Classification of Baseball Pitches through Supervised Learning

Seth Howells
seththowells@lewisu.edu
DATA-51000-002, FALL
Data Mining and Analytics
Lewis University

I. INTRODUCTION

Legendary manager, Connie Mack, once stated that “pitching is 75 percent of baseball [1].” America’s pastime has always been a sport of statistics from batting averages to closers’ performance in tight games. William Briggs, sport-enthusiast statistician, writes “for the first seven of the 21st century that found good pitching gives you about seven more victories over the course of a season, on average” in his article that discusses whether hitting or pitching wins more games [2]. While that discussion is up for debate, the key element of pitching – and moreover, the pitch type – is ultimately a determining factor.

Released in September 2019, the Major League Baseball (MLB) organization recorded MLB pitches from the 2015 season through to the 2018 season. The dataset consists of 65,535 instances (pitches) across 38 variables; however, the selected data is listed in SECTION II in more detail. The data is focused on the physics of a baseball in motion from the point of release to the point in which the baseball crosses Homeplate. The selected data is not concerned with the outcome of the baseball after it crosses Homeplate. That is, the umpire’s decision to call a pitch a Ball, Strike, or Foul is irrelevant to this analysis because the supervised learning models, further described in SECTION III, were applied to classify the type of pitch – not the outcome of the pitch.

The purpose of this analysis is to address whether the pitch type can be classified through supervised learning and how to identify the differences and similarities between each type of pitch.

The future sections of this report describe the dataset, the methodology, results along with a discussion, and a conclusion. Section II contains a description of the dataset used for this analysis. The methodology for analysis is presented in section III. In section IV, I report and discuss the results. Finally, section V provides conclusions.

II. DATA DESCRIPTION

The variables under investigations are listed in the table below. As mentioned in SECTION I, the dataset is comprised of 65,535 instances across 38 variables over a four-year period, which provides a sufficient amount of data for a classification analysis of multiple types of pitches. Each variable is directly associated with the physics of a ball in motion from a pitcher on the pitcher’s mound to Homeplate. Some of the attributes, like acceleration and velocity (denoted as data attributes: ax, ay, az, vx, vy, xz respectively), are represented in an *xyz* three-dimensional coordinate from the initial release of the ball.

TABLE I. SELECTED DATA ATTRIBUTES

Attribute	Type	Example Value	Description
pitch_type	Nominal (string)	CU	type of pitch (pitch code listed in TABLE II)
px	Numeric (real)	0.53	x-location as pitch crosses the plate. x=0 means right down the middle
pz	Numeric (real)	2.71	z-location as pitch crosses the plate. z=0 equates to ground level.
start_speed	Numeric (real)	91.80	speed of the pitch just as it's thrown
end_speed	Numeric (real)	84.60	speed of the pitch when it reaches the plate
break_angle	Numeric (real)	37.20	the angle, in degrees, of the vertical straight-line path from the release point to where the pitch crossed the

Attribute	Type	Example Value	Description
			front of home plate (as seen from the catcher's/umpire's perspective)
break_length	Numeric (real)	6.00	the measurement of the greatest distance, in inches, between the trajectory of the pitch at any point between the release point and the front of home plate, and the straight-line path from the release point and the front of home plate.
ax	Numeric (real)	-17.95	the acceleration of the pitch, in feet per second per second, in three dimensions, measured at the initial point.
ay	Numeric (real)	25.90	the acceleration of the pitch, in feet per second per second, in three dimensions, measured at the initial point.
az	Numeric (real)	-18.49	the acceleration of the pitch, in feet per second per second, in three dimensions, measured at the initial point.
vx0	Numeric (real)	11.63	the velocity of the pitch, in feet per second, in three dimensions, measured at the initial point.
vy0	Numeric (real)	-133.10	the velocity of the pitch, in feet per second, in three dimensions, measured at the initial point.
vz0	Numeric (real)	-4.94	the velocity of the pitch, in feet per second, in three dimensions, measured at the initial point.
x0	Numeric (real)	-2.59	the left/right distance, in feet, of the pitch, measured at the initial release point.
y0	Numeric (real)	50.00	initial release point of pitch.
z0	Numeric (real)	6.02	the height, in feet, of the pitch, measured at the initial release point.
pfx_x	Numeric (real)	-9.91	deviation (in inches) of pitch trajectory of vertical location.
pfx_z	Numeric (real)	6.72	deviation (in inches) of pitch trajectory of horizontal location.

Table II tabulates the pitch type distribution as well define the 7 different types of pitches being analyzed.

