To Switch or Not to Switch

Cleaning/Functions

I cleaned the data by dropping columns with over 95% of its values being NA. I did not want to lose a whole column if there were not many NA values in it, so I thought 5% was a good threshold. For the columns with less than 5% of its values being NA, I used forward fill to fill the NA values with new values based on the previous row.

I built a Decision Tree classifier function that would split the data into train and test sets with a 70% split, fit the data, make training and test predictions, print a classification report, print a confusion matrix, print the features with at least 3% feature importance, and plot those features in a bar plot. This function allows you to choose what criterion you want for the model when you call it.

I built a Random Forest classifier function that would split the data into train and test sets with a 70% split, fit the data, make training and test predictions, print a classification report, print a confusion matrix, print the features with at least 3% feature importance, and plot those features in a bar plot. This function allows you to choose the maximum depth for the tree for the model when you call it.

I built a Adaboost classifier function that would split the data into train and test sets with a 70% split, fit the data, make training and test predictions, print a classification report, print a confusion matrix, print the features with at least 3% feature importance, plot those features in a bar plot, and print a prediction for a data frame one wishes the model to make a prediction on. When calling the function, you can send a data frame for the model to make a prediction on and it allows you to choose the learning rate, number of estimators, and what algorithm for the model to use when you call it.

I chose these 3 classifiers since they give feature importance scores for the features in its model, and I liked the idea of using a boosting and forest classifier since people find success with those types of models.

Savant

I recoded the values in the description variable to positive, negative, and neutral based on what the outcome of the pitch resulted in and if that result was positive, negative or neutral for a pitcher. The values “ball”, “blocked_ball”, and “hit_by_pitch” were recoded into negative values. The values 'foul', 'called_strike', "swinging_strike", "bunt_foul_tip", "foul_bunt", "foul_tip", "missed_bunt", "swinging_strike_blocked", and "unknown_strike" were recoded into positive values. The values “pitchout” and “hit_into_play” were recoded into neutral. The value “hit_into_play” was made neutral since this value was likely positive sometimes and negative other times for a pitcher. I dropped the neutral values since I was not interested in predicting those. I then split the data into starters and relievers.

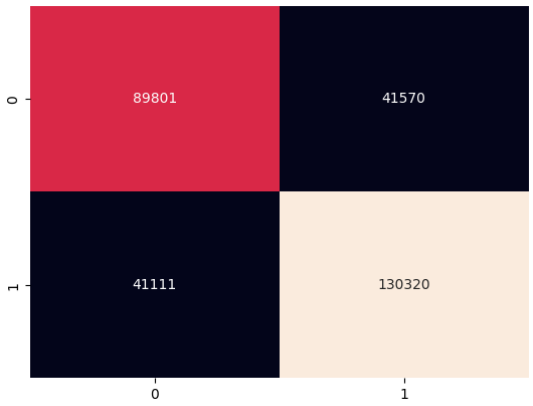
I decided to make classifiers that predict if the result of a pitch is positive or negative based on the pitch/game info.. One could find a starter that tended to have good values for the RP Decision Tree classifier’s 8 most important features in 2023. That player should have success as a starter if he puts up those values as a starter for those features. That is a big if, but it is what a team hopes for when changing

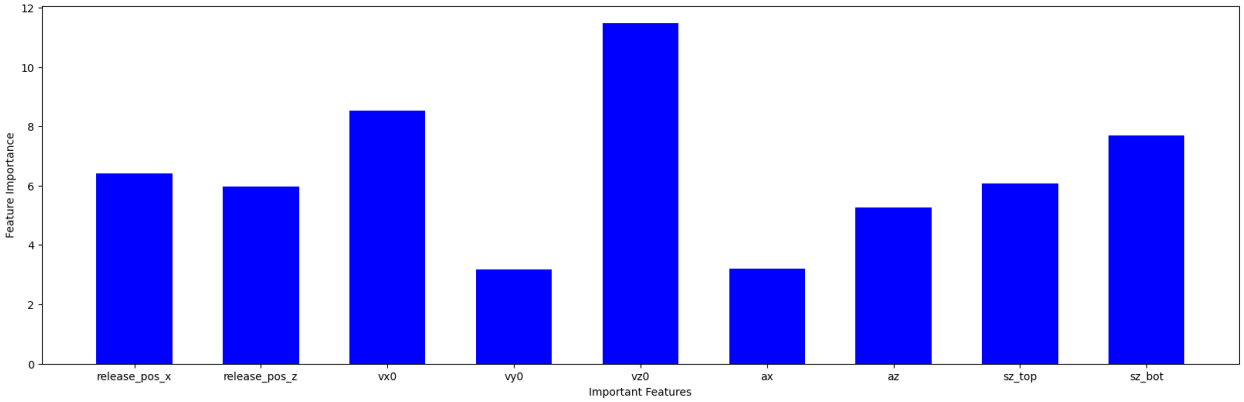
a player with good values for those features from reliever to starter. The opposite could be done with my SP decision tree classifier.

