Using PITCHf/x Data to Visualize Pitches and Predict Strikes

Chad Schupbach, David Jones, Rajeev Kumar, Ryan Miller

CSE 6242: Data & Visual Analytics, Spring 2019
Georgia Institute of Technology

Summary

What is the problem?
Current research is focused on classifying past pitches into different pitch types. There have been no experiments on predicting if future pitches will pass through the strike zone.

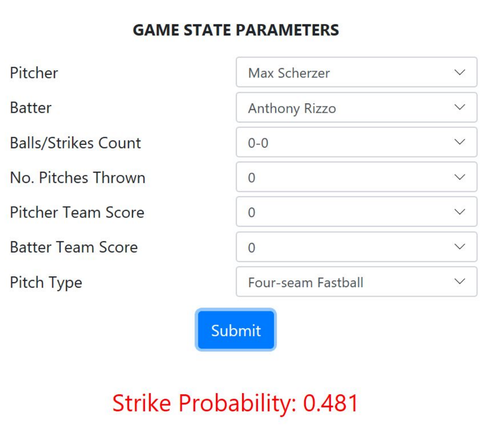
Why is it important and why should we care?
Prediction of future pitches developed with **machine learning techniques** can provide constructive assistance to batters by providing a better understanding of which pitches are more likely to appear in different game situations based on strike probability. An **interactive platform** that delivers these predictions along with **meaningful visualizations** of past pitch data as context can also engage fans, serious and casual, as an innovative way to interact with the game.

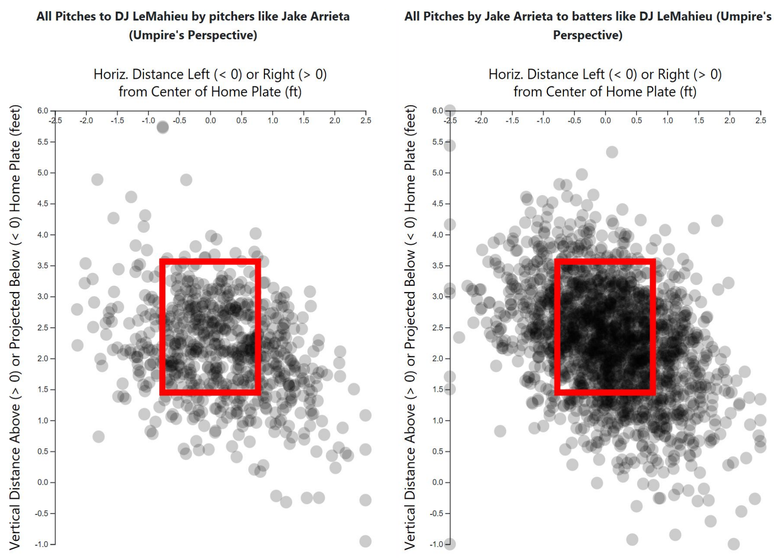
Methods

What is our approach?
Our approach involves:
- Clustering pitchers by the physics-based aspects of their pitches (e.g. spin rate) and batters by their performance statistics (e.g. strikeout rate) to account for sample size concerns in pitcher-batter matchups
- A machine learning model to predict the probability of a pitch passing through the strike zone
- An extremely flexible and tunable interactive user interface

How does it work?
We started by aggregating our data at the player level before using the K-Means algorithm to cluster the pitchers and batters in order to increase the number of observations for each matchup. Then, we determined which situational aspects of a pitcher-batter matchup are important predictors by using a Gradient-Boosting Decision Tree to iteratively remove the least important situational variables, performed Grid Search to determine the best hyper-parameters, and compared the models' 6-fold cross-validation prediction accuracy on the entire training set to choose the final model.

After selecting our final model, we output the predicted probability of a pitch being in the strike zone given the situation selected by the end user, which was fed into our visualization platform.
Through our interactive user interface, users can match up a pitcher and batter, and select other features on the screen to probabilistically predict whether future pitch would be strike. Visualization shows all historical pitchers between 2015 and 2018 in and around the strike zone for the pitcher and alike batters, and for the batter and alike pitchers.





Why do you think it can effectively solve your problem?
We believed that because of the millions of observations and wide range of features we had at our disposal, there was enough data for our machine learning algorithms to determine what each pitcher cluster will do in a given situation.

What is new in your approach?
While current research generates curiosity among viewers and helps them guess batter performance in upcoming matchups within the season, little research is available on predicting the pitches based on the historical data on batters against possible pitchers. This system predicts whether future pitches are within the strike zone, and help batters identify the scope of improvement based on the pitches in previous matches.

Data

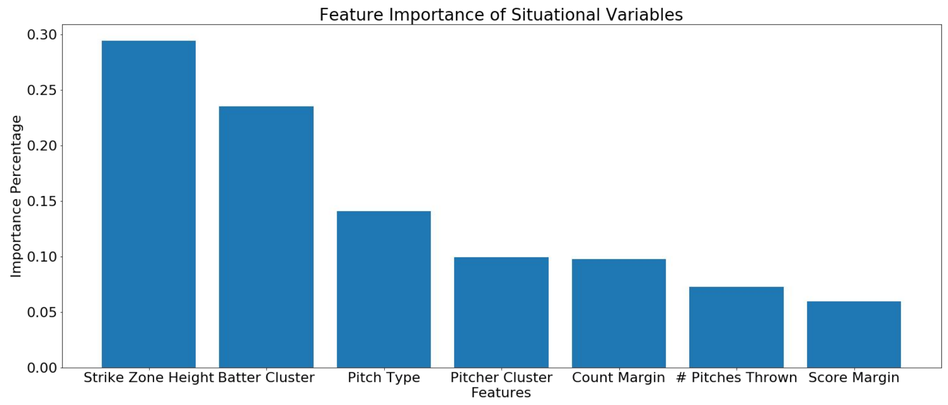
How did you get it?
- MLB pitch data from 2015 to 2018 was downloaded from Kaggle and included the following datasets: *pitches*, *games*, *atbats*, and *player_names*.
- Historical data for 3 pitchers and 3 batters from 2015 through 2018 was manually captured from *baseball-reference.com*.

