Predicting Swing and Miss Percentage for Pitchers Using Pitch f/x Data

Adam Yudelman

Introduction

For a Major League Baseball franchise, making the most of a limited payroll, is what separates franchises that do not have the financial resources of teams like the New York Yankees and Los Angeles Dodgers. In an effort to do this, teams are committed to the idea of arbitrage, where they attempt to identify undervalued players. In a method popularized by the movie *Moneyball*, teams use data analysis to find quantitative values of the players. Initially, this analysis focused on box score statistics calculate more advanced stats, like on-base percentage and slugging percentage, which built on the more basic average and home runs valuations. Next, analysts used play-by-play data to calculate expected run values in different situations and find players who either outperformed or underperformed the situations. About a decade ago, the amount of data available exploded with the introduction of Pitch F/X, which uses two cameras to track pitches and record statistics like speed, ball rotation, release point, and movement. More recent technology now simultaneously tracks the movement of all players on the field.

Making the Model

My project uses Pitch F/X data to try find undervalued pitchers. My goal is to build a model that predicts how well a pitcher can cause swing and misses. Baseball is a game that involves a lot of luck. A pitcher can only control the pitch that he throws, but not whether the hitter is looking for a specific pitch in a specific location. In that, a player may get unlucky. I wanted to build a model to find pitchers who showed the characteristics of a high strike out pitcher who may not have had the statistics that showed that. The data I used was the complete Pitch F/X data from the 2013, 2014, and 2015 seasons. This adds up to over 2 million separate pitches. I then broke this data down into just swings, classified as pitches where there was a swinging strike, a foul ball/tip, or a ball hit in play, which came out to roughly 600,000 individual observations. I made a dummy variable to identify contact (foul or ball in play) or swing and miss to use as a response variable. To build the model, I used roughly 66% of the data (the 2013 and 2014 seasons) as the training set and the remaining (the 2015 season) as the test set.

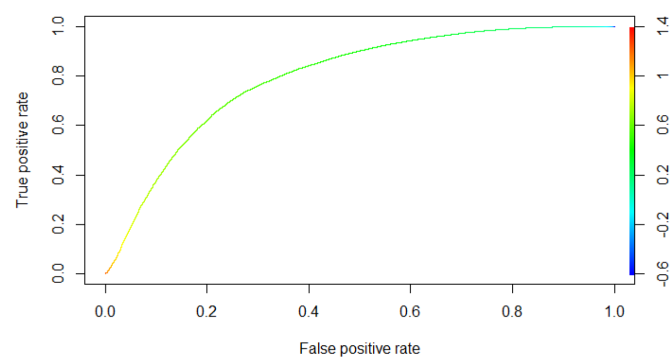
Before selecting a model, I looked at the variables to try select which ones could be predictive. I selected release velocity to measure the speed of the pitch, the movement of the pitch in both the horizontal axis and vertical axis, and the location of the pitch as it crossed the plate. All these variables are in complete control of the pitcher and his skill level. Because baseball has significant splits when pitchers are facing batters of a different hand, I added an indicator variable for whether the pitcher was facing a player of like-handedness. An additional problem I needed to solve was how to take into account different pitch types. A curveball has very different characteristics than a fastball; hence, a fastball that acted like a swinging strike would look very different than a curveball. To fix this problem, I decided to build different models for each pitcher.

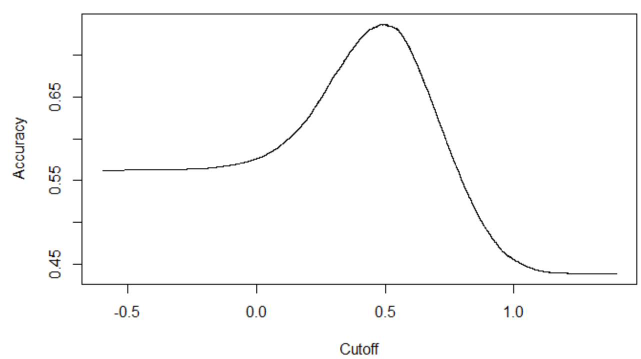
I approached my model knowing that I needed to use 2-type classification model. The possibilities included decision trees, support vector machines, neural networks, and logistic regression. After testing out all three, I ended up using a logistic regression. Using the variables described above, my model is as follows:

Pitches$Contact ~ Intercept + Pitches$releaseVelocity + Pitches$xmov + Pitches$ymov + Pitches$px*Pitches$SameHand + Pitches$pz*Pitches$SameHand + Error

Where xmov and ymov are defined as the delta in the location of the pitch from the first recording of pitch location (50 feet from the plate) and the pitch location as the ball crosses the plate.