My SP Decision Tree Classifier with Gini for the criterion had 73% accuracy. The top 3 most important features are vz0 (11.5%), vx0 (8.5%), and sz_bot (7.7%) (Feature importance score in parentheses):

	precision	recall	f1-score	support
negative	0.69	0.68	0.68	131371
positive	0.76	0.76	0.76	171431
accuracy			0.73	302802
macro avg	0.72	0.72	0.72	302802
weighted avg	0.73	0.73	0.73	302802

Most important features:
release_pos_x - 6.4%
release_pos_z - 6.0%
vx0 - 8.5%
vy0 - 3.2%
vz0 - 11.5%
ax - 3.2%
az - 5.3%
sz_top - 6.1%
sz_bot - 7.7%





My SP Adaboost classifier with 50 estimators, 1 learning rate and SAMME.R for the algorithm had 65% accuracy:

	precision	recall	f1-score	support
negative	0.64	0.42	0.51	131371
positive	0.65	0.82	0.73	171431
accuracy			0.65	302802
macro avg	0.65	0.62	0.62	302802
weighted avg	0.65	0.65	0.63	302802

Most important features:
release_speed - 4.0%
release_pos_x - 4.0%
balls - 4.0%
strikes - 4.0%
pfx_z - 4.0%
vx0 - 18.0%
vz0 - 20.0%
ax - 4.0%
az - 8.0%
sz_top - 12.0%
sz_bot - 12.0%

My SP Random Forest Classifier with 5 for max depth had 63% accuracy. My RP Decision Tree Classifier with Gini for the criterion had 72% accuracy. The top 3 most important from that this are vz0 (11.1%),

sz_bot (8.6%), and vx0 (8%). My RP Random Forest Classifier with Gini for the criterion had 63% accuracy. My RP Adaboost classifier with 50 estimators, 1 learning rate, and 'SAMME.R" for the algorithm had 60% accuracy. All the top variables in my most accurate model for relievers are the same as the top varoables in my most accurate model for starters, except the starters model has ax (acceleration in the x direction) as an important feature. Therefore, acceleration in the x of a pitch may be more important for starters than relievers.

Fangraphs

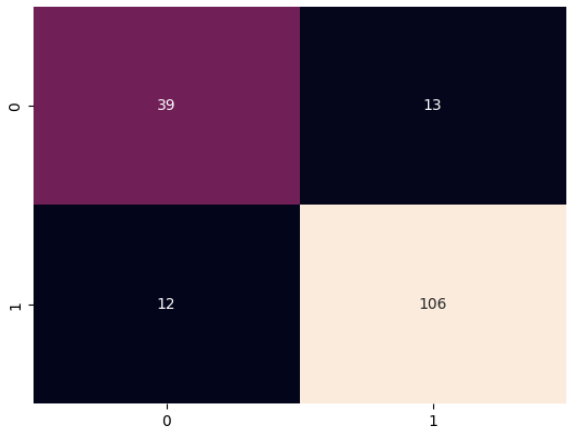
I recoded the WAR values in my SP and RP data frames to good or not good. I chose at least 2.3 WAR for starters to be considered good since I thought a little above the average of 2 would be good. I chose at least 0.4 WAR for relievers to be considered good since I learned 1 WAR was superb for relievers so I thought a little less than half of that would be good. I filtered out players who played less than 10 games. I dropped strings, totals (since good pitchers will pitch more) and many variables where it is pretty obvious that they would result in a high or low WAR.

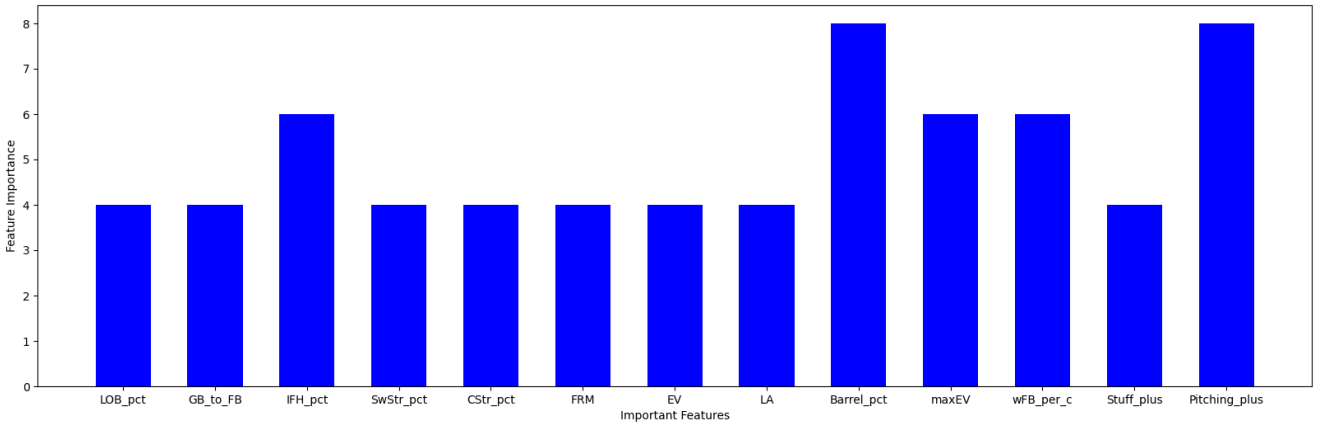
I made classifiers that identified if a pitcher was good or bad based on his stats from the season. This could be used to determine if a player would have been good if he put up those numbers in a different role.

