Using Naive Bayes to Predict the Outcome of a Plate Appearance

Margot Douillet

December 4, 2020

Abstract

There is an increasing effort in professional baseball to predict the outcome of a particular plate appearance (that is a batter-pitcher match up). This paper aims to take a naive Bayes approach to estimating the probability that a hitter will get on base in a given plate appearance. First, Bayes' theorem is used to estimate the posterior and posterior predictive distributions of the probability for a specific hitter getting on base against a certain pitcher. Then a naive Bayes classification method will be used on three separate models to attempt to classify the outcome of a plate appearance against a specific pitcher, given the pitches that the pitcher has thrown. The first naive Bayes classification model will be univariate, simply denoting whether or not the pitcher has thrown a certain pitch. The mutlivariate model will denote how many of each pitch the pitcher threw. The third model will be a variation of the multivariate model, where the entries represent proportions instead of counts (that is, the proportion of the pitches thrown that were of type j). The data will be randomly split into training and testing data sets and the classification error for both models will be analyzed.

Introduction

If managers in professional baseball could better predict the outcome of a plate appearance, this could drastically change the way in which the game is played. A manager could compare the probabilities of several different hitters getting on base against a given pitcher, and choose which hitter to put in the game accordingly. Similarly, if a manager knew the probability that any hitter reached base against their pitcher given that the pitcher had thrown certain pitches in the plate appearance, then the manager could tell which pitching sequences were the most unsuccessful. This information would greatly influence a pitcher's repertoire and pitch usage.

This analysis could just as easily be done for a hitter. If a team knew the probability of one of their hitters getting on base given the pitches he saw, then the team could determine which pitches the hitter struggled against. Teams could then review the repertoire of a pitcher they were about to face, and choose a lineup based on the hitters that performed best against this repertoire.

Related Work

There have been few attempts to predict the outcome of a plate appearance - with most attempts predicting multiple outcomes instead of just two. Davis Weir (2020) used several different logistic

regression models to try to predict outcomes such as a single, strikeout, or walk. For predictors, he used projected rates of the predicted outcome. For example, one of Weir's models attempted to model the probability that a batter would draw a walk (the outcome being binary, where 0 = the batter does not draw a walk and 1 = the batter does draw a walk). As predictors for this specific model, Weir used the projected probability that a hitter would get a walk, the projected probability a pitcher would give up a walk, and the projected league average probability of a walk occurring.

Weir's logistic regression model had varying rates of success. It predicted a strikeout with an error rate of 0.329, a walk with an error rate of 0.157, a home run with an error rate of 0.066, a single with an error rate of 0.261, a double with an error rate of 0.099, a triple with an error rate of 0.009, and a hit by pitch with an error rate of 0.021.[1]

Doo, et. al. (2018) used a Bayesian hierarchical log5 model to predict the outcome of an at-bat. As predictors, they used the probability of the hitter getting on base (On Base Percentage) and the likelihood of a pitcher letting a batter reach base (Opponent On Base Percentage). The model was created for several specific batter-pitcher match-ups. They achieved a test MSE of 0.0514. Improving upon this model, Doo et. al. changed the predictors for their log5 model, instead using a batter's number of hits, home runs, plate appearances, walks, strikeouts, and hit-by-pitches. The second model yielded a test MSE of only 0.0428.[2]

Joshua Silver (2020) used neural networks to try to predict the outcome of plate appearances. Neural networks are a type of complex artificial intelligence algorithm capable of finding patterns. The specific predictors used by Silver are unspecified, but his model was trained on data from more than 1.5 million plate appearances in Major League Baseball between 2009 and 2019. The two outcomes Silver was most interested in predicting were strikeouts, for which he yielded an error rate of about 0.21, and home runs, for which he yielded an error rate of about 0.46.[3] As compared to Weir's log5 model, the neural network predictions for strikeouts were more accurate, while the neural network predictions for home runs were less accurate.