What are its characteristics?
- The *pitches* dataset contains ball movement and situational data for 2.87M individual pitches. The *atbats* and *player_names* datasets were merged with *pitches* to provide further information about the matchup.
- Statistics were aggregated for individual batters and pitchers prior to clustering.
- Pitcher and batter clusterings were merged with historical pitch location data for 3 select players of each position to be used in a demo of the Strike Prediction System.
- The dataset used for our demo contains the situational strike probability of a hypothetical pitch between the 3 batters and 3 pitchers.

Experiments and Results

How did you evaluate your approach?
The experiment is designed to answer the following questions:
- Should a batter be expecting a strike based on the given situational factors?
- Which situational factors are most important in determining the strike probability of a given pitch?
- Because some batters tend to avoid swinging at the first pitch of an at bat, are opposing pitchers more likely to throw them first pitch strikes?
- Does a pitcher's probability of throwing a strike decrease when facing a batter who has a high walk rate?
- Do batters with high strikeout rates have a higher probability of being thrown a strike?
- Do high stress situations impact strike probability?
- Which algorithm will most accurately predict if a pitch will be a strike?

What are the results?
To determine which situational aspects of a pitcher-batter matchup are important predictors, we iteratively removed the least important situational variable (e.g. number of outs) until all variables were above a 5% threshold. We found the importance of each feature using a gradient-boosting decision tree that was repeatedly trained on half of the training dataset. We ended up with 7 important features:

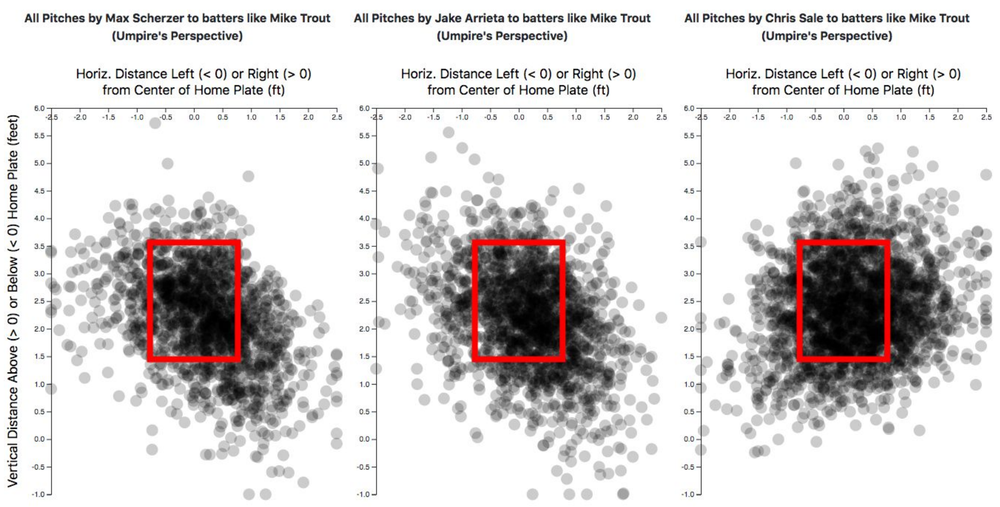


To select the best algorithm for our model, we split our data into two sets: training (80%) and testing (20%). For each algorithm, we performed Grid Search using 25% of the training dataset to determine the best hyper-parameters and compared the models' 6-fold cross-validated accuracy on the entire training set:
- The Gradient-Boosting Decision - 57.80%
- Random Forest Algorithm - 57.69%
- Artificial Neural Network - 57.48%.

Because it had the best cross-validated accuracy, we chose the Gradient-Boosting Decision Tree for our final model and tested it on our test dataset, where it achieved an accuracy of 58.00%.

We attribute the lack of success predicting whether a strike will pass through the strike zone due to the large variability in what individual pitchers will do in a given situation. By clustering the pitchers and batters to increase the observation count, we made it more difficult to accurately predict the outcome.

The pitch scatter plots offer additional insights and conclusions. The three charts below show pitch data for three different pitchers against the same cluster of batters. Chris Sale's data is a mirror image of the others', confirming his different handedness (he throws left-handed; the others throw right-handed.). In addition we can see that among all three, Jake Arrieta has the widest spread of outliers (more pitches further from the strike zone), Max Scherzer's non-strikes tend to fall directly to the right of home plate, while Chris Sale's non-strikes are more evenly distributed around the strike zone. This type of information, easily gleaned from the visualization, is very valuable to batters preparing for particular pitchers.



How do your methods compare to other methods?
Because there are currently no other models predicting the future probability of a strike, we have nothing to compare our results to.