Predicting Baseball Pitcher Efficacy Using Physical Pitch Characteristics

Tejas Oberoi

Abstract

The efficacy of baseball pitchers can be predicted from prior pitching data using machine learning models. Previous machine learning works relating to baseball have primarily involved predicting outcomes of baseball games and a thrown pitch. This paper is the first work that uses sixteen game-independent features, which describe a pitcher's set of thrown pitches, to predict multiple pitcher efficacy metrics, like walks/hits allowed per inning (WHIP), batting average against (BAA), and fielding independent pitching (FIP). We hypothesized that these sixteen "physical features," which are measured by the use of sensors, can explain >50% of the variance while predicting pitcher efficacy. We applied the Neural Network model to predict the efficacy metrics using all sixteen features, while we used the Linear Regression model to analyze the individual impact of each feature for predicting the efficacy metrics. We observed from the Neural Network and Linear Regression models that the "ballFrequency" feature was the most impactful in predicting the WHIP for any pitcher. For the BAA and FIP metrics, the Linear Regression models showed that none of the features were impactful; however, we observed that the Neural Network model improved the prediction of the BAA and FIP metrics. Based on our evaluations, the machine learning models could not prove our hypothesis, as the results accounted for <50% of the variance when predicting the pitcher efficacy metrics. Professional scouts can still use the results of our individual feature analysis to select better pitchers who have never played a game at the professional level.

Introduction

In the sport of baseball, professional teams rely heavily on advanced statistics related to the performances of batters and pitchers in order to maximize their success. To aid these teams, scientists have created advanced statistics, such as Fielding Independent Pitching (FIP) and Walks and Hits per Innings Pitched (WHIP), to describe the impact of the performances of baseball players on a team's success with more detail than the original metrics like Earned Run Average (ERA) could describe. The creation and analysis of these advanced statistics in baseball are called sabermetrics (1).

Researchers have used machine learning (ML) models to predict different aspects of baseball using sabermetrics. Lee et al. and Hickey et al. applied ML models to predict a thrown pitch's outcome (2, 3). Connor Heaton and Prasenjit Mitra applied ML models to predict baseball game outcomes based on the performance of a specific player (4). On the other hand, Huang et al. applied three separate ML models to predict the outcome of a game (5). Watkins applied ML models to predict the efficacy of a major league baseball (MLB) batter for the next year based on his performance during the current season (6). Bock used sabermetrics and ML models to predict both the short-term and long-term efficacy of pitchers on their particular teams (7). All the above studies incorporated "non-physical features," like ball-strike or on-base percentages or batting averages, to predict only a single outcome metric. Unlike these previous studies, this study evaluates a single or a combination of “physical features” of a pitch, features that can be described qualitatively or can be measured using sensors as the input data to the ML models to predict multiple output metrics.

“Physical features” use advanced sensors or human eye ability to measure/describe the feature of a pitch thrown either in a game or non-game setting; however, a “non-physical feature,” such as the ball-strike count, must be measured in a game setting with or without the use of sensors. Therefore, using these ‘physical features”, scouts and recruiters can evaluate pitchers who have never played a game in the MLB. Accurately predicting sabermetrics like WHIP, BAA (Batting Average Against), and FIP would be crucial for determining a pitcher’s future success, as a lower value of these metrics would imply more efficient innings with fewer base runners and runs allowed (8). Even if a pitcher might seem enticing because of his high velocity and diverse set of pitches, he would be ineffective in games with high values of these metrics. A knowledge of these statistics for a pitcher prior to them being selected or pitching in their first game at the professional MLB would be pivotal for professional team scouts and managers. In addition, with the knowledge of which feature impacts a pitcher's efficacy, scouts could emphasize the important features while evaluating a pitcher.

In our study, we tested the hypothesis of whether physical pitch characteristics can predict greater than 50% of the variance, defined by the term ‘r^2,’ in the efficacy of a pitcher. To test this hypothesis, we applied Neural Network (NN) and Linear Regression (LR) models to predict three output metrics, WHIP, BAA, and FIP, using the sixteen “physical features.” When predicting the metrics, the models did not account for more than 50% of the variance (r^2). However, the NN models for the WHIP and FIP metrics still provided statistically significant results based on statistical analysis. Additionally, when we added a ‘non-physical feature’ like WHIP to the input space, the NN model accounted for more than 50% of the variance when predicting the BAA. Surprisingly, we observed a few characteristics, such as how hard a pitcher throws and the types of pitches, were insignificant in predicting the efficacy of a pitcher. Our findings are contrary to popular belief among baseball scouts and recruiters who place a significant emphasis on these two characteristics while determining the efficacy of a pitcher (9).

Methods

The following section describes the data collected and how the data was prepared for the models to use.