Results from my SP Adaboost classifier with 85% accuracy, 50 estimators, and 1 learning rate are below. The top 5 features are Barrel_pct (8%), Pitching_Plus (8%), wFB_per_c (6%), maxEV (6%); IFH_pct (6%). (Feature importance score in parentheses):

	precision	recall	f1-score	support
good	0.76	0.75	0.76	52
not good	0.89	0.90	0.89	118
accuracy			0.85	170
macro avg	0.83	0.82	0.83	170
weighted avg	0.85	0.85	0.85	170

Most important features:
LOB_pct - 4.0%
GB_to_FB - 4.0%
IFH_pct - 6.0%
SwStr_pct - 4.0%
CStr_pct - 4.0%
FRM - 4.0%
EV - 4.0%
LA - 4.0%
Barrel_pct - 8.0%
maxEV - 6.0%
wFB_per_c - 6.0%
Stuff_plus - 4.0%
Pitching_plus - 8.0%





My SP Random Forest classifier had 85% accuracy with 20 max. depth. The top 3 features are wFB_per_c (6.7%), Pitching_plus (4.8%), and LOB_pct (4.2%). My SP Decision Tree classifier had 77% accuracy with Gini criterion. Combining the top 3 most important variables from the random forest classifier and the top 5 from the Adaboost classifier results in Barrel_pct, Pitching_Plus, wFB_per_c, maxEV, IFH_pct; LOB_pct being the 6 most important variables that determine success or failure for starters. The variables here that were not in the top 6 for relievers are maxEV, IFH_pct; LOB_pct. Therefore, maxEV, IFH_pct; LOB_pct may be more important for starters than relievers.

Results from my RP Adaboost classifier with 79% accuracy, 60 estimators, and 0.2 learning rate are below. The top 3 features are Barrel_pct, wFB_per_c; Contact_pct_sc:

	precision	recall	f1-score	support
good	0.77	0.62	0.68	130
not good	0.80	0.90	0.85	230
accuracy			0.79	360
macro avg	0.79	0.76	0.77	360
weighted avg	0.79	0.79	0.79	360

```
Most important features:
WP - 3.3%
RS_per_9 - 6.7%
LOB_pct - 3.3%
OSwing_pct - 3.3%
SwStr_pct - 5.0%
CStr_pct - 3.3%
FRM - 5.0%
LD_pct_plus - 3.3%
Pull_pct_plus - 3.3%
Barrel_pct - 11.7%
maxEV - 3.3%
wFB_per_c - 10.0%
Contact_pct_sc - 10.0%
Pitching_plus - 6.7%
```

My RP random forest classifier had 78% accuracy and a max depth of 4. The top 3 features are botOvr (7.9%), Pitching_plus (6.9%), and wFB_per_c (6.8%). My RP decision tree classifier had 74% accuracy. The top 3 features are botOvr (18%), Contact_pct (7.5%); Barrel_pct (9.2%). Combining the top 3 from each model results in botOvr, Contact_pct, Barrel_pct, Pitching_plus, wFB_per_c, Contact_pct_sc being the 6 most important variables that determine success or failure for relievers. The variables here that were not in the top 7 for starters are botOvr, Contact_pct; Contact_pc_sc.

The Switch

I sent potential candidates to my SP Adaboost model with 85% accuracy. This model had the highest accuracy out of all my models, so I decided to use this one. I got relievers who were not too old since they may be unwilling to change roles at an old age, relievers who had a high barrel% since that was the most important feature, and relievers who played at least 30 games in 2023. I used 2023 since that was the most recent season. I sent the 10 relievers from this group who had the best Ptching_plus, since that was tied for the most important feature for this model. The model classified Tanner Scott as a good starter in 2023. This likely comes from his 111 Pitching Plus, 3.4% barrel pct, and 1.4 wFB per c last season. I think him being a starter could be good since he is still in his 20's so he is less likely to get injured from the change. The model classified Tyler Rogers as a good starter in 2023. This likely comes from his 112 Pitching Plus last season and 2.2% Barrel%. I think the Giants should try this since they were in the bottom 10 in the league in SP WAR last year and Rogers could help them with that. This likely has to do with his 6.1% barrel%.

I sent all the Reds relievers who pitched at least 30 games in 2023 and who were not too old to the model. The model classified Ian Gibaut as a good starter in 2023. This is likely from his 107 Pitching_plus, 6% Barrel%, and 1.3 wFB per c. The Reds should try to use him as a starter because their SP WAR last year was also in the bottom 10 last year and he could help them.

Sources:

https://www.fangraphs.com/leaders/major-league?pos=all&lg=all&qual=y&type=8&season=2023&month=0&season1=2023&ind=0&stats=sta&team=0%2Cts&sortcol=20&sortdir=default&pagenum=1

https://library.fangraphs.com/misc/war/