The nature of logistic regression requires the use of ROC to find the cutoff point for the predicted value and the correlation matrix to see the accuracy of such a model. Below is an example of the ROC curve and accuracy plot for the changeup model:





Using these plots and the optimal cut off point to maximize accuracy, I was able to store the corresponding cut off for each model to be used in the testing set. The following shows the accuracy for each model for each pitch for the training set.

Pitch	CU	CH	FC	FF	FT	KC	SI	SL	FS
Acc	.737	.677	.674	.732	.789	.778	.787	.7105	.703

We can see that the models are most predictive for two seam fastballs (FT), sinkers (SI), and knuckle curves (KC). The models struggle to predict change ups (CH) and cutters (FC). The following shows the confusion matrix for the entire testing set:

	Miss	Hit	Acc
Miss	59669	24257	0.71097157
Hit	86345	240130	0.735523394
		Total acc:	0.730502606

Given these diagnostics, I am content with my model. In order to scale the values of the model predictions so that I could compare pitches to each other, I normalized each value. To further look at the diagnostics, I wanted to look at which pitchers in the training set are rated highly and rated poorly. Given my response variable, players who have low values are more likely to get swings and misses and high values are pitchers prone to contact. The following show the top ten for players who threw over 500 pitches with swings over the two seasons:

Player	Value	Pitches
Craig Kimbrel	0.475806407	604
Greg Holland	0.495318432	671
Shawn Kelley	0.518790296	523
Cody Allen	0.522090642	675
Al Alburquerque	0.522239695	566
Aroldis Chapman	0.550608747	580
David Robertson	0.554313312	553
Mike Dunn	0.556003545	643
Trevor Rosenthal	0.56388352	800
Carlos Carrasco	0.565498894	880

To try filter out relievers, the following shows the top ten players who threw over 1000 pitches with swings:

Player	Value	Pitches
Francisco Liriano	0.56880326	1661
Madison Bumgarne	0.586004207	2181
Alex Cobb	0.586531638	1454
Gerrit Cole	0.589634957	1170
Max Scherzer	0.591253319	2177
Jordan Zimmermar	0.592812947	1900
Stephen Strasburg	0.596752591	1807
Tyson Ross	0.597641995	1523
Clayton Kershaw	0.600532036	2073
Anibal Sanchez	0.604731551	1584
A.J. Burnett	0.605569981	1832

To contrast this, we can also look at the bottom ten players who threw over 1000 pitches with swings:

Player	Value	Pitches
Doug Fister	0.734378854	1727
Bronson Arroyo	0.733381165	1231
Bartolo Colon	0.727956295	1710
Henderson Alvarez	0.727677296	1255
Mark Buehrle	0.725778425	1824
Jered Weaver	0.717302152	1727
Felix Doubront	0.709114006	1144
Brandon McCarthy	0.707773895	1531
Travis Wood	0.707430319	1669
Jhoulys Chacin	0.705775689	1155

Given my knowledge of baseball, all three charts pass the eye test. When evaluating the test data set, however, I will look at the pitcher's Strikeouts per 9 innings and contact% and see how well is correlates with my ratings.

