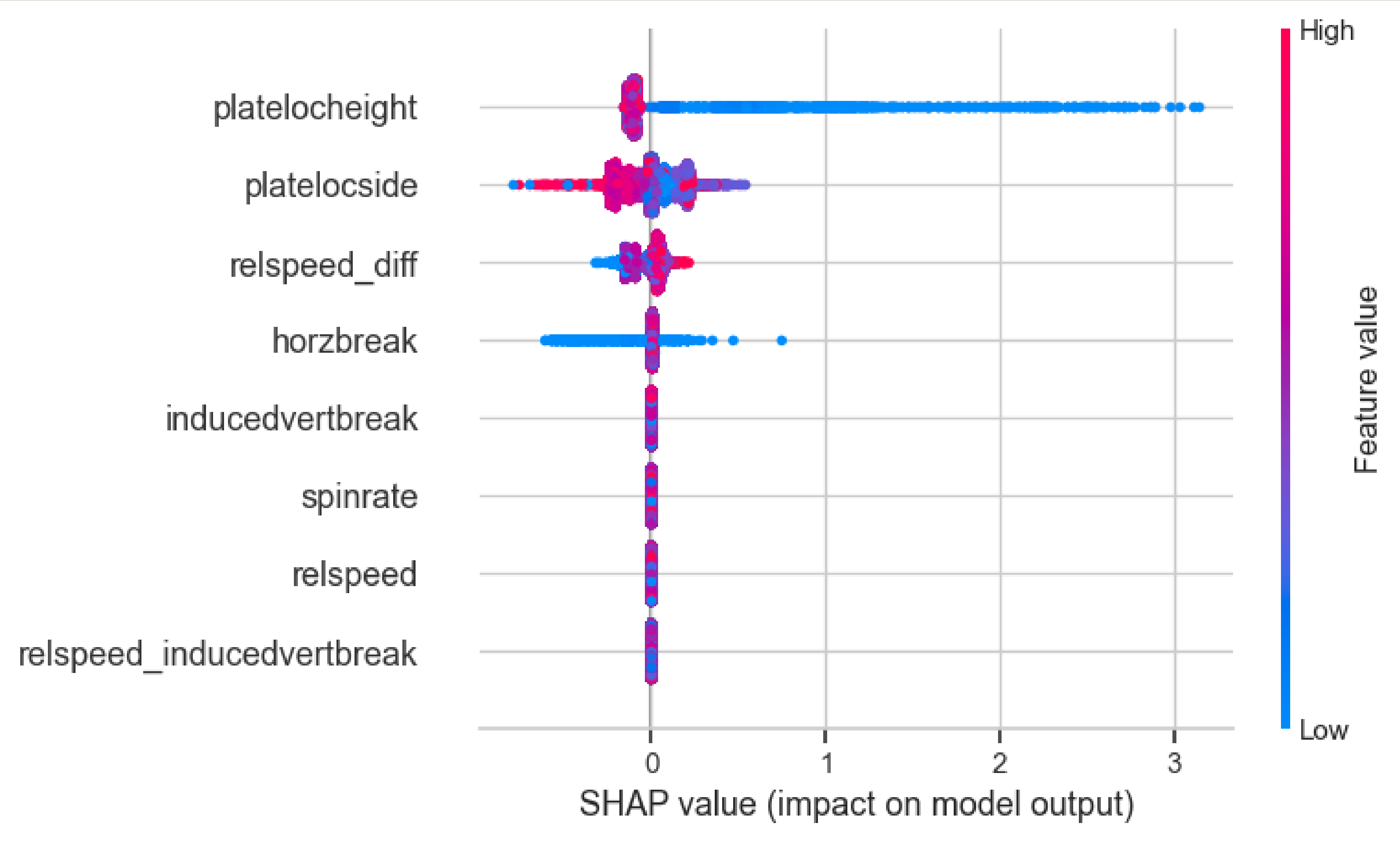
$$\boxed{\phi_i(v)} = \sum_{\boxed{S \subseteq N \setminus \{i\}}} \boxed{\frac{|S|!\,(|N| - |S| - 1)!}{|N|!}} \boxed{\left(v(S \cup \{i\}) - v(S)\right)}$$

- The SHAP value for a specific feature, {i}, given any model v: The **contribution** of a specific feature to the prediction.
- Considers all possible subsets of the set N (all features) that **excludes** the feature of interest {i}
- **The Weight -** Likelihood of a feature's contribution across all possible coalitions/combinations
- The **marginal contribution** of the feature {i} to the overall model, takes the **difference in the probability of predicting strikeout with/without {i}**