Another approach to predict the outcome of plate appearances was Nuclear Norm Penalty (an alternative to the Ridge Penalty) for multinomial regression, as used by Powers et. al. (2019). Powers et. al. attempted to predict nine different outcomes: a strikeout, walk, hit-by-pitch, ground out, fly out, single, double, triple, and home run. Similar to Doo et. al., they tailored the model to specific batter-hitter matchup, using the tendency of the batter to produce the outcome of interest, the tendency of the pitcher to give up the outcome of interest, the tendency of the ballpark to produce the outcome of interest, the league-wide frequency of the outcome, and the increased likelihood of the outcome, given the batter's handedness (right vs. left), as predictors. Powers et. al. yielded a test mean loss of approximately 1.25.[4]

Overall, log5 appears to be the preferred model for predicting the outcome of a plate appearance, both because it is not very computationally expensive, and because it yields fairly accurate predictions. While neural networks appears to be a fairly accurate method of predicting the outcome of a plate appearance, both the collection of the data and the training of the model are enormously computationally expensive. This paper simply aims to examine the effectiveness of naive Bayes classification in predicting a binary outcome of a plate appearance (where 0 indicates the hitter does not get on base, and 1 indicates the hitter does get on base), as compared to other methods previously used. Furthermore, unlike similar research, this study aims to use the pitches seen in the plate appearance as predictors to better understand the effectiveness of different pitches that the pitcher of interest throws.

Data

The pitcher of interest, Yu Darvish throws nine different pitches: a four-seam fastball, two-seam fastball, cut fastball (cutter), split-finger fastball (splitter), sinker, slider, changeup, curveball, and slow curve. For the purposes of this analysis, the slow curve was defined as any curveball whose velocity was 74 mph or less. There are patterns in pitch usage from one season to the next, and one game to the next. For this reason, the game and inning from which the data was drawn were randomly selected. Game logs for Darvish's starts are available from 2012-2020, but the years 2012 and 2013 were not sampled from, as pitch usage from these years was unreliably available.

For the univariate model, the data from 244 plate appearances was collected and randomly split into 194 training plate appearances, and 50 test plate appearances. For the multivariate model, the data from 290 plate appearances was recorded, with 50 randomly selected plate appearances used as the test data, and the remaining as the Training data. The data was sampled in the same method as for the univariate model. The proportional model used the same raw data as the multivariate model, but where each entry represents a proportion $\frac{x_{ij}}{\sum_{i=1}^{j} x_{ij}}$. That is, instead of k being the count of pitches j that a hitter saw in a plate appearance, k would be the proportion of pitches in the plate appearance that are of type j.The first nine plate appearances of the univariate data, multivariate data, and proportional data are shown in figures 1, 2, and 3 below, respectively.

Table 1: Univariate Data

Pitch Type	**PA 1**	**PA 2**	**PA 3**	**PA 4**	**PA 5**	**PA 6**	**PA 7**	**PA 8**
4-Seam Fastball	1	0	1	1	1	1	1	1
2-Seam Fastball	0	0	0	0	0	0	0	0
Sinker	0	0	0	0	0	0	0	0
Cutter	1	1	1	0	0	0	0	0
Curveball	0	0	0	0	0	0	0	0
Slider	0	1	1	0	0	0	0	0
Slow Curve	0	0	0	0	0	0	1	1
Changeup	0	0	0	0	0	1	1	0
Splitter	1	0	0	0	0	0	0	0

Table 2: Multivariate Data

Pitch Type	**PA 1**	**PA 2**	**PA 3**	**PA 4**	**PA 5**	**PA 6**	**PA 7**	**PA 8**
4-Seam Fastball	1	0	2	1	0	0	0	1
2-Seam Fastball	0	0	0	1	1	1	0	1
Sinker	0	0	0	0	0	0	0	0
Cutter	1	2	2	1	2	4	1	2
Curveball	0	0	0	0	0	0	0	1
Slider	0	1	1	0	0	0	0	0
Slow Curve	0	0	0	0	0	0	0	0
Changeup	0	0	0	0	0	0	0	0
Splitter	0	0	0	0	0	0	0	0