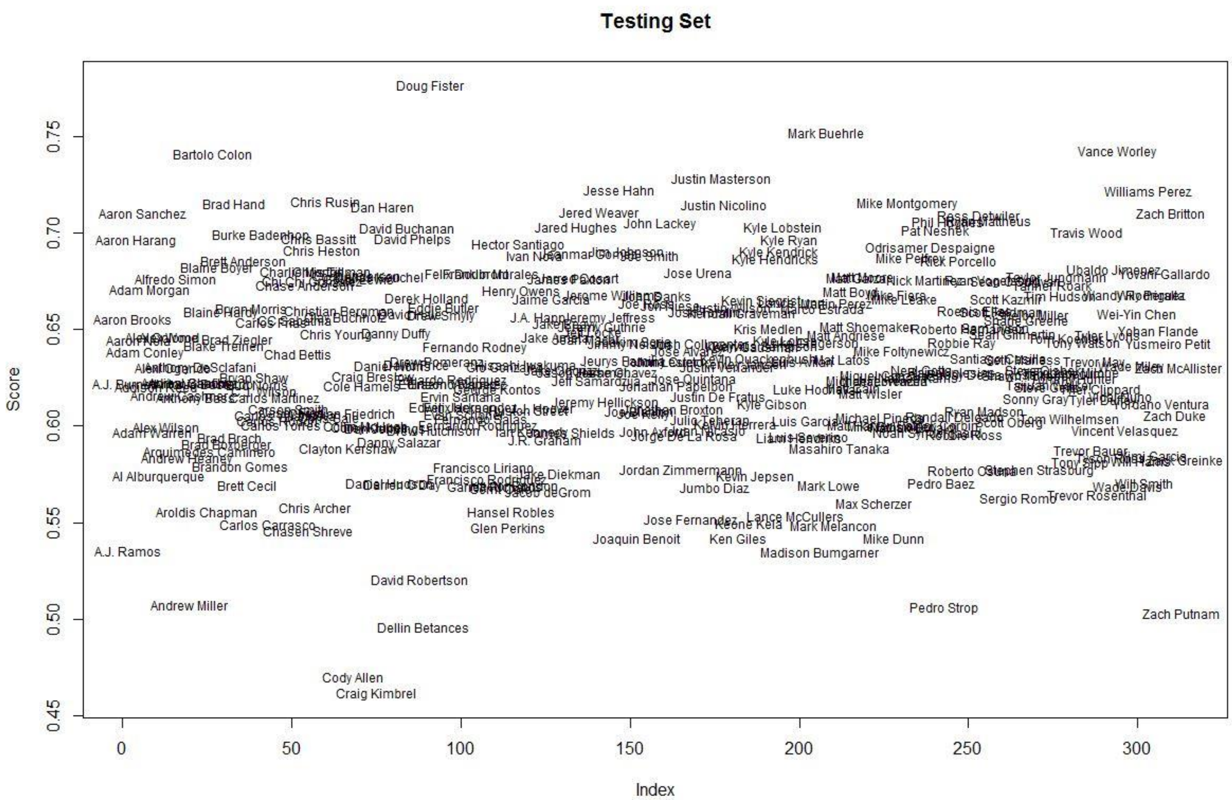
<u>Testing the Model</u>

Moving on to my test model, I looked at the 2015 data. The following again shows the accuracy for each pitch model and to confusion matrix:

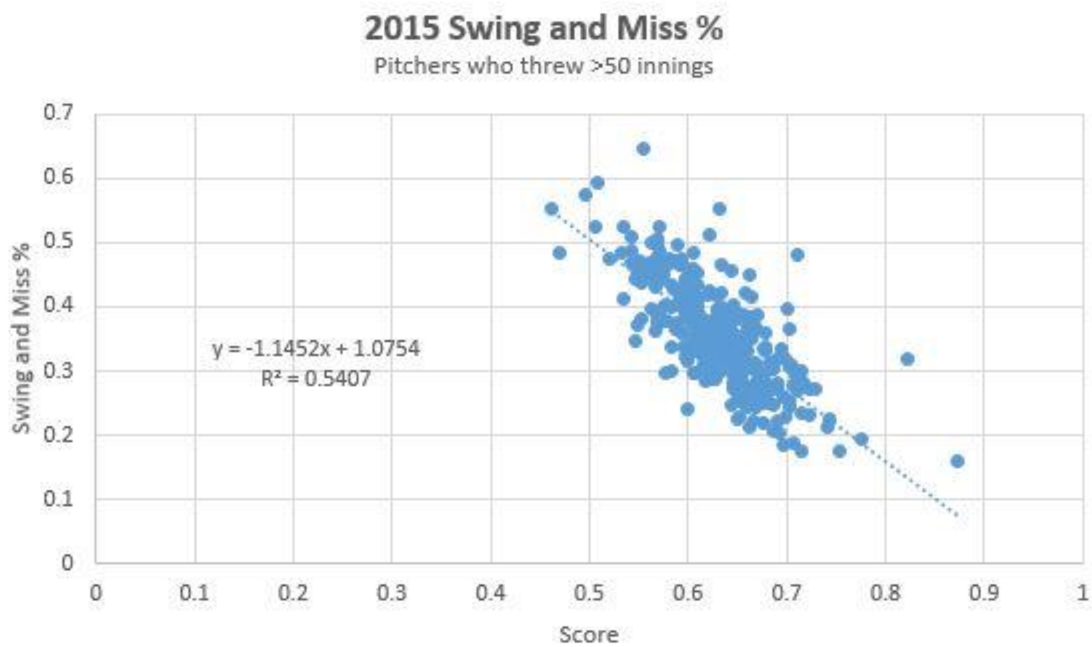
Pitch	**CU**	**CH**	**FC**	**FF**	**FT**	**KC**	**SI**	**SL**	**FS**
Acc	.749	.685	.690	.742	.803	.789	.805	.775	.718