Dataset Preprocessing

The dataset used for this research was obtained from Statcast, which used an advanced camera-driven tracking system that was installed in every MLB stadium to extract advanced features for each pitch, such as its velocity, spin rate, exit velocity, pitch movement, pitch location, and more (10). For this research, we downloaded the Statcast dataset, which was comprised of the pitch-by-pitch data from 2017–2021, from the Kaggle website, and combined it into one Pandas data frame that yielded 3,149,505 rows of pitch data and 92 columns of pitch features (11). We then initially modified this dataset by deleting pitches that resulted in extremely rare outcomes (such as pickoffs) and only using data from 777 pitchers that had pitched at least 1000 pitches over the five seasons. As a result, we reduced the number of rows to 2,835,562 rows (90% of the original data). Additionally, we also reduced the number of columns from 92 to 83 by removing nine deprecated columns.

The modified dataset consisted of an “events” column, which described the outcome of an at-bat, and a “description” column, which described if the resulting pitch was a ball, strike, foul, or hit into play. We combined these two columns into a new and more detailed “description” column that contained the results of the at-bat from the “events” column and the results of the other pitches from the original “description” column. Additionally, some of the infrequently occurring variables in the new “description” column, which were a subset of a more commonly occurring result, were also combined to obtain the modified final dataset.

The modified dataset was used to extract the sixteen “physical” features, as defined below, that were used as input for our ML models (Table 1) from our current assortment of 83 columns. The “pitch_type” feature determined the type of thrown pitch and was classified using the one hot encoding method as “0” if it was not the pitch type thrown and “1” if it was the pitch type thrown. We created five categories for the “pitch type” feature–fastball, curveball, change-up, slider, and other pitches (12). The four pitch types selected were the most commonly thrown pitches, while the “other_offspeed” category comprised all other pitches (knuckleball, forkball, etc.) that were rarely thrown. The “zone” feature determined the location of the thrown pitch, whether it was thrown “high” (above batter’s waist) or “low” (below batter’s knee). The “release_ extension” feature measured the horizontal extension in feet of the pitcher’s arm before the ball was released (called the “release extension” of the pitch). The “release_spin_rate” and “release_speed” features described the spin rate and the velocity of the thrown pitch, respectively (13). We used the “p_throws” feature to classify right-handed pitchers as 0 and left-handed pitchers as 1. Unlike the above features based on a single-thrown pitch, the “ballFrequency” and “pitchTypeEntropy” features are based on all the pitches thrown by a specific pitcher from the dataset of 777 pitchers. The "ballFrequency" of each pitcher was calculated as follows:

$$ballFrequency = \frac{pitches\ thrown\ as\ balls}{total\ pitches\ thrown}$$

To compute the "pitchTypeEntropy", which described the distribution of the type of pitch thrown for each pitcher, the Shannon Information Entropy value E(x) was used on the five one-hot encoded “pitch type” features to estimate a value by applying the formula shown below:

$$E(x) \ = \ - \sum_{i=1}^{n} P(x_i) * log_2(P(x_i))$$

where, P(x_i) = probability of x_i occurring

A higher E(x) value implied that the pitcher threw a larger variety of pitches compared to a small E(x) value that implied the pitcher throws fewer types of pitches.

Creating the Input Features for each Pitcher Using the data given for each thrown Pitch

After grouping the rows by the pitchers, a Pandas GroupBy object was created from the original data frame with the one-hot encoded categorical columns containing 777 groups that each contain all the pitches thrown by a particular pitcher. For each pitcher group, we took the 5th and 95th percentiles of each quantitative feature in all the pitches. For each of the one-hot encoded qualitative features, relative frequencies were computed of their occurrences in the pitches for each pitcher group. As a result, the input features for the ML models (Table 1) were created.

"PHYSICAL" FEATURES	EXPLANATIONS
pitch_type_FF (PTFF)	Fastball (FF)
pitch_type_SL (PTSL)	Slider (SL)
pitch_type_CH (PTCH)	curve ball (CH)
pitch_type_CU (PTCU)	change-up (CU)
pitch_type_other_offspeed (PTOS)	offspeed pitches (OS)
zone_high_zone (HZ)	% of high zone pitches (above waist)
zone_low_zone (LZ)	% of low zone pitches (below knee)
release_extension_5th (RE5)	5th percentile of pitch release
release_extension_95th (RE95)	95th percentile of pitch release
release_spin_rate_5th (RSR5)	5th percentile of pitch spin rate
release_spin_rate_95th (RSR95)	95th percentile of pitch spin rate
release_speed_5th (RSPD5)	5th percentile of pitch speed
release_speed_95th (RSPD95)	95th percentile of pitch speed
p_throws (PT)	left or right-handed pitchers
ballFrequency (BF)	frequency of balls
pitchTypeEntropy (PTE)	distribution of pitches

TABLE 1. Input "physical" features. *Pitch_type – type of pitch thrown; zone – Location of pitch between knee and waist of the batter. release_extension – Horizontal extension of the pitcher's arm from the starting point. 5th and 95th percentile mean 5% had arm extended below and above the average extension, respectively. "release_spin_rate" – Spin rate of thrown pitch for a single pitcher. 5th and 95th percentile reflect 5% of thrown pitches had less spin and 5% had more spin than average spin rate, respectively. "release_speed" – Velocity of the thrown pitch for a single pitcher. 5th and 95th percentile reflect 5% of thrown pitches had less and 5% had higher velocity, respectively, than average velocity. "p_throws" – Classifies right-handed pitchers as 0 and left-handed pitchers as 1. "ballFrequency" – Pitches thrown as "balls" / total number of pitches thrown for a given pitcher. "pitchTypeEntropy" – Distribution of type of pitches (whether FF, SL, CH, CU, OS)*