Table 3: Proportional Data

Pitch Type	PA 1	PA 2	PA 3	PA 4	PA 5	PA 6	PA 7	PA 8
4-Seam Fastball	0.5	0	0.4	0.333	0	0	0	0.2
2-Seam Fastball	0	0	0	0.333	0.333	0.2	0	0.2
Sinker	0	0	0	0	0	0	0	0
Cutter	0.5	0.667	0.4	0.333	0.667	0.8	1	0.4
Curveball	0	0	0	0	0	0	0	0.2
Slider	0	0.333	0.2	0	0	0	0	0
Slow Curve	0	0	0	0	0	0	0	0
Changeup	0	0	0	0	0	0	0	0
Splitter	0	0	0	0	0	0	0	0

Methods

Under the naive Bayes model, the posterior distribution is defined as $p(\theta|y) \propto \frac{p(y|\theta)p(\theta)}{p(y)}$, where $p(y|\theta)$ is the likelihood, and $p(\theta)$ is the prior. In this case, the prior is the probability that the hitter gets on base in general. In baseball, this probability is referred to as a player's On Base Percentage (OBP). The likelihood is the probability that the hitter gets on base against a specific pitcher, given that the hitter has gotten on base. The probability $p(y)$ is defined as the probability that the given hitter faces the given pitcher.

The example hitter selected for this study is Seattle Mariner's third baseman, Kyle Seager. The example pitcher for the purpose of this study is Texas Rangers' pitcher Yu Darvish. He has been selected due to the fact that he throws many different kinds of pitches (this would only make the model more interesting to interpret).

In order to sample from the posterior, a sample was first drawn from the prior, which is of the form $N(\mu, \sigma)$, where the mean is Seager's season OBP, $\mu = 0.316$, and the standard deviation σ is the standard deviation of his season OBPs from 2011 to 2020. Then a sample is drawn from the likelihood, which is of the form $N(\mu, \sigma)$, where $\mu = 0.003$, and σ is sampled from a uniform hyperprior on [0.001, 0.0025]. $p(y)$ is also sampled from a normal prior, with a mean $\mu = 0.018$, and a uniform hyperprior on [0.001, 0.0025].

Using the posterior distribution, we can sample from the posterior predictive. The posterior predictive, $p(\tilde{y}|y)$ represents the probability that Seager would have gotten on base had he faced Darvish a 13th time in 2012. The posterior predictive is a normal distribution, with mean $E[\tilde{y}|y] = \mu$ that is the posterior mean. The variance of the posterior predictive is the variance of the posterior plus some unknown certainty, drawn from a uniform distribution on [0.0001, 0.00025].

With the same framework, it is possible to also estimate Seager's probability of getting on base against Darvish in n new plate appearances in the 2021 season. The prior mean in this case would be selected from a normal distribution with mean equal to the mean of his career On Base Percentages for each of his previous seasons. The standard deviation hyperprior is selected from a uniform distribution. The likelihood would be sampled from a normal distribution with mean equal to the mean number of times on base in previous seasons divided by the mean number of times Seager got on base against Darvish in previous seasons. The probability $p(y)$ would have mean equal to the mean number of plate appearances from previous seasons divided by the mean number of plate appearances against Darvish in previous seasons.

Under naive Bayes classification, the probability of a hitter getting on base against Darvish is defined as $p(y|x) \propto p(x|y)p(y)$, where $p(y)$ is the prior, defined as the probability a hitter gets

on base. The likelihood of getting on base, $p(x|y=1)$, is defined as $\prod_{j=1}^{d} \phi_j|_{y=1}^{x_j} \ (1-\phi|_{y=1})^{1-x_j}$, where ϕ_j is the probability that a hitter gets on base given that they have seen a certain pitch. The likelihood of not getting on base, $p(x|y=0)$, is defined as $\prod_{j=1}^{d} \phi_j|_{y=0}^{x_j} \ (1-\phi|_{y=0})^{1-x_j}$, where ϕ_j is the probability that a hitter does not get on base given that they have seen a certain pitch.