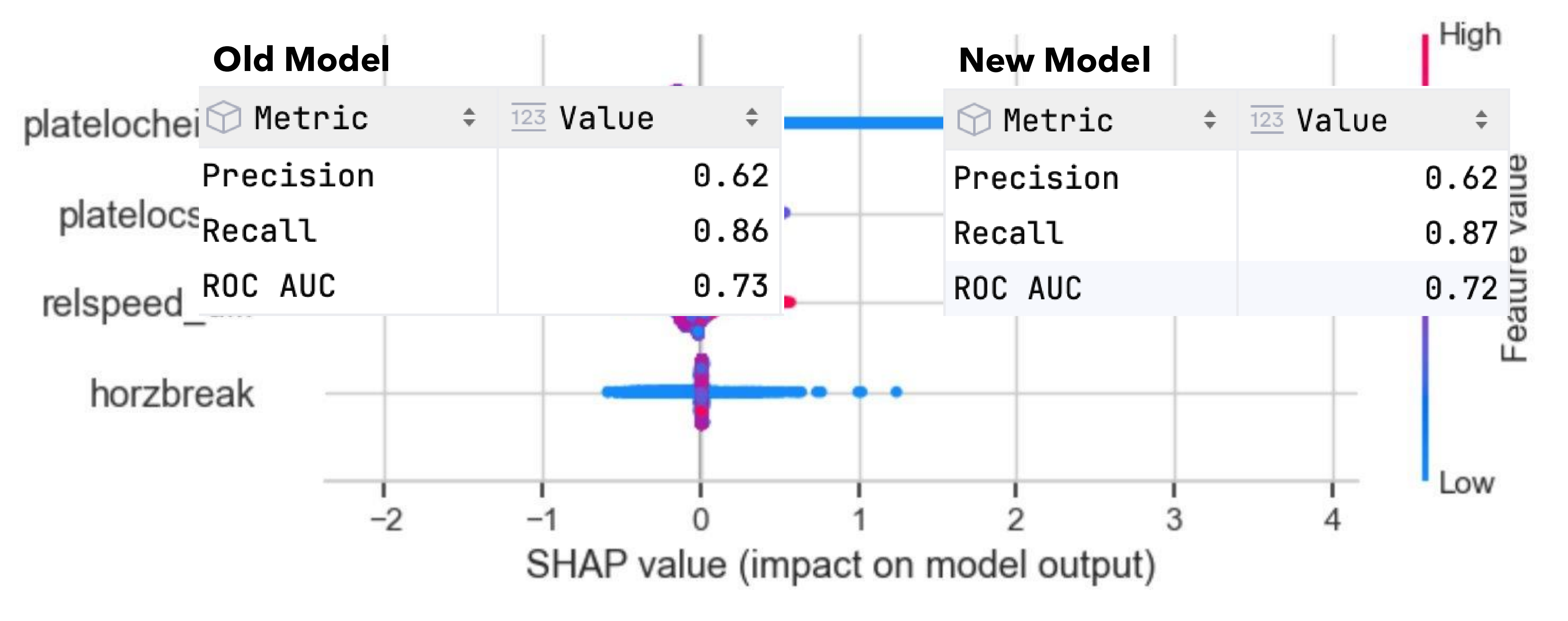
Example: Finding SHAP(Release Speed)



Results - Aggregated



Results - Aggregated



Old Model

Metric	Value
Precision	0.62
Recall	0.86
ROC AUC	0.73

New Model

Metric	Value
Precision	0.62
Recall	0.87
ROC AUC	0.72

Actual Value

P(Strikeout) - Log odds

Application - Interpretation

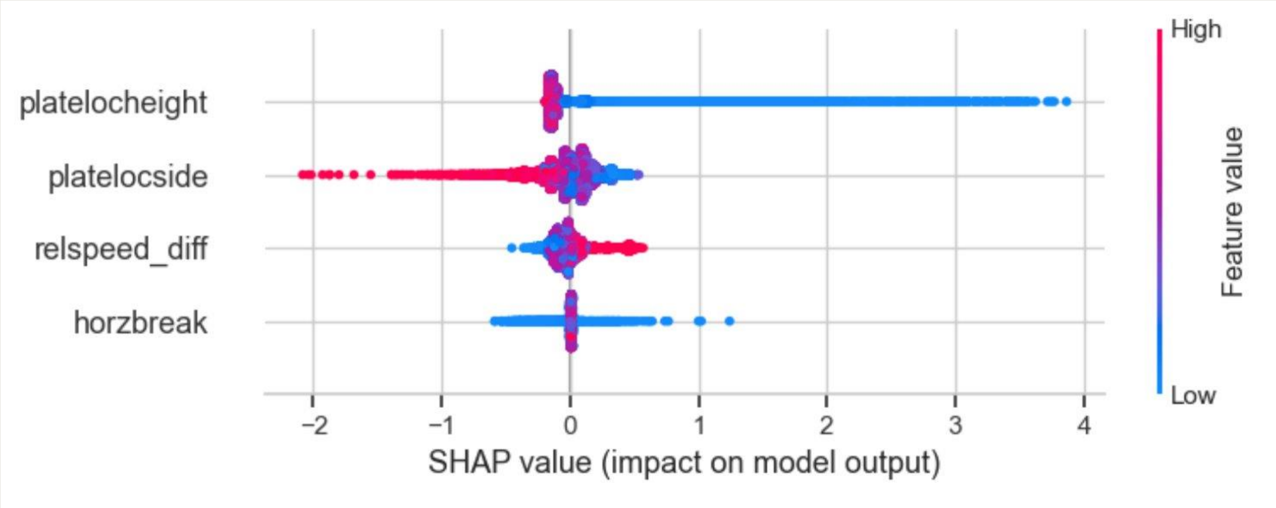
- Feature Importance – Plate location
- Feature impact

