Whiff Probability as a Metric to Grade Pitch Quality

Josh Wintroub

**Executive Summary**

Objective: Create a new metric to grade the quality of a given pitch or pitcher based on trackman metrics. This metric will aid in recruiting pitchers and projecting performance in season based on pitch quality.

Key Findings:

* Using Random Forest feature selection, the model was able to determine that for fastballs, the most important metrics are Vertical Approach Angle (VAA), Induced Vertical Break (IVB), Release Speed (Velocity), Horizontal Approach Angle (HAA), Release Side, Horizontal Break, and Extension.
* Breaking Balls most important metrics are VAA, HAA, Velocity, IVB, Spin Rate, and Extension.
* Offspeeds most important metrics are VAA, HAA, Velocity, IVB, Horizontal Break, and Extension

Recommendations:

* Implement this model into recruiting practices in order to better decide what pitchers to recruit and spend NIL money on to get them to pitch for the school.
* Use insights from the model to better train pitchers to focus on the metrics that contribute more to swing and misses and therefore better pitch quality.

Scope:

There were 6 models created, a lefty and righty fastball, lefty and righty breaking ball, and lefty and righty off-speed. The data was acquired from trackman data of college baseball in 2024.

Limitations:

This model does not account for environmental factors, nor does it account for the location of the pitch. This was done to focus on the quality of the metrics of the pitch.

Decision Problem:

The decision problem addressed in this analysis is understanding and quantifying the factors that go into a pitch resulting in a swing and miss. By building a predictive model that identifies key pitch characteristics that influence the likelihood of a swing and miss being the result of the pitch. Using this model will help Kansas State Baseball more successfully recruit pitchers that are more likely to succeed at the Division 1 level. Additionally, this model can be used to determine if a change in a pitcher’s arsenal was a benefit to his pitch characteristics or not. This removes some guessing work out of coaching a pitcher’s change and can quantify how much of a difference the change will make to a pitcher’s success.

Data Collection:

The data source used for this analysis was sourced from Trackman, which provides advanced tracking data from baseball games. In order to create the data set, all division one baseball games from March 14th, March 22nd, and March 23rd were downloaded from the Trackman FTP server and merged into one CSV file. These dates were chosen at random, and in total 96,787 pitches are included in the analysis. From there the csv file was uploaded into Jupyter notebooks for cleaning and analysis.

In Jupyter, a second data frame was created which only included variables necessary for the analysis. The variables included are Pitcher, PitcherThrows, AutoPitchType, PitchCall, RelSpeed, VertRelAngle, HorzRelAngle, SpinRate, SpinAxis, Tilt, RelHeight, RelSide, Extension, InducedVertBreak, HorzBreak, VertApprAngle, HorzApprAngle, EffectiveVelo. Two additional variables were created, PitchGroup: which mapped each pitch type into either fastball, breaking ball, or off-speed. And SwingandMiss, which is a binary variable in which a swing and a miss is a 1 and any other pitch outcome is a 0. This became the dependent variable in the analysis. Any pitch with a null value in any column was dropped from the data frame.

Having cleaned the data frame and added necessary additional variables, the data frame was split into six separate data frames: lefty fastballs, righty fastballs, lefty breaking balls, righty breaking balls, lefty off-speed, and righty off-speed. The data frames were split by handedness of the pitcher because the trackman data uses a grid system for movement so a righty and a lefty will have different movement values based on which hand they throw with. Due to this, lefty and righty were split in order to make the model more accurate. The three different pitch groups were separated to make a model for each group of pitches, because a good fastball will look very different metrically from a good breaking ball or off-speed pitch.

Analysis:

Having split the data into the six separate data frames, Random Forest feature selection and correlation analysis was employed to determine which variables to include in the model. Due to the larger number of observations in the right-handed data frames, those were used for feature selection for the three pitch types.

For fastballs the correlation analysis to swing and miss was:

VertApprAngle 0.092114

RelSpeed 0.044613

VertRelAngle 0.041805

InducedVertBreak 0.041335

EffectiveVelo 0.039944

SpinRate 0.027500

HorzRelAngle 0.026145

HorzApprAngle 0.007928

RelSide 0.002029

RelHeight -0.000725

Extension -0.006998

HorzBreak -0.025131

SpinAxis -0.032955

The correlations in addition to the feature selection prompted the use of are Vertical Approach Angle(VAA), Induced Vertical Break (IVB), Release Speed (Velocity), Horizontal Approach Angle (HAA), Release Side, Horizontal Break, and Extension for use in the model.