The likelihoods ϕ_j were further improved through Laplace smoothing. This is a technique used to smooth out categorical data, and solves the problem of small probabilities. The prior for the data under Laplace smoothing is $\frac{\sum_{i=1}^{n} 1\{y_i=C\}+1}{N+K}$, where K is the number of different values x can take. For the univariate model, we let $C=1$, and x can take two values: 0 and 1, so $K=2$. For the likelihood, Laplace smoothing is $\frac{\sum_{i=1}^{n} 1\{y_i=C,x=1\}+1}{\sum_{i=1}^{n} 1\{y_i=C\}+K}$, where again, $C=1$ and $K=2$.

It may be that the classification can be improved upon by using a multivariate model instead of a simple univariate one. Instead of using a '1' to denote the presence of a certain pitch in a plate appearance, the multivariate model uses an integer, denoted k, to represent the number of those pitches seen in the plate appearance. There are two $p \times K$ matrices of parameters $\phi_{y|k}$, where K is the maximum value of k in the training data. The first matrix is for plate appearances that end in getting on base, $y=1$, and the second is for plate appearances that end in an out, $y=0$. The probability that a batter gets on base is $\prod_{i=1}^{j} \phi_{y=1|k}\phi_{y=1}$. As with the univariate model, we let $\phi_{y=1}$ be the probability that the batter gets on base.

Again, Laplace smoothing will be used for the multivariate model. Recall that the prior for the data under Laplace smoothing is $\frac{\sum_{i=1}^{n} 1\{y_i=C\}+1}{N+K}$, where we let $C=1$, and K is the number of different values y can take. For the multivariate model, we find that $K=6$. The likelihood, with Laplace Smoothing, is $\frac{\sum_{i=1}^{n} 1\{y_i=C,x=1\}+1}{\sum_{i=1}^{n} 1\{y_i=C\}+K}$, where again, $C=1$ and $K=6$.

Multivariate data does not always tell the whole story. Anywhere between 1 and upwards of 18 pitches can be seen by a hitter in any given plate appearance. Therefore, it may be that the multivariative model can be further improved upon by calculating proportions. The equations are the same as for the prior and likelihood for the multivariate model, but the Training and test datasets are first converted to matrices of proportions instead of counts. Laplace smoothing can again be used for the proportional model, with $C=1$. In the data, the number of different proportions in 21, so therefore $K=21$.

One method often used for reducing test error is boosting. Boosting is meant to decrease the classification error by fitting a series of weak models, analyzing the misclassifications after each fit, adjusting weights for points based on misclassification, and refitting accordingly. In order to easily perform boosting, the Bayes classification process can first be converted to a decision tree. A decision tree classifies each plate appearance as either a success (hitter gets on base), or a failure (hitter does not get on base), based on the number of times the hitter saw each different kind of pitch.

Results

During the 2012 season, Kyle Seager had a 0.316 On Base Percentage against all opposing pitchers. He got on base 200 times during the 2012 season, and 6 of those times were against Yu Darvish, so $p(y|\theta) = \frac{6}{200} = 0.003$. Out of Seager's 651 plate appearances during the 2012 season, 12 of those appearances were against Darvish. Thus $p(y) = \frac{12}{651} = 0.018$. A 95% posterior interval is (0.425, 0.606). The distribution of the posterior is shown in figure 1.