	Miss	Hit	Acc
Miss	40698	13915	0.745207
Hit	38084	116828	0.754157
		Total acc:	0.751824

The following plot shows the scores for the players with the player names for pitchers who threw at least 250 pitches with swings:

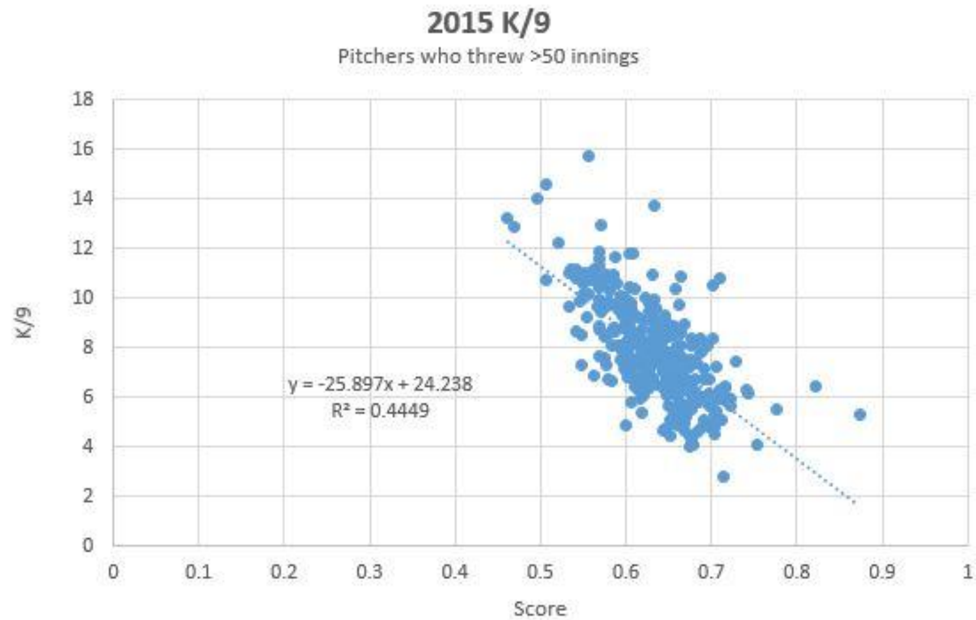


As we can see, Kimbrel, Allen, Betances, and Miller are among the best while Fister, Colon, Buehrle, and Worley are easily the worst. While these scores are good, they mean nothing if they are not predictive of swing and misses for the pitcher. I plotted their actual swing percentage against the predictive values:



I am very happy with these results. Players who are underneath the regression line underperformed their pitch repertoire, so I would suggest they would get more swings and misses in the 2016 while pitchers above the line over performed their repertoire. If I were a team, I would target these underperforming pitchers in free agency and trades, for they are potentially undervalued given their performance.

I also wanted to see how these predictions translated to K/9, so I plotted the scores against the K/9 for the 2015 season and found a predictive equation:



Again, I am happy with the results. This suggests that this model's predictions translate to tangible success.

<u>Extending the Project</u>

If I could extend this project, I would add more variables to control for the batter's skill in the regression. I would do this so as not to punish pitchers who are pitching against more difficult lineups. I also could have built models specifically for left-handed vs left-handed situations, left-handed vs right-handed situations, etc. to see quadrupling the number of models would improve the accuracy of the predictions despite losing sample sizes in the training set. I also would have changed my value for "Contact" to 0 and "Swing and Miss" to 1 so that a higher score would mean a "better" pitcher. I think this would make interpreting the graphs a bit easier.

I would like to use more data from more seasons to see whether my pitchers who are outperforming their score do indeed regress the following year. I also would like to use a more powerful computer to try some different machine learning techniques. Often I found that my computer would freeze when trying to use SVM. I think there are plenty of different ways to model this question, but I just did not have the time or computing power to do so.