Platelocheight: lower plate locations (colored blue) lead to more strikeouts – attack the lower part of the strike zone

Platelocside: centered SHAP values indicate that slight deviances to the left side hold predictive power

Variation from the average fastball speed

Aggregated Model



Application - Rankings

Bottom 10

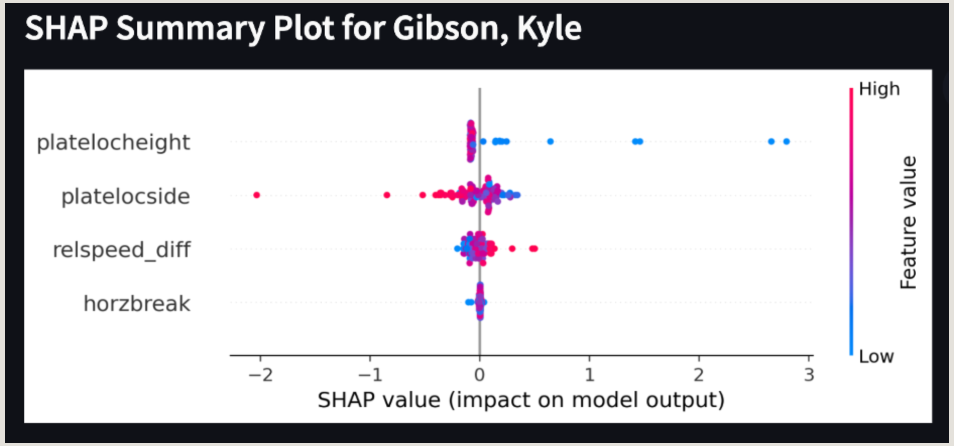
pitcher	strike_percentage
Gibson, Kyle	0.108434
Blackburn, Paul	0.134831
DeSclafani, Anthony	0.145833
Espino, Paolo	0.148148
Heasley, Jon	0.155844
Civale, Aaron	0.155963
Plesac, Zach	0.157025
Gray, Josiah	0.166667
Senzatela, Antonio	0.168142
Thompson, Zach	0.171429

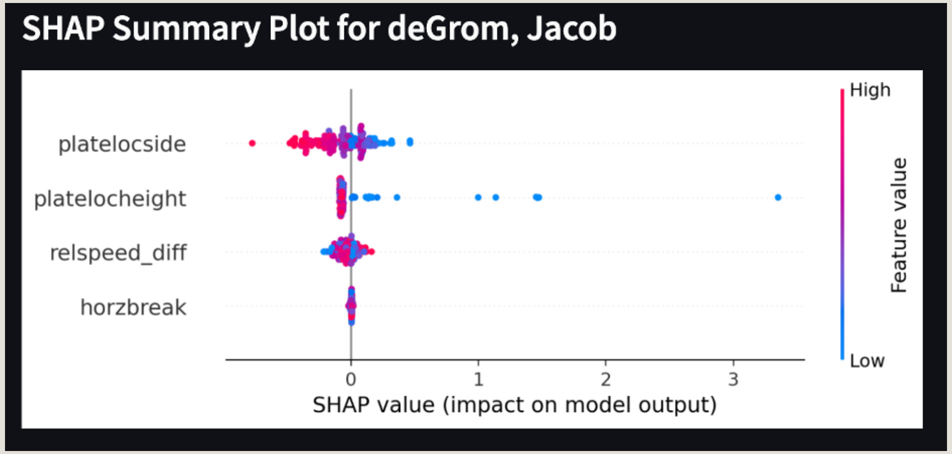
Top 10

pitcher	strike_percentage
deGrom, Jacob	0.573770
Fairbanks, Peter	0.522936
Gausman, Kevin	0.516014
Wick, Rowan	0.500000
Neris, Hector	0.491667
Crawford, Kutter	0.479675
Vest, Will	0.477273
Iglesias, Raisel	0.471154
Sewald, Paul	0.468354
Cisnero, Jose	0.467890

Application – Example

Kyle Gibson	Both	Jacob deGrom
Has a lot more non-strikeout predictions Plate Location Height was the most important predictor	*Same General Patterns* *Emphasis allows for feature importance*	Probability of strikeout is higher Plate Location Side was the most important predictor





DEMO

Limits and Beyond

Limitations:

- Size of data
- Number/Quality of predictors

Further Research/Applications:

- More advanced modeling to explore the interaction between characteristics
- Investigating developing player's weak spots with an emphasis on our certain characteristics

THANK YOU!