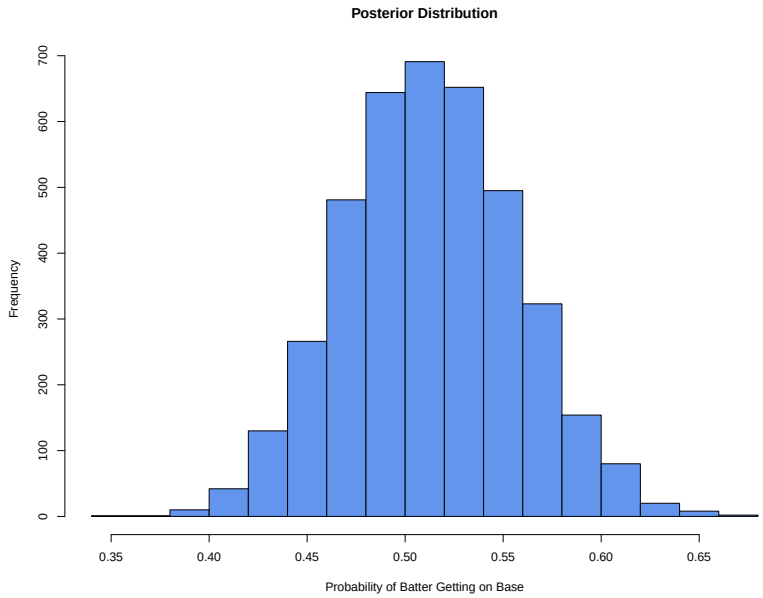

Figure 1: Posterior distribution for probability of Seager reaching base against Darvish

Using the posterior distribution, the posterior predictive distribution was estimated. Using 4000 draws from the posterior, the posterior predictive mean is $\mu = 0.513$, with a 95% posterior predictive interval of (0.290, 0.656). This interval is much wider than the posterior interval, which may in part be due to the added uncertainty. The distribution is shown in figure 2.

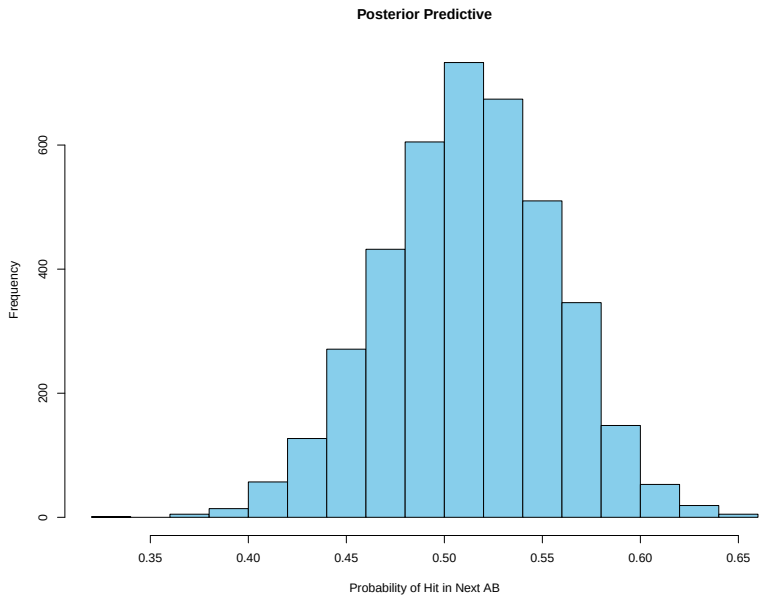

Figure 2: Distribution of probability that Seager gets on base in an additonal plate appearance against Darvish in 2012

The distribution of the posterior predictive for the probability of Seager getting on base against Darvish during the 2021 season is shown below in figure 3.

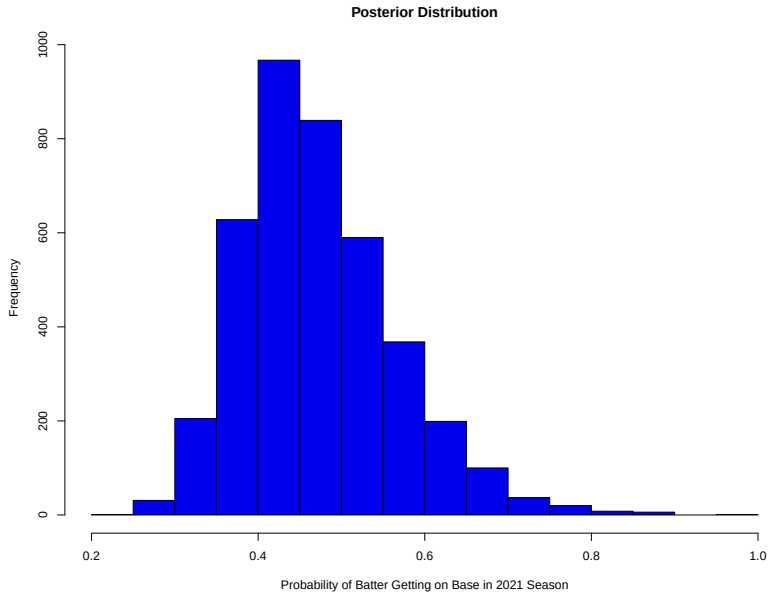

Figure 3: Distribution of probability that Seager gets on base against Darvish in future plate appearances during the 2021 season