Output Pitcher Efficacy Metric Explanations

For the initial simulations, we only evaluated the WHIP output metric, as defined below, because it is one of the most well-known sabermetrics and is easily computed (14).

$$WHIP = \frac{walks+hits}{IP}$$

Where, "walk + hits" = Total number of pitches in the dataset whose description column contained a walk or a hit, which is defined as a single, double, triple, or home run;

IP = Total innings pitched by each pitcher = $\frac{total\ number\ of\ outs}{3}$

The WHIP measured how "effective" a pitcher is by determining how many baserunners the pitcher allowed per inning and is unrelated to other confounding factors which are not in control of the pitcher, such as team defense. Therefore, a low WHIP implied that the pitcher was effective and did not allow many baserunners, which resulted in the pitcher pitching more innings and having lower chances of allowing many runs.

However, WHIP was not the only metric that measured a pitcher's efficacy, and when paired with other sabermetrics, it can provide more accurate predictions (15). Therefore, two other sabermetrics, Batting Average Against (BAA) and Fielding Independent Pitching (FIP), were also evaluated.

Unlike the WHIP, BAA measures pitcher efficacy based on each batter's ability, and takes into consideration only hits and no walks (16). BAA is defined as:

$$BAA = \frac{Hits}{At-bats}$$

where, At-bat = Any event excluding a walk, a hit by pitch (HBP), a sacrifice, and catcher's interference (17)

To compute the BAA for each pitcher, we used the same GroupBy object and data frame as the WHIP computations. However, instead of summing up the walks and hits, we only summed up the hits and ignored any value in the "description" column that was a walk, an HBP, a sacrifice event, or a catcher's interference.

For evaluation of the FIP metric, only home runs (HR), strikeouts (K), hit by pitch (HBP), and walks (BB) were considered as these were completely fielder-independent metrics and were controlled by only the pitcher.

$$FIP = \frac{13HR-3(BB+HBP)-2K}{IP} + C$$

where, IP = Innings pitched

C = Constant that is assigned a value of 3.18, which is the average of all C values from 2017 – 2021

Overall, all three of these metrics interpreted the same outcome, a pitcher's efficacy, while minimizing the dependence on confounding factors (such as fielders, errors, game situations, and more). The main difference between these three metrics is the events that they considered.

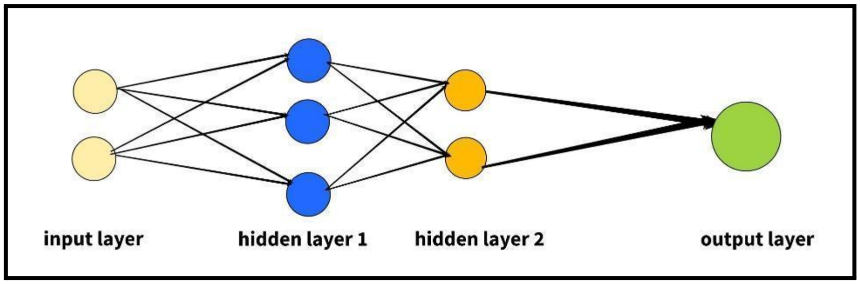
Training/Testing/Validation Data

The dataset of 777 pitchers was split randomly to produce a training dataset of 582 pitchers and a validation dataset of 195 pitchers. The NN and LR models were trained on the training dataset of 582 pitchers, and then all the accuracy and significance metrics were computed on the validation dataset of 195 pitchers

Neural Network Model

A two-hidden layer NN model (Figure 1) was used to evaluate the dataset by varying both of the hidden layer widths from 4-14 and trying every possible combination of the layers using a linear search with increments of 2 (4,6,8,10,12,14). In addition, we used the leaky_relu activation function on each hidden layer. The initial results indicated that the accuracies were similar regardless of the width of the hidden layers. So for the main model, we decided to set the sizes of the two hidden layers to 8 and 6, respectively.

Figure 1. Typical 2-hidden layer neural network model



Linear Regression Model

LR models were applied to determine the correlation between every “physical” feature and the three-output metrics. We compared LR and the NN model results because we suspected that the NN model only used a few of the 16 features to make its predictions.

Neural Network Model Training and Validation Process

Before training the model, the training and validation input datasets were normalized by using the z-score normalization formula:

$$x_{new} = \frac{(x_i - mean)}{SD}$$

where, x_{new} = transformed dataset value
x_i = initial dataset value

A training loop was run on the 582-pitcher training dataset using 50 epochs on a CPU, batch sizes of 4, and an Adam optimizer with a learning rate of 0.001 and a weight decay of 0.01. In addition, a validation dataset of 195 pitchers was evaluated. After the model ran 50 loops, the same validation data was used to compute the accuracy and root mean squared error (RMSE).

