Can NBA Draft Combine Measurements be Used to Predict Players' Box/Plus Minus?

A MACHINE LEARNING APPROACH

Predicting BPM with Anthropometric and Strength Agility Measurements

- The NBA Draft Combine is an annual chance for the league's scouts and decision-makers to collect critical intel about players.
- Box Plus/Minus (BPM) estimates a player’s contribution in points above league average per 100 possessions.

NBA draft combine measurements from 2000-2019 are downloaded from nbaathlete.com.

Player performances from the 2000-2001 season to the 2019-2020 season are web scraped from basketball-reference.com.

The two datasets are joined based on the player-name field.

The combined data set is cleaned.

The preprocessed data includes data on 543 players.

Three Regression Strategies

01

1. One model per position: Lasso regression was performed as variable-selection method. Non-zero coefficients were fed into a regular multiple linear regression model.

02

Separate models, one for each position: Ridge regression was performed

03

One model for all: multiple machine learning regression algorithms were performed.

At least a sixth men?

01

The minority class is up-sampled to the same level as the majority class.

02

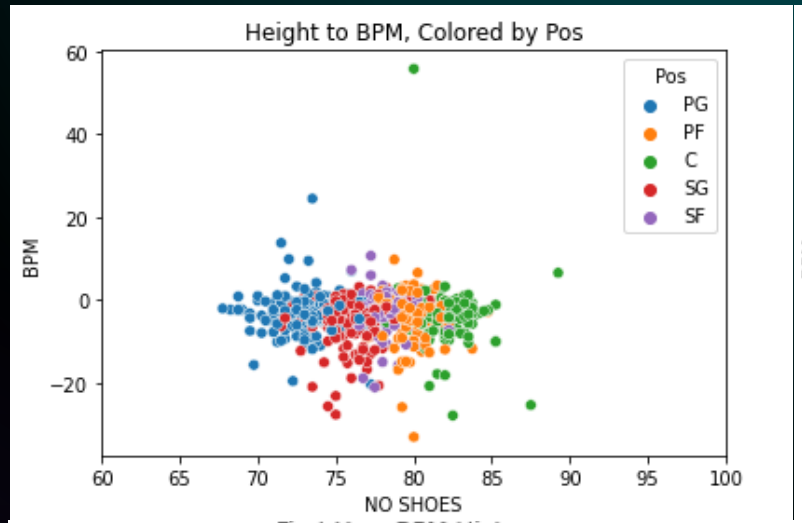
Multiple classification algorithms were performed on the binary target.