For the univariate naive Bayes classifiction model, the classification error was approximately 0.28. The confusion matrix for the univariate model is shown below, where the columns represent the actual outcomes, and the rows represent the outcomes as predicted by the univariate naive Bayes classifier. The obvious weakness of the univariate model is that it tends to classify plate appearances which ended with the batter getting on base (y = 1) as failures (y = 0), in which the batter failed to get on base. For the test data, the univariate model classified all test plate appearances as failures (y = 0).

In all cases in which the batter did not actually get on base, the univariate model predicted the outcome correctly, but in all cases in which the batter did not actually get on base, the model did not predict the outcome correctly. Thus, for a true outcome of $y = 0$, the model's classification error was 0, but for a true outcome of $y = 1$, the model's classification error was 1.

	Actual	
Estimate	**Not on Base**	**On Base**
Not on Base	36	14
On Base	0	0

Using the data collected for this model, it is also possible to evaluate Darvish's most and least successful pitches. Given that Darvish throws a four-seam fastball in a plate appearance, a hitter has a 75.5% chance of getting on base. This is not surprising. Four-seam fastballs are by far the most common pitch in baseball. 78.1% of the plate appearances in the training dataset included a four-seam fastball. Four-seam fastballs also tend to be the easiest to hit, due to the fact that

there is little to no movement on the pitch. The second least sucessful pitch that Darvish threw was his cut fastball, which hitters had a 41% chance of getting on base given that they had seen it. His split-finger fastball appeared to be the most successful. Opposing hitters that saw Darvish's splitter only had a 2% chance of getting on base. This should be taken with a grain of salt, however, because he only threw 9 splitters in the training plate appearances.

For the multivariate naive Bayes classification mode, the classification error was 0.26. The confusion matrix for the multivariate model is shown below, where the columns represent the actual outcomes, and the rows represent the outcomes as predicted by the multivariate naive Bayes classifier. For plate appearances in which the batter did not actually get on base (that is, true y = 0), the multivariate model had an error rate of 0.081. For plate appearances in which the batter did actually get on base (that is, true y = 1), the model had an error rate of 0.375.

	Actual	
Estimate	**Not on Base**	**On Base**
Not on Base	31	10
On Base	3	6

Using similar data to the univariate model, it is possible to estimate the most successful patterns of pitches for Darvish. Given that Darivsh has thrown 2 cut fastballs in a plate appearance, the batter has a 31.25% chance of getting on base. Given that a hitter sees 1 four-seam fastball in a plate appearance, the batter has a 31.25% chance of getting on base. Interestingly enough, given that Darvish throws 2 cut fastballs in a plate appearance, the batter has a 25% chance of not getting on base.

The proportional model used the same Training and Test data as the multivariate model, but again, where each entry represented a proportion instead of a count. The testing error for the proportional model was 0.32. The confusion matrix for the proportional model is shown below, where the columns represent the actual outcomes, and the rows represent the outcomes as predicted by the proportional naive Bayes classifier.

As with the univariate model, the proportional model predicted all outcomes as failures (y = 0). For plate appearances in which the hitter actually did not get on base (that is, true y = 0), the proportional model had a classification error of 0. However, for plate appearances in which the hitter actually did get on base (that is, true y = 1), the proportional model had an error rate of 1. The proportional model thus appears to have the same weakness as the univariate model, predicting all outcomes to be failures.

	Actual	
Estimate	**Not on Base**	**On Base**
Not on Base	34	16
On Base	0	0

In order to attempt to reduce the testing error for both the multivariate and proportional models, the models were first converted into decision trees. Boosting was then performed on both decision trees. The boosted test error for the multivariate model was 0.396, while the boosted test error for the proportional model was 0.313.