Linear Regression Model Training and Validation Process

For the LR model, we used the testing data as the prediction term to calculate the RMSE. In addition, we made 15 scatterplots of the physical features in the testing data vs. the actual pitcher metrics to visualize any potential correlations.

Results

We performed these experiments to determine if the ML models could accurately predict the metrics enough to satisfy the hypothesis. For all three-output metrics, we defined the NN accuracy as the % of predicted values within 10% of the respective output metric data range away from the actual value. In addition, we used the RMSE to evaluate the quality of our predictions, the correlation coefficient (r) to describe the LR model scatter plots, and the measure of variance (r^2) to predict the variance of the observed values from the predicted values for the ML models and to evaluate our hypothesis. We calculated the measure of variance (r^2) using the formula:

$$RMSE = \sqrt{\left(1 - r^2\right)} * SD$$

where, RMSE = Root Mean Squared Error of the model
SD = Standard Deviation of the output metric

Due to the NN model producing slightly different results for every model iteration using identical parameters, we trained the NN model 5 times, and reported the accuracies as ranges, and the RMSE as the average of all the five iterations. We first ran a baseline prediction for each experiment to obtain the baseline values for the output metric's RMSE and accuracy value. Following the baseline prediction simulation, we ran the ML models to obtain the efficacy metrics. For each model-produced RMSE value, we then ran an F-Test to determine if the variance of the model is significantly different from the baseline results, implying that the model fits the output data better than the baseline model.

Predicting WHIP Metric

For the baseline prediction, we observed that using the mean WHIP value in the validation dataset as the constant value yielded 48.2% accuracy and an RMSE of 0.158. Following the baseline prediction, we ran the best NN model, which yielded a validation accuracy of 53-57%, an RMSE of 0.134, and an r^2 value of 0.28. Finally, we ran an F-test on the RMSE of the NN and the SD of the WHIP values, which yielded a p-value of 0.0190, which implied that the NN produced statistically significant results for predicting WHIP.

In addition, we trained and tested LR models to determine the RMSE for the relationship between each feature and the WHIP. Then, we ran F-Tests comparing the RMSEs of each input feature with the variance of the WHIP to obtain the p-values (Table 2). All features, except for the "ballFrequency" feature, produced statistically insignificant results ($p > 0.05$, F-test, Table 2). We observed the "ballFrequency" feature to be statistically significant with a p-value of 0.024. For the trained "ballFrequency" LR model, the testing accuracy was 55.8%, the RMSE was 0.137, the correlation coefficient (r) was 0.498, and the variance (r^2) was 0.248, which implied a low correlation between the frequency of balls thrown and the WHIP value (Figure 2a). The graphs did not show any correlation due to the randomly scattered data points for the non-statistically significant LR models we evaluated using the other 15 physical features (Figure 2b).

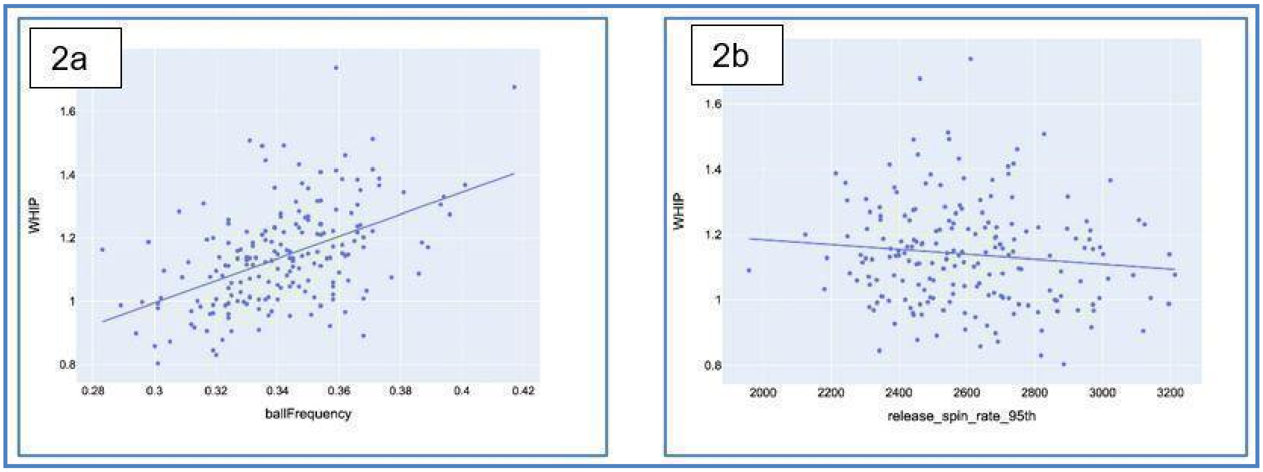


Figure 2. Input "physical" features vs. WHIP. *Figure 2a shows the scatter plot of the validation dataset when the Linear Regression (LR) model was run for the "ballFrequency" feature and the Walks and Hits Per Inning (WHIP) metric. The F-Test resulted in a p-value of 0.024. The correlation coefficient (r) was 0.498, and the variance (r^2) was 0.248. Figure 2b shows the scatter plot for one of the non-correlated and non-statistically significant features (release_speed_95th, $r < 0.1$, $r^2 < 0.01$) and the WHIP metric when the LR model was run. The F-Test resulted in a p-value of 0.465.*

Predicting BAA Metric

For the baseline prediction, we observed that using the mean BAA value in the validation dataset as the constant value yielded 52.8% accuracy and an RMSE of 0.0245. Following the baseline prediction, we ran the best NN model, which yielded a validation accuracy of 55-61%, an RMSE of 0.0223, and an r^2 value of 0.172. Finally, we ran an F-test on the RMSE of the NN and the SD of the BAA values that yielded a p-value of 0.0815.