For breaking balls the correlation analysis to swing and miss was:

RelSpeed 0.051972

EffectiveVelo 0.047510

SpinRate 0.039066

SpinAxis 0.021899

InducedVertBreak 0.017134

Extension 0.016368

RelHeight 0.009988

HorzBreak 0.008442

RelSide -0.007131

HorzApprAngle -0.053014

HorzRelAngle -0.064703

VertApprAngle -0.073315

VertRelAngle -0.125667

These correlations in addition to the feature selection prompted the use of VAA, HAA, Velocity, IVB, Spin Rate, and Extension for use in the model

For off-speed the correlation analysis to swing and miss was:

SpinAxis 0.054259

HorzBreak 0.043496

HorzApprAngle 0.042268

RelHeight 0.018039

HorzRelAngle 0.005685

Extension -0.001380

RelSide -0.012507

EffectiveVelo -0.043475

SpinRate -0.043482

RelSpeed -0.044589

InducedVertBreak -0.053571

VertRelAngle -0.057801

VertApprAngle -0.119447

These correlations in addition to the feature selection prompted the use of VAA, HAA, Velocity, IVB, Horizontal Break, and Extension for use in the model.

Next, the extraneous variables were dropped from the respective data frames, and swing and miss was set as the dependent variable for the model. To address the imbalance in the positive and negative class, SMOTE (Synthetic Minority Oversampling Technique) was used to enable the model to learn more effectively from both classes. From there, each of the six data frames were split into train and test data, with 30% of the data going into the test class. A Random Forest model was used in making the model and using the probability of a swing and miss that the model scored the pitch with, that probability was multiplied by 100 and used as the “Whiff score” which is the measurement of the probability that a given pitch will result in a strikeout. The data frames were then recombined into a singular data frame for use.

Limitations:

The first limitation in this analysis is the under sampling of the positive value of the dependent variable. The use of SMOTE, while widely used to address this problem, does have some limitations. The first of which is the risk of overlapping classes. SMOTE generates synthetic samples by interpolating between existing minority class samples. If the minority and majority classes are not well-separated in the feature space, this can lead to overlapping samples, which may confuse the model and reduce classification accuracy. In future analysis, it may be beneficial to use a dependent variable that has a more even distribution in the classes in order to avoid using under sampling procedures.

Another limitation that occurs in this analysis is in the data collection. Trackman tags the pitch based on what pitch type the trackman unit determines it to be. Due to this, some issues are typically present in the pitch type. In this case the model would classify the pitch into the wrong group, and that pitch would be incorporated into the incorrect model. This inconsistency will limit the model in its ability to score each pitch.

Insights and recommendations:

The insights that this analysis gives are mainly in the metrics that the model deemed important to creating a swing and miss. By knowing these metrics, pitchers can be trained to increase the performance of their pitches by throwing pitches with higher probabilities of a swing and miss. The main benefit of this analysis is the creation of the model that can grade a pitch based on the probability of a swing and miss. This model can be used in recruiting to determine a pitcher’s ability to perform at the division one level. One main problem that Kansas State has is the inability to predict performance from a lower level of baseball when they compete at the division one level. The introduction of this model will provide predictive power into how a pitcher could perform at this level.

Appendix:

Variable Explanations:

Pitcher: Name of the pitcher throwing the pitch

PitcherThrows: The handedness of the pitcher

AutoPitchType: The pitch type that the trackman tagged the pitch as

PitchCall: the result of the pitch

RelSpeed: Velocity of the pitch

VertRelAngle: the angle that the ball was released on the vertical axis

HorzRelAngle: the angle that the ball was released on the horizontal axis

SpinRate: How many RPMs the pitch had

SpinAxis: the axis that the ball was spinning on

Tilt: the tilt of the ball at release

RelHeight: the height from the ground the ball was released from

RelSide: how far to the side the ball was released from

Extension: the distance from the mound to the release point towards home plate

InducedVertBreak: the amount of vertical movement the ball had in a gravity free environment

HorzBreak: how much horizontal movement the ball had

VertApprAngle: the vertical angle at which the ball entered the strike zone

HorzApprAngle: the horizontal angle at which the ball entered the strikezone

EffectiveVelo: the perceived velocity of the pitch based on its location

A graph of a number of people

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