Discussion

For the posterior distribution of the probability of Kyle Seager getting on base against Yu Darvish in the 2012 season, because the sample size of the observed data, y, is so small, the only way to do posterior analysis is through external validation. In the 2012 season, Seager in fact had 12 plate appearances against Darvish, and got on base in 6 of those chances, for an On Base Percentage of $\frac{6}{12} = 0.500$. The posterior distribution fairly accurately reflects this, with a center slightly higher than 0.500. Similarly, the 95% posterior interval of (0.425, 0.606) fairly accurately reflects a reasonable distribution of Seager's probability of getting on base against Darvish

The mean of the posterior predictive for an additional plate appearance in the 2012 season is about 0.512, which is very close to the mean of the posterior distribution. It would make sense that if Seager would have had one more plate appearance in the 2012 season, his probability of getting on base would be similar to his overall probability of getting on base against Darvish that season. The 95% posterior predictive interval for an additional plate appearance in the 2012 season was (0.290, 0.656). This is a much wider interval than that of the posterior distribution. While these seem to be fairly appropriate bounds, the spread is quite large, likely due to the uncertainty that was introduced in the estimation of the posterior predictive.

The posterior predictive mean for the 2021 season is around 0.482, which is lower than the posterior mean and 2012 posterior predictive mean. This makes sense, as 2012 was a very successful season for Seager against Darvish, so regression towards the mean would be expected. The 95% posterior predictive interval for 20201 is (0.323, 0.689). This interval is much wider than the posterior interval, but had similar spread to that of the posterior predictive for 2012, which may in part be due to the added uncertainty. There were several unrealistically high outliers, and as a whole, the distribution is skewed to the right.

The test error of the univariate model was 0.28. The multivariate model was expected to be lower than the univariate model, due to the fact that the multivariate model provides model details of the plate appearance. The multivariate model did yield a lower test classification error of 0.26. The proportional model provides still more detailed information about the plate appearance. Nevertheless, it yielded a 0.32 error rate. From the confusion matrices, it appears that the univariate model and proportional model were better at predicting the outcome of plate appearances that actually ended with the hitter not getting on base (that is, true y = 0). The multivariate model, however, appeared to be better at predicting the outcome of plate appearances that actually ended with the hitter getting on base (that is, true y = 1).

The differences between the test classification errors for the three models would suggest that the multivariate model is the best, followed by the univariate model. However due to the small sample size of both the training and test data, the difference in classification errors is not necessarily significant. Due to the fact that the univariate model and proportional model appeared to assign each test plate appearance as a failure (y = 0), and the multivariate model did not, it may be that the multivariate model would be the most accurate on a larger test sample.

It was hypothesized that the test error of the multivariate and proportional models could potentially be reduced through boosting. After boosting, however, the test error for the multivariate model was not reduced. The test error of the proportional model was reduced through boosting, but by an almost negligible amount. For the multivariate model, the boosted test error was 0.396, and the boosted test error for the proportional model was 0.313. The inability of boosting to reduce the classification error is as to be expected. Ting et. al, 2002, concluded that the Bayes classifier cannot be significantly improved through boosting.

There are definite limitations to this analysis. The data was collected from pitch-by-pitch data from games beginning in 2014, when this type of data was first recorded. Pitch identification is not an exact science, and thus it is entirely possible that some pitches got misclassified. With some pitches in particular, it is easy to confuse different types. Furthermore, the inning number and game number from which the data was selected was generated through a random number generator, but that does not mean the data was perfectly randomly sampled.

Due to the random selected of the game and inning number, the data collection was extremely computationally expensive. It took approximately 1 hour to collect data on 50 plate appearances. This was the greatest drawback of the analysis. The test error for all models could likely be reduced with more data.

As compared to other models, the naive Bayes classifier in this analysis fared poorly. Doo et. al., 2018, used a log5 model to estimate the outcome of plate appearances. They achieved a test MSE of 0.0514 for their initial model, and a test MSE of only 0.0428 for the second model [2]. Weir, 2020, used logisitic regression to predict the outcome of a plate appearance. In the Weir model, however, outcomes were not recorded as a binary "on base" vs. "not on base", but rather as multiple different outcomes, including a walk, single, double, triple, etc.