In addition, we trained and tested LR models to determine the RMSE for the relationship between each individual feature and the BAA. Then, we ran F-Tests comparing the RMSEs of each input feature with the variance of the BAA to obtain the p-values (Table 2). We observed that no features produced statistically significant results ($p > 0.05$, F-test, Table 2). However, we still analyzed the LR model results for the feature with the lowest p-value, "release_speed_95th". We obtained a testing accuracy of 55.8% and obtained an RMSE, SD, 'r', and r^2 of 0.0233, 0.0245, -0.310, and 0.0961, respectively (Figure 3a). The low correlation coefficient implies that no strong correlation existed between the 95th percentile of a pitcher's release speed and the BAA value. Since no feature produced statistically significant results, we could not conclude which LR models fit the data better than the baseline model for the BAA metric.

F-TEST RESULTS (FEATURES vs OUTPUT METRICS)			
Input Feature	**P-Value**		
	WHIP	**BAA**	**FIP**
p_throws	0.445	0.462	0.502
release_speed 5th	0.496	0.404	0.344
release_speed_95th	0.486	0.226	0.152
release_spin_rate_5th	0.453	0.364	0.362
release_spin_rate_95th	0.465	0.343	0.379
release_extension_5th	0.428	0.454	0.328
release_extension_95th	0.502	0.521	0.476
pitch_type_CH	0.490	0.413	0.436
pitch_type_CU	0.485	0.509	0.478
pitch_type_FF	0.476	0.474	0.486
pitch_type_SL	0.487	0.427	0.542
pitch_type_other_offspeed	0.451	0.470	0.410
zone_high_zone	0.517	0.532	0.526
zone_low_zone	0.517	0.532	0.526
ballFrequency	0.024	0.437	0.464
pitchTypeEntropy	0.495	0.437	0.274

***TABLE 2. F-Test results for WHIP, BAA, and FIP output metrics.** Table 2 shows the results of the performed F-Tests. The F-Tests use the variance of the input features and the output metrics. Walks and Hits Per Inning (WHIP); Batting Average Against (BAA); Fielding Independent Pitching (FIP).*
P-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant. Shaded cell shows significant p-value.

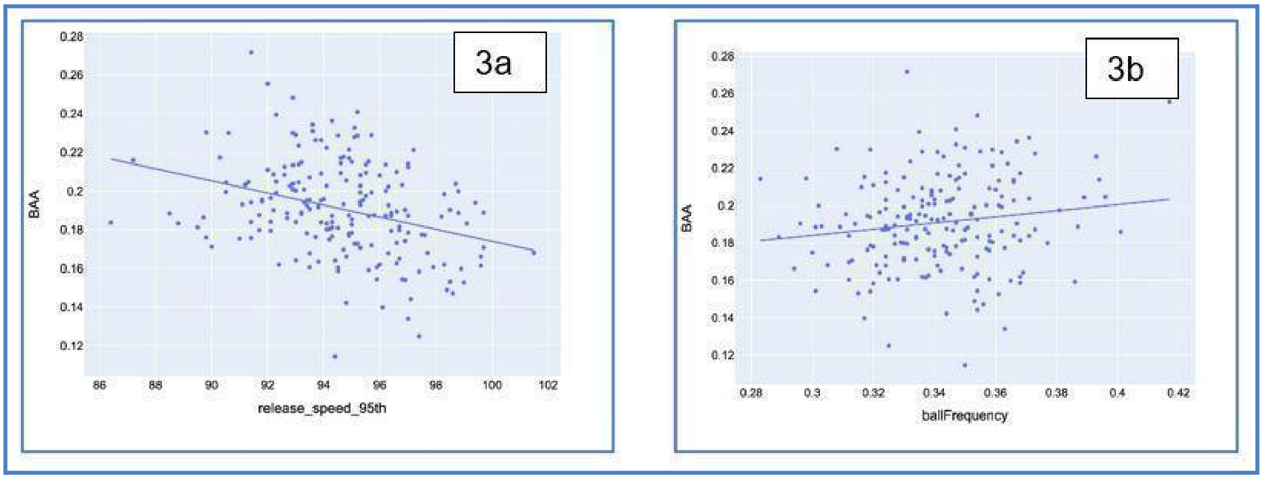


Figure 3. Input "physical" features vs. BAA. *Figure 3a shows the scatter plot of the validation dataset when the Linear Regression (LR) model was run for the "release_speed_95th" feature and the Batting Average Against (BAA) metric. The F-Test resulted in a P-value of 0.226. The correlation coefficient (r) was -0.310, and the variance (r^2) was 0.0961. Figure 3b shows the scatter plot for another non-correlated and non-statistically significant feature (ballFrequency, r = 0.150, r^2 = 0.025) and the BAA metric when the LR model was run. The F-Test resulted in a p-value of 0.437.*