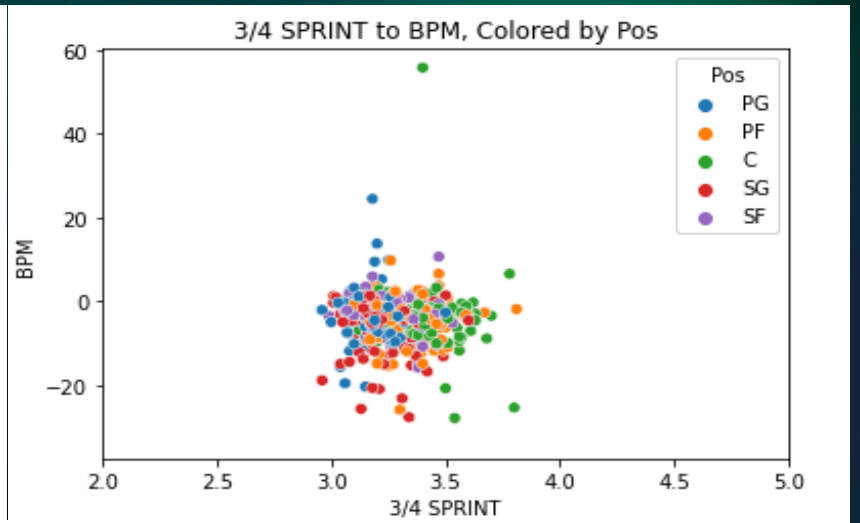
03

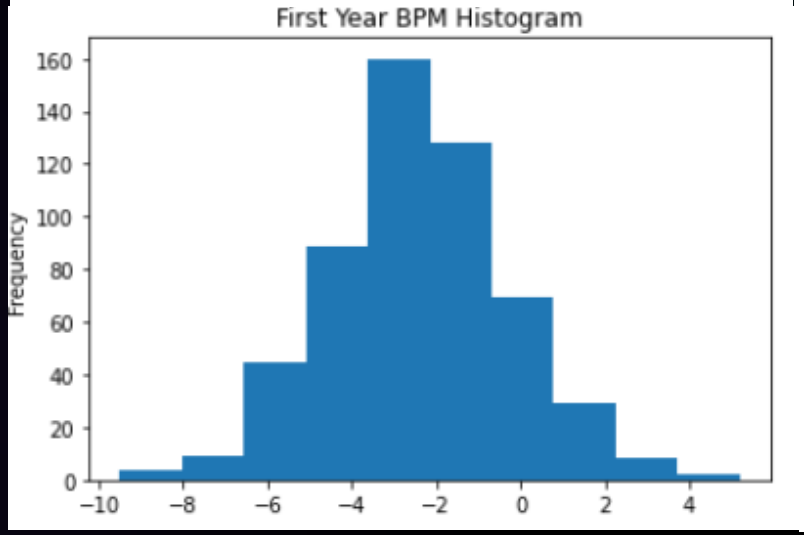
Shapley scores were obtained from the best model to see which variables have the biggest influence on the target.

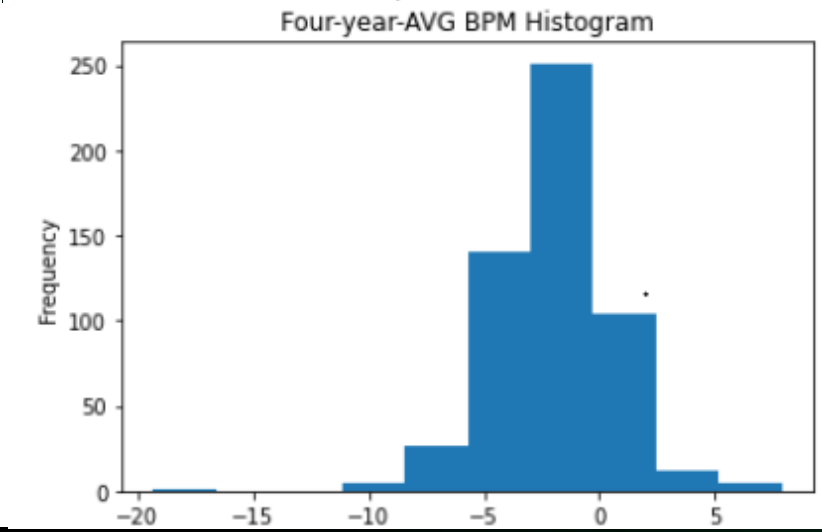
No anthropometric measurements are linearly correlated with BPM

BPM	Avg_BPM
-0.007803	0.027175
-0.000646	0.035829
0.013904	0.014807
-0.005542	0.046498
0.033978	0.048577
-0.023454	-0.031933
-0.011898	-0.023983
-0.015006	-0.029097
-0.018192	-0.033933
0.015353	-0.011755
0.025157	-0.001800
-0.035951	-0.018980
-0.041762	-0.022603
-0.010835	0.006380



