Using Machine Learning Algorithms to Identify Undervalued Baseball Players

Tatsuya Ishii, twishii@stanford.edu

Predicting

The objective of this project is to use clustering algorithms to identify undervalued players in Major League Baseball (MLB). These players provide outsized contribution to teams relative to their financial commitments and teams can gain financial edge in roster construction. The guiding principle of the project is to focus on process over results: the assumption is that struggling players who are similar to successful players in terms of process are prime candidates for improvement in the subsequent season. By processing data on player process, the clustering algorithms were able to identify intriguing players and with the addition of new features generated in this step, it opened the way to build an end to end model by predicting player improvement.

Data

With the advent PITCHf/x and Statcast, teams have access to very granular data from all 2430 games that take place over the course of the season. Both of these data sources are hosted by BaseballSavant and contain wealth of information on player process. To evaluate the clustering algorithms, Fangraphs hosts a varied assortment of advanced analytics measures that give insight to player value.

Features

To analyze player process, the data from PITCHf/x and Statcast lend themselves to focus on pitchers. In evaluating pitcher process, some of the variables that was looked at can be roughly grouped to:

- Movement
- Velocity Component
- Acceleration Component
- Spin Rate
- Effective Velocity

Per pitch type. In all, 15 variables were used for player identification clustering analysis.

Models

Player Identification: Clustering Algorithms

- K-means Clustering

$$Cluster\ Centroid: \mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

- Hierarchical Clustering

$$Complete\ Linkage:\ d_{CL}(G,H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'} \quad for\ pairwise\ dissimilarity\ d_{ii'}$$

Performance Improvement: Supervised Learning

- Boosting
- Random Forest

Evaluate the predictive models using Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y_i})^2}$$

Results

Both the K-means and Hierarchical clustering algorithms produced similar set of players at the top. To determine which algorithm yielded better quality of results, examine the ranking in which the players were deemed to be most likely candidates for improvement between the two algorithms.

WAR/IP Delta			WAR/IP 2016			WAR 2016		
Pitch	KM	HC	Pitch	KM	HC	Pitch	KM	HC
FF		X	FF		X	FF	X	
SL		X	SL	X		SL		X
CU	X		CU	X		CU		X
CH		X	CH		X	CH		X
SI		X	SI		X	SI		X
FC	X		FC	X		FC	X	
KC	X		KC	X		KC	X	
FT		X	FT	X		FT		X

Discussion

Even though the two algorithms both identified similar top candidates, its success can be further broken down into what the end goal is and how to define undervalued. If the goal is to identify who was the most valuable on a per innings basis, then the K-means algorithm produced the better set of results. On the other hand, if the objective is to identify who improved the most or was the most valuable, then Hierarchical clustering had the upper hand.

Regardless, both algorithms successfully identified intriguing players who improved their performances after having a down year. The results obtained from the clustering analysis however did not dramatically improve the prediction performance of the predictive models. Nevertheless, the results demonstrate the potential of PITCHf/x and Statcast data and the process oriented approach to evaluating baseball players.

Future

One of the most fertile grounds for further exploration is the method by which these undervalued players are identified after grouping them into clusters. Incorporating other advanced metrics such as Fielding Independent Pitching (FIP) and ERA- in the identification process may lead to more refined results. In addition, there is more potential to pursue clustering based on pitch repertoire rather than on individual pitch basis.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York, NY: Springer, 2009.

Data Sources:
Baseball Savant: https://baseballsavant.mlb.com/
Fangraphs: http://www.fangraphs.com/