Predicting FIP Metric

For the baseline prediction, we observed that using the mean FIP value in the validation dataset as the constant value yielded 53.3% accuracy and an RMSE of 0.856. Following the baseline prediction, we ran the best NN model, which yielded a validation accuracy of 55-60%, an RMSE of 0.720, and an r^2 value of 0.293. Finally, we ran an F-test on the RMSE of the NN and the SD of the FIP values that yielded a p-value of 0.0057, which implied that the NN model produced statistically significant results for predicting FIP.

In addition, we trained and tested LR models to determine the RMSE for the relationship between each feature and the FIP. Then, we ran F-Tests comparing the RMSEs of each input feature with the variance of the FIP to obtain the p-values (Table 2). We observed that no features produced a statistically significant correlation ($p > 0.05$, F-test, Table 2). However, we analyzed the LR model results for the feature with the lowest p-value, "release_speed_95th". We obtained a testing accuracy of 53.3% and an RMSE, SD, 'r', and r^2 of 0.797, 0.856, -0.365, and 0.133, respectively (Figure 4a). The low correlation coefficient implies that no strong correlation existed between the 95th percentile of a pitcher's release speed and the FIP value. Since no feature produced statistically significant results, we could not conclude which LR models fit the data better than the baseline model for the FIP metric (Figure 4b).

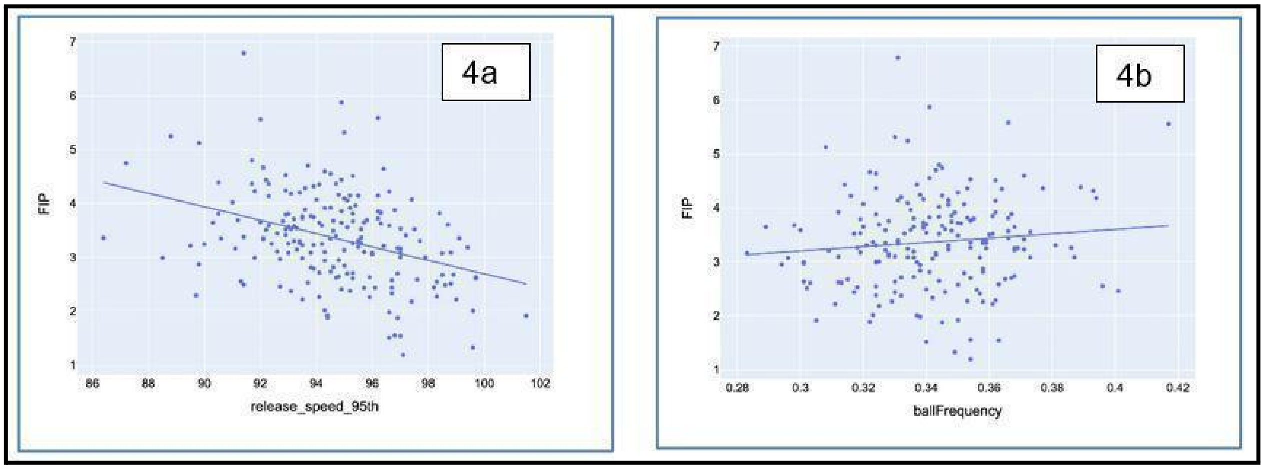


***Figure 4. Input "physical" features vs. FIP.** Figure 4a shows the scatter plot of the validation dataset when the Linear Regression (LR) model was run for the "release_speed_95th" feature and the Fielding Independent Pitching (FIP) metric. The F-Test resulted in a P-value of 0.152. The correlation coefficient (r) was -0.365, and the variance (r^2) was 0.133. Figure 4b shows the scatter plot for another non-correlated and non-statistically significant feature (ballFrequency, r = 0.103, r = 0.0106) and the FIP metric when the LR model was run. The F-Test resulted in a P-value of 0.464.*

Using WHIP/BAA/FIP as input features

To determine if the output metrics correlated with each other, we performed F-tests for the regression analysis between the RMSEs of each output feature. Since WHIP and BAA produced significant results, we added WHIP to the input space to analyze how much the NN model improved in predicting BAA and also evaluate if a non-physical feature, like WHIP, helped the NN model to account for more than 50% of the variance in its predictions (p = 0.0003, F-test, Table 3). With these added features, we ran the NN model on 100 epochs instead of 50 because it took longer to train the dataset. In addition, we tested the NN with added dropout layers of p=0.3 between the hidden layers of the model.