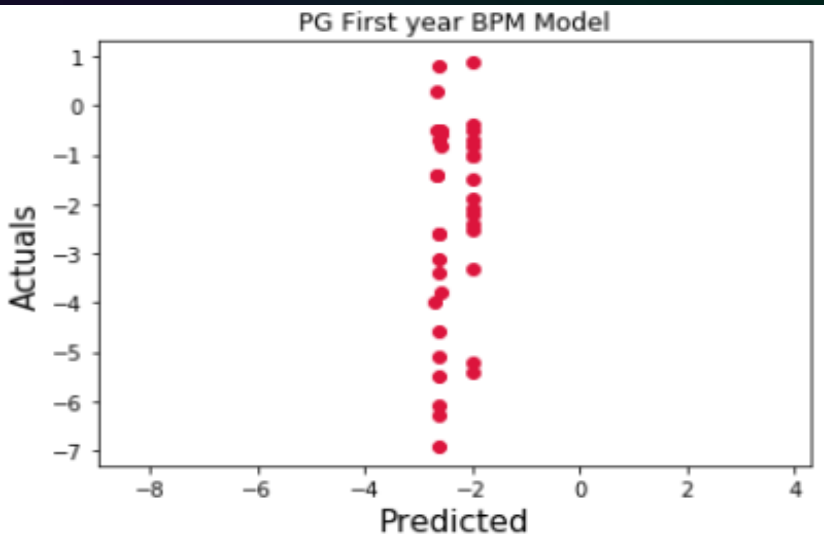


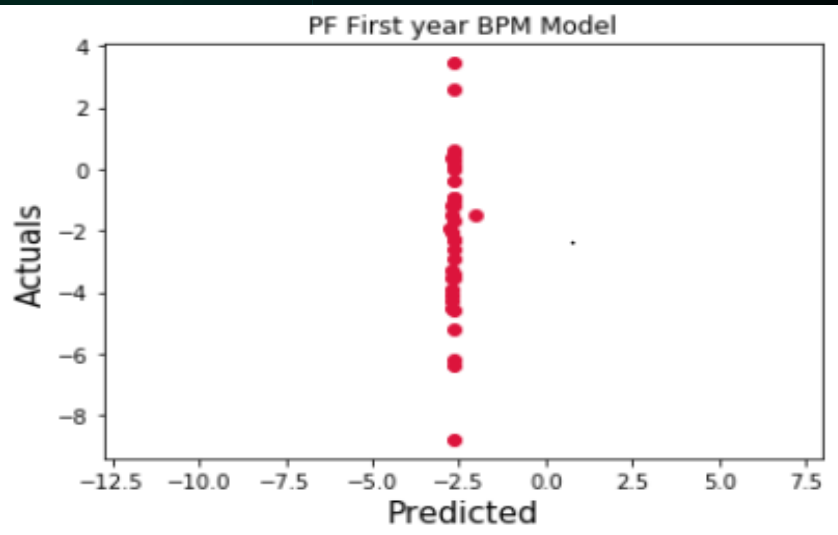


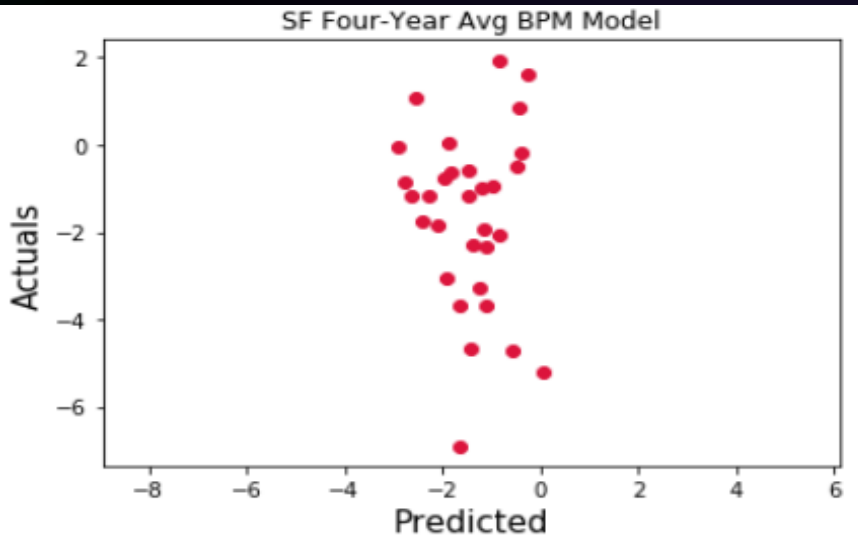


Results: Lasso and multi-Linear Regression

- RMSE range from 1.9 to 2.5 for the individual position's first-year BPM models
- RMSE range from 2.0 to 2.7 for the individual position's four-year-average BPM models.
- It is more difficult to predict four-year-average BPM compared to first-year BPM.
- Model consistently predicts negative values.
- "Best" model always predicts around -2.5.