Weir's model predicted a strikeout with an error rate of 0.329, a walk with an error rate of 0.157, a home run with an error rate of 0.066, a single with an error rate of 0.261, a double with an error rate of 0.099, a triple with an error rate of 0.009, and a hit by pitch with an error rate of 0.021. It is difficult to compare Weir's model to the naive Bayes model because the outcomes of interest between the two models are different. The multivariate naive Bayes model produced the lowest test error, of 0.26.

Silver (2020) used a neural networks model that improved upon the log5 method. The two outcomes Silver was most interested in predicting were strikeouts, for which he yielded an error rate of about 0.21, and home runs, for which he yielded an error rate of about 0.46.[3] Again, it is difficult to compare Silver's model to the naive Bayes model in this paper, as the two approaches attempt to predict slightly different outcomes.

It should be noted that these models all assume that the independence assumption applies to plate appearances in baseball. That is, these models all assume one plate appearance in independent - or has no effect - on the last. This is not widely accepted as true. Often the last plate appearance effect what pitches the pitcher throws to the next batter, and how effective those pitches are. Therefore, some of the test error in the analysis may be due to the effect of previous plate appearances that was not quantified.

Based on the confusion matrices and classification errors, it appears that a naive Bayes classification model, particularly a multivariate model, does a fairly good job of predicting the binary outcome of a plate appearance. Further research could improve upon the models used in this paper to predict more than one outcome of a plate appearance, as done in the other studies referenced. The naive Bayes models focused on opponents' plate appearances against one specific pitcher, but the same models could just as easily be used in a future study for the plate appearances of one hitter against a variety of pitchers - this would allow for the analysis of the pitches a hitter performs best against.

Sources

[1]Doo, Woojin, and Heeyoung Kim. "Modeling the Probability of a Batter/Pitcher Matchup Event: A Bayesian Approach." PLOS ONE, vol. 13, no. 10, Oct. 2018, p. e0204874. PLoS Journals, doi:10.1371/journal.pone.0204874.

Jensen, Shane. "Hierarchical Bayesian Modeling of Hitting Performance in Baseball." Bayesian Analysis, vol. 4, no. 4, 2009, pp. 631–52.

[4]Powers, Scott, et al. "Nuclear Penalized Multinomial Regression with an Application to Predicting at Bat Outcomes in Baseball." Statistical Modelling, vol. 18, no. 5–6, 2018, pp. 388–410. PubMed Central, doi:10.1177/1471082X18777669.

S, Yang. "An Introduction to Naïve Bayes Classifier." Medium, 8 May 2020, https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf.

[3]Silver, Joshua. "Singlearity: Using A Neural Network to Predict the Outcome of Plate Appearances." Baseball Prospectus, 9 July 2020, https://www.baseballprospectus.com/news/article/59993/singlearity-using-a-neural-network-to-predict-the-outcome-of-plate-appearances/.

Ting, Kai Ming, and Zijian Zheng. "Improving the Performance of Boosting for Naive Bayesian Classification." Methodologies for Knowledge Discovery and Data Mining, edited by Ning Zhong and Lizhu Zhou, Springer, 1999, pp. 296–305. Springer Link, doi:10.1007/3-540-48912-6_41.

[2]Weir, Davis. "Predicting MLB Plate Appearances With Logistic Regression." Medium, 10 Aug. 2020, https://medium.com/@ndavisweir/predicting-mlb-plate-appearances-with-logistic-regression-

Zhang, Zixuan. "Boosting Algorithms Explained." Medium, 7 Aug. 2019, https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30.

"Yu Darvish Game by Game Stats and Performance." ESPN, https://www.espn.com/mlb/player/gamelog/_/id/32055/year/2017. Accessed 4 Dec. 2020.

BrooksBaseball.Net Player Card: Yu Darvish. http://www.brooksbaseball.net/landing.php?506433. Accessed 4 Dec. 2020.

Kyle Seager - Stats - Batting — FanGraphs Baseball. https://www.fangraphs.com/players/kyle-seager/9785/stats?position=3B. Accessed 4 Dec. 2020.