TABLE II. DISTRIBUTION OF TARGET VARIABLE: PITCH TYPE

Type	Distribution	Description
CH	14.42%	Changeup
CU	11.28%	Curveball
FC	8.52%	Cutter
FF	49.86%	Four-seam Fastball
FS	2.18%	Splitter
KC	3.37%	Knuckle-curve
SI	10.37%	Sinker

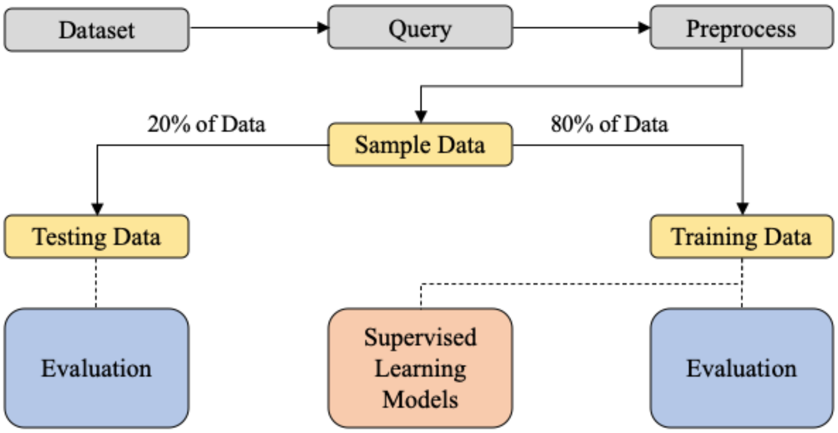
III. METHODOLOGY

The methodology is broken into three parts: data preprocessing and data sampling, supervised learning models, and model evaluation.

Data Preprocess and Sample

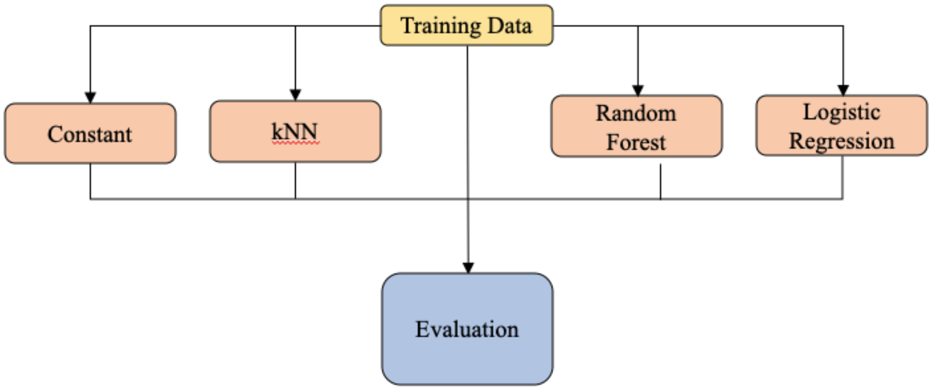
1. Query attributes that only pertain the pitch characteristics
    a. Selected Data in TABLE I
2. Query 7 most common types of pitches, all others were not introduced into the model
    a. Pitch Types in TABLE II
3. Impute missing values (3.2%) with average of vector
4. Normalize attributes: $\mu = 0$ and $\delta^2 = 1$

5. Sample 20% of queried data table from Step 3 for Testing Data.
6. Sample remaining 80% of queried data table from Step 3 for Training Data.



Supervised Learning Models

7. Constant: Chooses predicted class at random because there are two or more majority classes.
8. kNN: Set number of neighbors to 10
    a. Metric: Euclidean Distance
    b. Weight: Distance
9. Random Forest: Set number of trees to 100
    c. Limit the number of attributes at each split to 5
    d. Do not split subsets smaller than 20
10. Logistic Regression: Ridge L2 regularization type
    e. Strength set to C=10



Model Evaluation

11. All supervised learning models applied on the training set enter testing and scoring (evaluation)
12. Testing set, derived from Step 5, enters evaluation
13. Assess Confusion Matrix
    a. Area Under the Curve (AUC): a measure of separability in classification. Closer to 1.0 indicates a higher degree of separability and closer to 0 indicates a lower degree of separability [3].
        i. 0.9 to 1.0 = good
        ii. 0.8 to 0.9 = fair
        iii. 0.7 to 0.8 = poor [4]
    b. Classification Accuracy (CA): percentage of correctly predicted classes