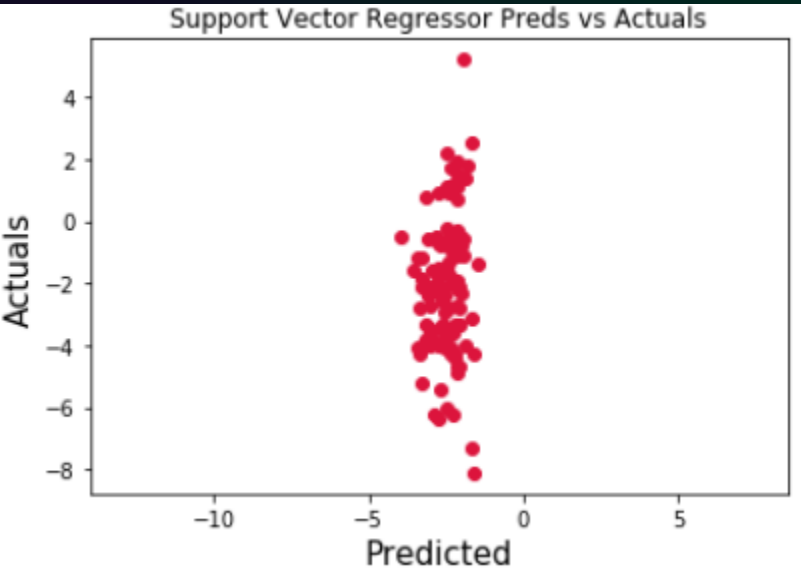


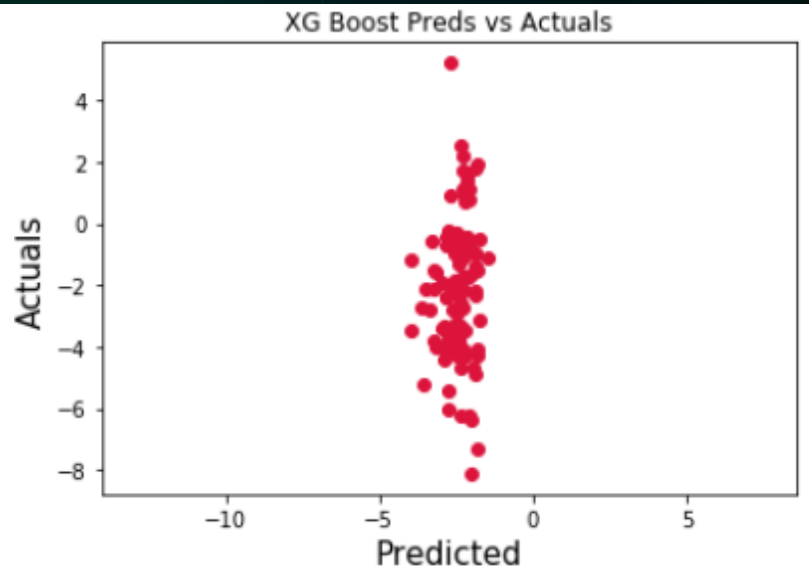


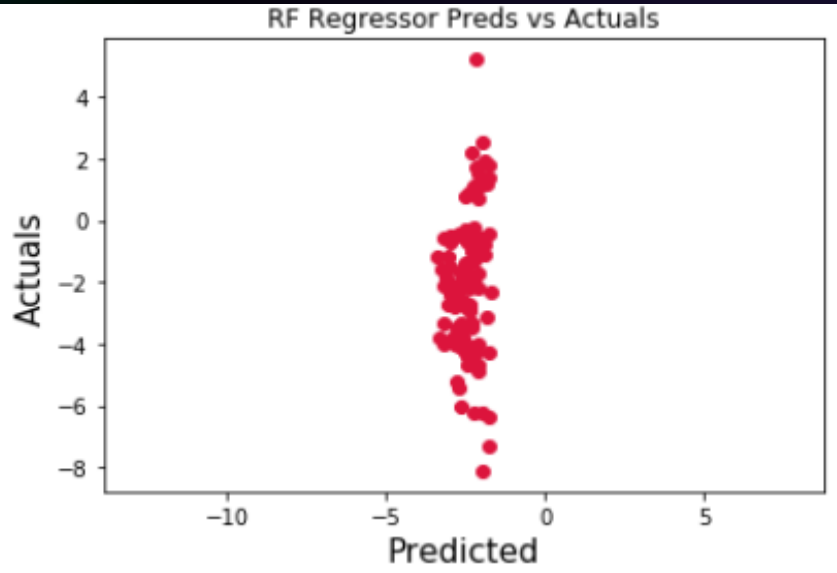


Other regression algorithms not so good either

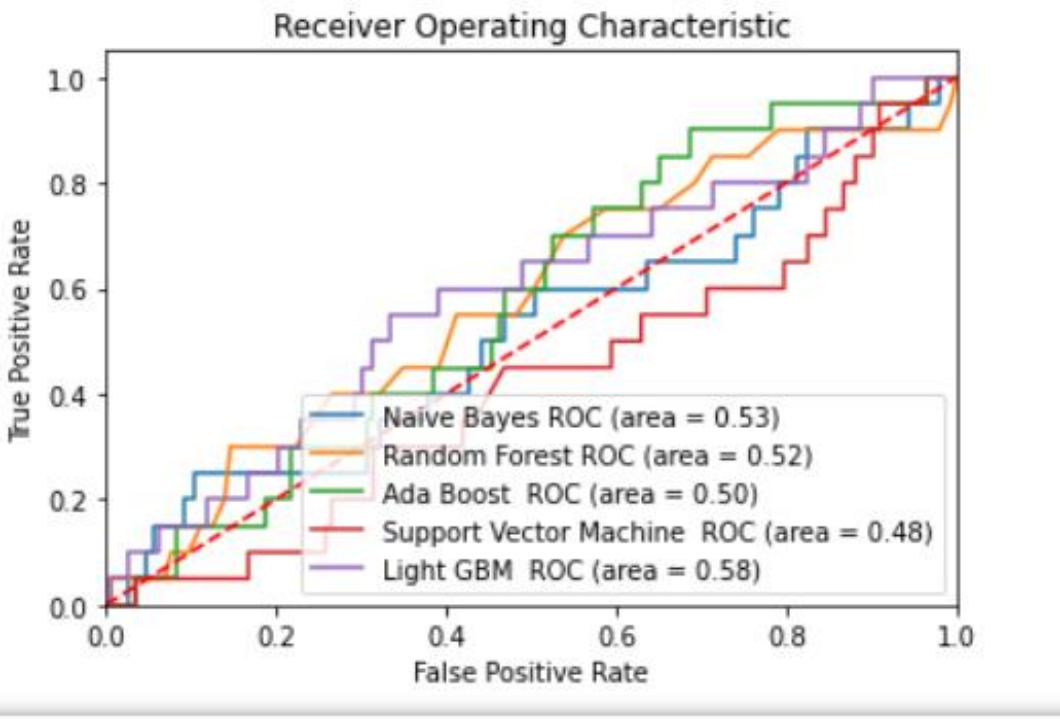
- RMSE from machine learning regression algorithms are similar to, if not worse, just using Lasso and multi-linear regression. They are between 2 and 3. (This could be the difference between bench player and starters).
- Best model is support vector machine with RMSE 2.13.
- Like previously, models still consistently predicts negative values.





