Predicting Whiffs, by Cody Stumpo

Objective

Use data on 1.2M swings from major league baseball to form a model that predicts and explains what characteristics of a pitch lead to a swing and a miss. As I don't have data on all pitches, just swings, I can't explain what induces a batter to swing in the first place. Only given that they decide to swing, what kind of pitch will make them miss. Without complete pitch data, I also can't look into the effect of pitch sequencing on whiffs.

But I do have 30 attributes of each pitch from pitch f/x about the position, speed, and acceleration in 3 dimensions, as well as spin and some contextual data around who the pitcher and batter were. Pitch f/x itself categorizes the pitches into types (fastball, curveball, etc...), so when you put this all together it's quite a rich dataset and lets us build a fairly good model in terms of both predictive power on a test set and in terms of explanatory power.

Exploratory Data Analysis

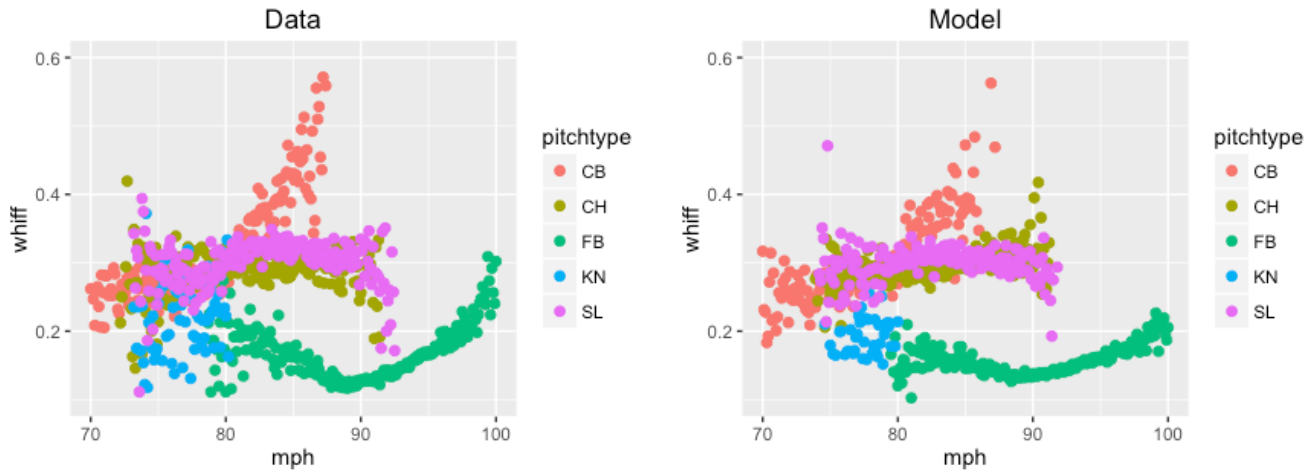
First, I made some plots of the data to get a sense of it, using my knowledge of baseball to guide what dimensions might be useful to look at together. Right away, we get some strong predictors. High velocity is critical for getting whiffs on curveballs and fastballs, but not other types of pitches. For curveballs and fastballs furthermore, the relationship has some interesting hingepoints we'll want the model to capture. There is not much connection between slider or changeup or knuckleball speed and whiffs, although it is important to note that there are strong differences between simply the expected whiff rate by pitchtype. The overall whiff rate across the whole dataset, by the way, is close to 20%.

Looking at some other attributes a few at a time, I learned spin matters, especially for knuckleballs (more is better) and fastballs (similar to velocity, a V-shape with a bottom around 1700 rpm). If you can get someone to swing at balls out of the strike zone in either dimension, of course that leads to more whiffs. But this is especially true for breaking pitches in or off the plate. I learned some useful patterns in the data that I can leverage in the modeling, by transforming the variables to ones that nudge the model fitting toward finding the relationships highlighted above.

Modeling

After using baseball knowledge to transform the data and boost predictive power (feature engineering, in the trade), I applied several types of models to find the relationship between the original 30 variables and the swinging strike rate. I selected some powerful, but quick-to-run algorithms that are well suited to this sort of problem. In particular, I chose Multivariate Adaptive Regression Splines (MARS), Generalized Linear Models (GLM), and Recursive Partitioning and Regression Trees. The three models performed about equally well, and taking a blend of the three improved predictability slightly.

I was able to measure the models success by holding back a random 20% of the training sample to test against, so that I could iterate on modeling parameters. Taking different random samples and repeating the tests ensured I was not 'overfitting' the data, which might lead to poor predictability on the true test set. In addition to common statistical measures coming out well, I was pleased to see that I could largely reproduce the exploratory data analysis above on new data by using the model!



Pitch Attribute Significance

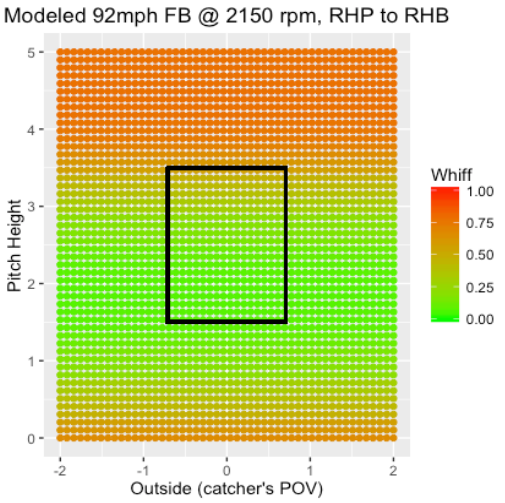
In addition to predictive power, the models help explain the relationships in the data. The models agree the pitch characteristics that most control whiffs are pitch-height, then some combination of being outside, being a fastball or offspeed, and the contact-tendency of the batter. Through the model, I find nuggets in the data like swings at very low offspeed pitches lead to whiffs 90% of the time while swings on high pitches to high-contact batters lead to whiffs only 10% of the time.

Conclusion

The model can be used in any number of practical ways for a Major League Baseball team.

While it is hard to translate minor-league performance to major-league performance, this study lets us get a very good estimate of how a pitcher's stuff will play in the big leagues. Given his typical measurements on the pitch f/x variables for each of his pitches, we can with high confidence gauge how that will translate to whiffs. Whiffs are known to be a primary predictor of strikeouts, and strikeouts a primary predictor of pitching success in general.

Also, we can get insight into how major league pitchers might be able to use the stuff they have to generate more whiffs by targeting certain locations with certain pitches (e.g. if your fastball is 92 mph with a spin rate of 2150 rpm, where are certain classes of batters likely to hit it vs. miss it). Or we could use the model to understand the impact an adjustment might have on a pitcher's whiff rate, given the impact of the adjustment on his pitch f/x. These points are true even for combinations of pitch attributes with little support in the data, as the model interpolates and extrapolates to fill in the gaps. You could form a game plan for a pitcher against a lineup he hadn't faced much before.



Given the time constraints of the exercise, I did not use any computationally intensive techniques, like neural networks or random forests, but those often perform even better than the algorithms I did use. Model-building is a process, and the initial analysis here is promising enough to invest further research.