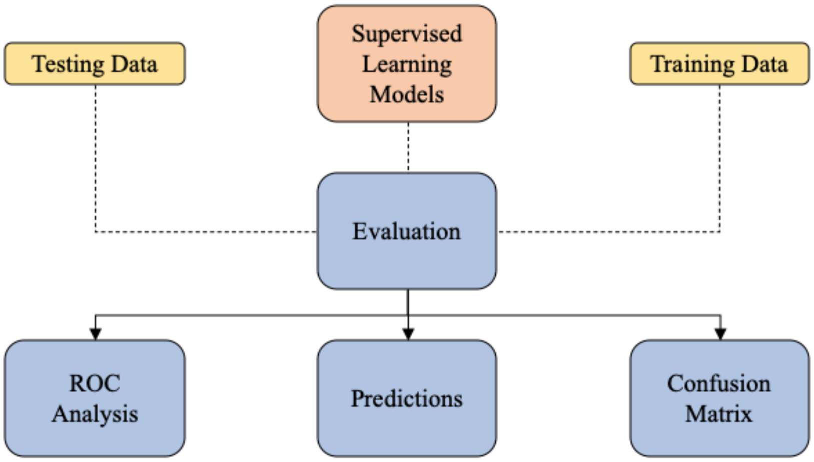
c. Precision: the number of true positives divided by the number of true positives plus false positives [5]

   i. Example: $\frac{pitches\ correctly\ identified}{pitches\ correctly\ identified\ +\ pitches\ incorrectly\ labled\ as\ a\ different\ pitch}$

d. Recall: the number of true positives divided by the number of true positives plus false negatives

   i. Example: $\frac{pitches\ correctly\ identified}{pitches\ correctly\ identified\ +\ pitches\ incorrectly\ labled\ as\ not\ the\ same\ pitch}$

14. Analyze ROC
15. View predictions of the testing dataset



IV. RESULTS AND DISCUSSION

The supervised learning models applied on the MLB 2015 season to 2018 season testing data, which consisted of 20% of pitches from the overall dataset as shown in Step 5 in SECTION III. The purpose of sampling the dataset into two data tables, 80% training data and 20% testing data, was for model evaluation through a series of testing and scoring of the models. The 20% sample for testing was initially withheld for cross validation [1]. In return, the training set with tuned parameters of the supervised learning models (see SECTION III) would be applied to unseen instances from the testing sample and scored on their accuracy. Scoring for classification models is determined by the metrics in SECTION III.

The evaluation results listed in Table III show how well the model was able to classify the pitch types. All models scored high in AUC which indicates that the models had a high degree of classifying the type of pitch thrown based on the characteristics of a baseball in motion. Classification accuracy (CA), on the other hand, showed greater variability with Nearest Neighbors (kNN) and Random Forest with above 90%. That is, +90% of the pitches thrown can be classified by the attributes listed in Table I.

TABLE III. EVALUATION RESULTS

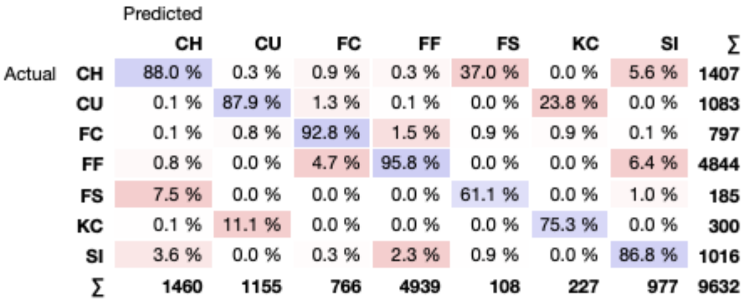
Model	AUC	CA	F1	Precision	Recall
kNN	0.989	0.917	0.914	0.913	0.917
Random Forest	0.992	0.911	0.906	0.906	0.911
Logistic Regression	0.983	0.863	0.851	0.851	0.863
Constant	0.500	0.503	0.337	0.253	0.503

Fig 1 is a confusion matrix that dives deeper into the Classification Accuracy (CA) with a table of correctly predicted types of pitches for Nearest Neighbors (kNN). For best interpretation, the matrix was converted to percentages rather than the number of instances. Likewise, Fig 2 shows the results of Random Forest.

Both Nearest Neighbors (kNN) and Random Forest experienced a high degree of misclassifications between Splitter and Changeup as well as Curveball and Knuckle-curve as indicated in Table IV. Attributes that did not affect the separability between the misclassified pitch and the actual pitch are visualized through box plots in Fig 3, Fig 4, and Fig 5. That is, all other attributes listed in Table II can be attributed to the misclassifications because of their similarities in distributions.