OUTPUT METRICS	**P-VALUE**
WHIP vs BAA	0.0003
FIP vs WHIP	0.230
FIP vs BAA	0.0005

***TABLE 3. F-Test results for all three metrics.** Table 3 shows the results of the performed F-Tests, where the variance of different output metrics is used to evaluate their relationship with each other. Walks and Hits Per Inning (WHIP); Batting Average Against (BAA); Fielding Independent Pitching (FIP). P-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. A p-value of 0.05 or lower is generally considered statistically significant. Shaded cells show significant p-values.*

We observed that for predicting BAA using the existing physical features and adding WHIP to the input space, the accuracy increased to 72%-77% with the dropout layers (69%-74% without), while the RMSE decreased to 0.0160. In addition, we performed an LR analysis of WHIP vs BAA, and we observed that the accuracy decreased to 67% with an RMSE of 0.0190. With the WHIP metric as an added input feature, we computed the p-value of the NN model results to be extremely low (approximately 0), which implied that the NN model significantly improved the accuracy for predicting BAA when we added the WHIP metric (Figure 5).

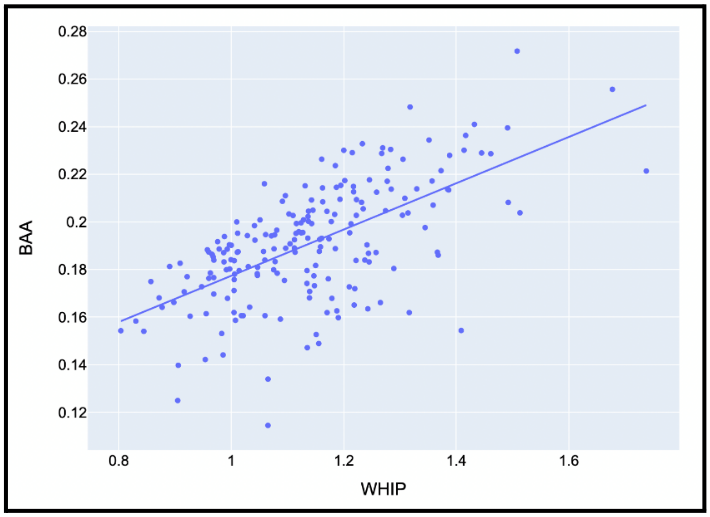


Figure 5. WHIP vs. BAA. *Figure 5 shows the scatter plot for predicting Batting Average Against (BAA) using Walks and Hits Per Inning (WHIP). Using the Linear Regression (LR) model with the WHIP and the BAA metrics, the correlation coefficient (r) was 0.610, and the variance (r^2) was 0.372. The F-Test resulted in a p-value of 0.00003.*

Conclusion and Discussion

For our research, we used the RMSE metric because of its usefulness in significance tests to make statistically backed solutions. When we analyzed the WHIP output metric, we observed that the NN model RMSE and the LR model RMSE with the “ballFrequency” feature yielded similar p-values of 0.024 and 0.019, respectively. This implied that the NN model did not provide any improvements in predictions compared to simply using an LR model with only the ”ballFrequency” feature.

Based on the similar RMSEs and p-values from the F-test for the NN and LR models, the NN model did not use the other 15 features to its advantage and most likely implemented a linear regression-like function using the “ballFrequency” feature instead of a more sophisticated function using multiple features. Furthermore, the low correlation values and high p-values for the other 15 features implied that they do not add any meaningful value to predicting WHIP.

We observed from the ML model runs that the other 15 “physical” features did not add meaningful value to predicting WHIP. This directly contrasts with the popular belief among professional scouts and coaches who consider that “physical” features such as “release_speed” (velocity of a pitch thrown), “pitchTypeEntropy” (mixing up pitch types frequently), and release_spin_rate (the amount a pitch spins when thrown) are better indicators of a pitcher’s efficacy than other “physical” features.

However, unlike the WHIP metric, we observed that the BAA and FIP metrics had much lower RMSEs for the NN model than for any of the LR models. In addition, the FIP NN model predictions displayed a statistically significant result for the F-Test that we ran between the NN model’s RMSE and the SD of the FIP values ($p = 0.005$, F-test, Table 2). Since the NN model that predicted the FIP did not contain any statistically significant input features, the NN model probably created a function that used a combination of the “physical” features to significantly improve its performance for the FIP predictions ($p > 0.05$, F-test, Table 2).
Based on all the experiments conducted, we concluded that the current “physical” feature pool and the dataset are insufficient in explaining more than 50% of the variance of all three pitcher efficacy metrics ($r^2 < 0.5$ for all models). However, when we extended the feature pool to include a non-physical feature, WHIP, in the input space, we observed that the NN model's predictions accounted for more than 50% of the BAA's variance. The WHIP's influence on the model did not surprise us because it is more correlated with the BAA than any of the other input features and is essentially measuring a similar property, the pitcher's efficacy (Figure 5).

In addition, we concluded that the ”ballFrequency” physical feature is the most important “physical” feature in determining WHIP because it had the lowest p-value in the F-Test and had the highest linear correlation. Even though ”release_speed_95th” had the lowest p-values in both F-Tests with the BAA and FIP output metrics, we observed that the p-values were not low enough for them to be considered statistically significant. This result implied that throwing a high percentage of strikes is important to determine a pitcher’s efficacy and that the current trend among scouts and coaches of using pitch velocity and a wide variety of pitches, though crucial attributes, to determine a pitcher’s efficacy may not be as important.