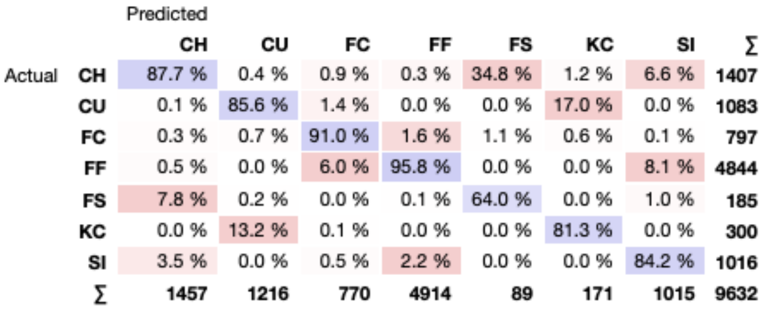
TABLE IV. MOST COMMON MISCLASSIFICATIONS

Predicted	Actual	Average Misclassification
Splitter (FS)	Changeup (CH)	35.9%
Curveball (CU)	Knuckle-curve (KC)	20.4%
Knuckle-curve (KC)	Curveball (CU)	12.2%
Changeup (CH)	Splitter (FS)	7.7%



	Predicted							
Actual	CH	CU	FC	FF	FS	KC	SI	∑
CH	88.0 %	0.3 %	0.9 %	0.3 %	37.0 %	0.0 %	5.6 %	1407
CU	0.1 %	87.9 %	1.3 %	0.1 %	0.0 %	23.8 %	0.0 %	1083
FC	0.1 %	0.8 %	92.8 %	1.5 %	0.9 %	0.9 %	0.1 %	797
FF	0.8 %	0.0 %	4.7 %	95.8 %	0.0 %	0.0 %	6.4 %	4844
FS	7.5 %	0.0 %	0.0 %	0.0 %	61.1 %	0.0 %	1.0 %	185
KC	0.1 %	11.1 %	0.0 %	0.0 %	0.0 %	75.3 %	0.0 %	300
SI	3.6 %	0.0 %	0.3 %	2.3 %	0.9 %	0.0 %	86.8 %	1016
∑	1460	1155	766	4939	108	227	977	9632

Fig. 1. Confusion Matrix for kNN: proportion of predicted pitch type vs actual pitch type



	Predicted							
Actual	CH	CU	FC	FF	FS	KC	SI	∑
CH	87.7 %	0.4 %	0.9 %	0.3 %	34.8 %	1.2 %	6.6 %	1407
CU	0.1 %	85.6 %	1.4 %	0.0 %	0.0 %	17.0 %	0.0 %	1083
FC	0.3 %	0.7 %	91.0 %	1.6 %	1.1 %	0.6 %	0.1 %	797
FF	0.5 %	0.0 %	6.0 %	95.8 %	0.0 %	0.0 %	8.1 %	4844
FS	7.8 %	0.2 %	0.0 %	0.1 %	64.0 %	0.0 %	1.0 %	185
KC	0.0 %	13.2 %	0.1 %	0.0 %	0.0 %	81.3 %	0.0 %	300
SI	3.5 %	0.0 %	0.5 %	2.2 %	0.0 %	0.0 %	84.2 %	1016
∑	1457	1216	770	4914	89	171	1015	9632

Fig. 2. Confusion Matrix for Random Forest: proportion of predicted pitch type vs actual pitch type