However, there are some limitations in the experiments conducted. For instance, a lack of pitcher data (only 777 pitchers) could have contributed to the NN model being unable to find a pattern more sophisticated than one exhibited by an LR model.

The intentional feature selection of only including game-independent “physical” features for this study resulted in the models only knowing about the properties of the thrown ball, and nothing about the “non-physical” features like a batter’s batting average, game score, ball/strike count, fielder positions, and more. In prior sabermetric studies, many experiments involved knowing some of these “non-physical” features, especially the opposing batter’s statistics against a particular pitcher, which provided useful information that determined the likelihood of the pitcher allowing a hit or getting an out. Overall, the research dataset was very noisy because of the inherent variance of successes of MLB pitchers, since at the professional level, most batters can hit most pitchers, regardless of their unique physical attributes. With more pitchers in a dataset and more useful physical features, an NN model can potentially predict a pitcher’s efficacy without seeing the pitcher perform in an MLB game setting because the data might become less noisy. Due to the existing statistically significant results with the WHIP and FIP metrics with the current dataset and feature pool, more data and features can potentially result in significant results for BAA and improve the RMSE of the FIP and WHIP NN models as well.

Instead of attempting to predict a pitcher efficacy metric by using a training dataset of professional baseball pitchers, one could analyze each pitcher individually. For instance, one could break up each individual pitcher by each pitch that they throw. With each thrown pitch, one can use 1D LR analysis and hypothesis testing for each physical feature as the input and the pitch result (i.e., hit or no hit) as the output to determine if these features convey useful information at the pitch-by-pitch level for that individual pitcher.

In addition, one could explore other methods in ML with more pitcher data (i.e. 20 years instead of 5 years) to predict the pitcher efficacy metrics more accurately. This would provide a breakthrough in the professional field of baseball, particularly recruiting, as scouts would be able to solely use “physical” features to predict a pitcher’s success prior to being drafted in MLB.

One could also extend this pitcher efficacy evaluation using ML models to the high school and collegiate levels. In such a setting, the ML NN model could potentially yield more accurate results for predicting the pitcher efficacy metrics.

Acknowledgments

I would like to thank Dr. Sam Saarinen for mentoring me in this project, as he was extremely helpful and supportive throughout my journey.

References

1. Beneventano, et al. "Predicting run production and run prevention in baseball: the impact of Sabermetrics." *Int J Bus Humanit Technol* 2.4 (2012): 67-75.

2. Lee, Jae Sik. "Prediction of pitch type and location in baseball using ensemble model of deep neural networks." *Journal of Sports Analytics* Preprint (2022): 1-12.

3. Hickey, Kevin, et al. "Dissecting moneyball: Improving classification model interpretability in baseball pitch prediction." (2020).

4. Heaton, Connor, and Prasenjit Mitra. "Learning to Describe Player Form in the MLB." *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer, Cham, 2022.

5. Huang, Mei-Ling and Yun-Zhi Li. "Use of machine learning and deep learning to predict the outcomes of major league baseball matches." *Applied Sciences* 11.10 (2021): 4499.

6. Watkins, Christopher. *Novel statistical and machine learning methods for the forecasting and analysis of Major League Baseball player performance*. Diss. Chapman University, 2020.

7. Bock, Joel R. "Pitch sequence complexity and long-term pitcher performance." *Sports* 3.1 (2015): 40-55.

8. Rogers, Michael. “What Is WHIP in Baseball? (Fully Explained).” *Nations-Baseball*, Nations-Baseball, 29 Aug. 2022, www.nations-baseball.com/whip-in-baseball/.

9. Destefano, Christine. *What Scouts Look For*. baseballakademie.de/wp-content/uploads/2016/02/whats_scouts_look_for.pdf.

10. “Statcast | Glossary | MLB.com.” *MLB.com*, 2022, www.mlb.com/glossary/statcast.

11. “MLB Statcast Data.” *Kaggle.com*, 2015, www.kaggle.com/datasets/s903124/mlb-statcast-data.

12. “What Are the Pitch Types Generated from MLB Statcast for Speed of Pitch?” *Daktronics.com*, 2022, www.daktronics.com/en-us/support/kb/DD3312647.

13. “Statcast Search CSV Documentation.” *Baseballsavant.com*, 2022, baseballsavant.mlb.com/csv-docs.

14. Slowinski, Piper. “WHIP.” *Sabermetrics Library*, 2022, library.fangraphs.com/pitching/whip/.

15. Slowinski, Piper. “FIP.” *Sabermetrics Library*, 2022, library.fangraphs.com/pitching/fip/.

16. “Batting Average | Glossary | MLB.com.” *MLB.com*, 2022, www.mlb.com/glossary/standard-stats/batting-average.

17. Nelson, Steve. “What Is an At-Bat in Baseball?” *Baseball Training World*, Baseball Training World, 8 Sept. 2021, baseballtrainingworld.com/what-is-an-at-bat-in-baseball/.

Appendix

Code used to run our experiments - https://github.com/toberoi05/BaseballResearch