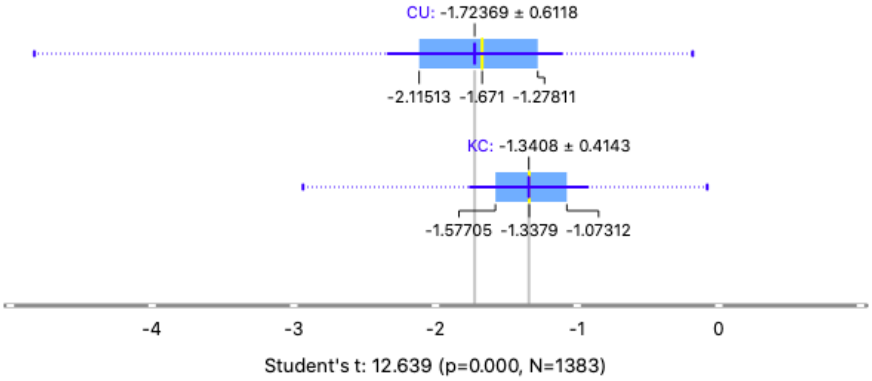


Fig. 3. Box Plot: Start Pitch Speed between Curveball and Knuckle-curve

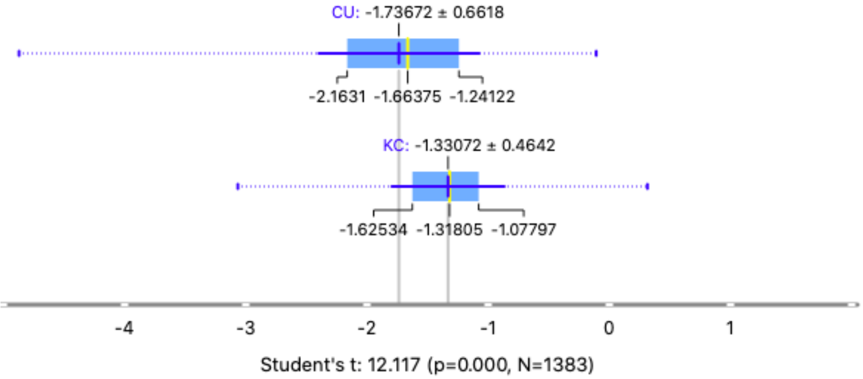


Fig. 4. Box Plot: End Pitch Speed between Curveball and Knuckle-curve

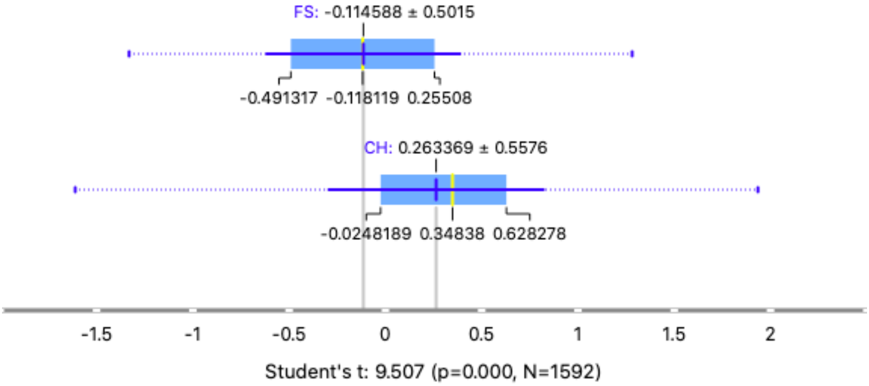


Fig. 5. Box Plot: Break Angle between Splitter and Changeup

V. CONCLUSIONS

Classifying pitch types, albeit impossible to the naked eye at real-time speed, based on a baseball in motion serves a new perspective in the 9.9 billion-dollar industry given that the sport is entirely dependent on the type of pitch [6]. Within 1.1 to 1.2 seconds, the ball is released from pitching mound and crosses Homeplate [7]. Despite small variability of a ball in motion within that time window, supervised learning models, like Nearest Neighbors (kNN) and Random Forest, can ultimately classify each pitch with around 90% accuracy with relatively low Type I and Type II errors.

REFERENCES

[1] “Is Pitching Really 75% of the Game: Pitching Facts & Opinions: Baseball,” Is Pitching Really 75% of the Game?, 03-Apr-2017. [Online]. Available: https://imaginesports.com/news/pitching-75-percent-game. [Accessed: 09-Oct-2020].

[2] Briggs, Hitting or Pitching. Which wins more games?, 28-Apr-2008. [Online]. Available: http://wmbriggs.com/post/120/. [Accessed: 09-Oct-2020].

[3] R. H. El Khouli, K. J. Macura, P. B. Barker, M. R. Habba, M. A. Jacobs, and D. A. Bluemke, “Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast,” Journal of magnetic resonance imaging : JMRI, 01-Nov-2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935260/. [Accessed: 09-Oct-2020].

[4] M. Kearns, “A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split.” [Online]. Available: http://papers.nips.cc/paper/1070-a-bound-on-the-error-of-cross-validation-using-the-approximation-and-estimation-rates-with-consequences-for-the-training-test-split.pdf.

[5] W. Koehrsen, “Beyond Accuracy: Precision and Recall,” Medium, 03-Mar-2018. [Online]. Available: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c. [Accessed: 09-Oct-2020]

[6] C. Gough, “MLB revenue by team/franchise 2019,” Statista, 29-Apr-2020. [Online]. Available: https://www.statista.com/statistics/193645/revenue-of-major-league-baseball-teams-in-2010/. [Accessed: 09-Oct-2020].

[7] J. J. Cooper, “Art Of The Steal,” College Baseball, MLB Draft, Prospects - Baseball America, 12-Aug-2012. [Online]. Available: https://www.baseballamerica.com/stories/art-of-the-steal/. [Accessed: 09-Oct-2020].