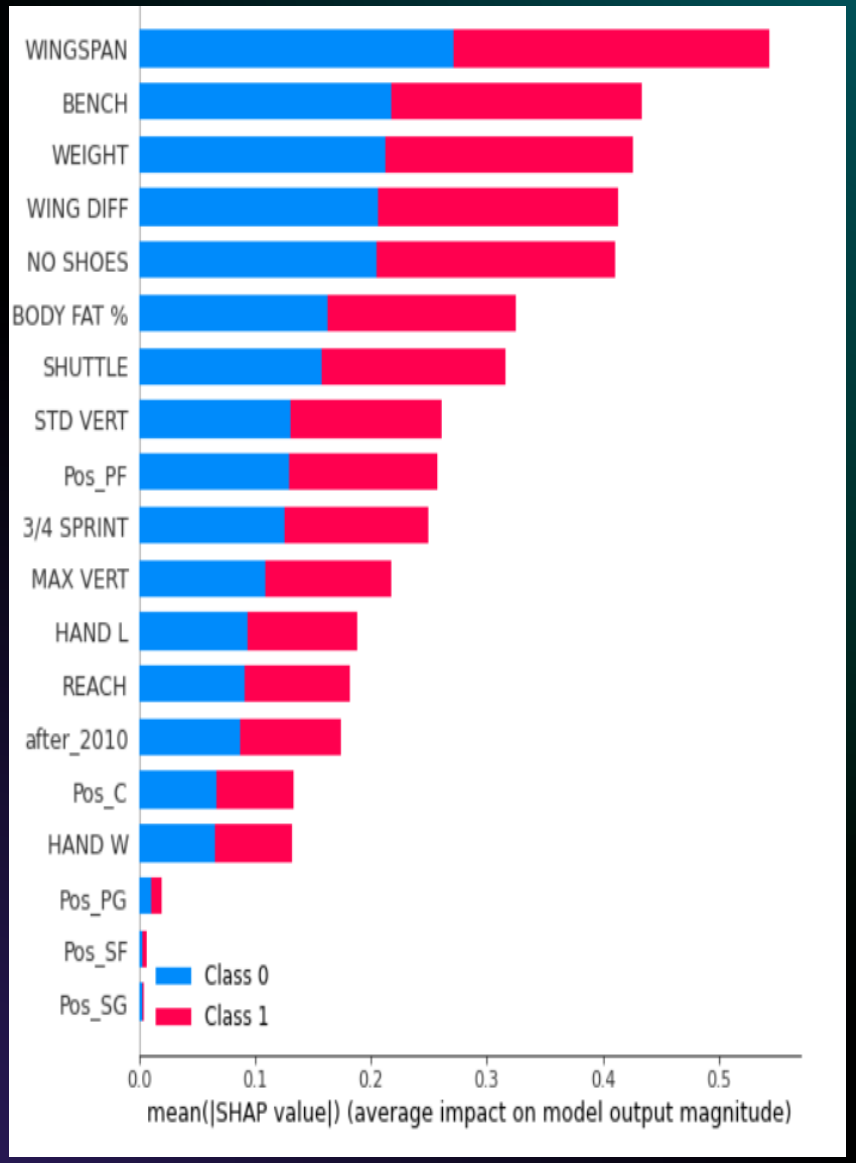


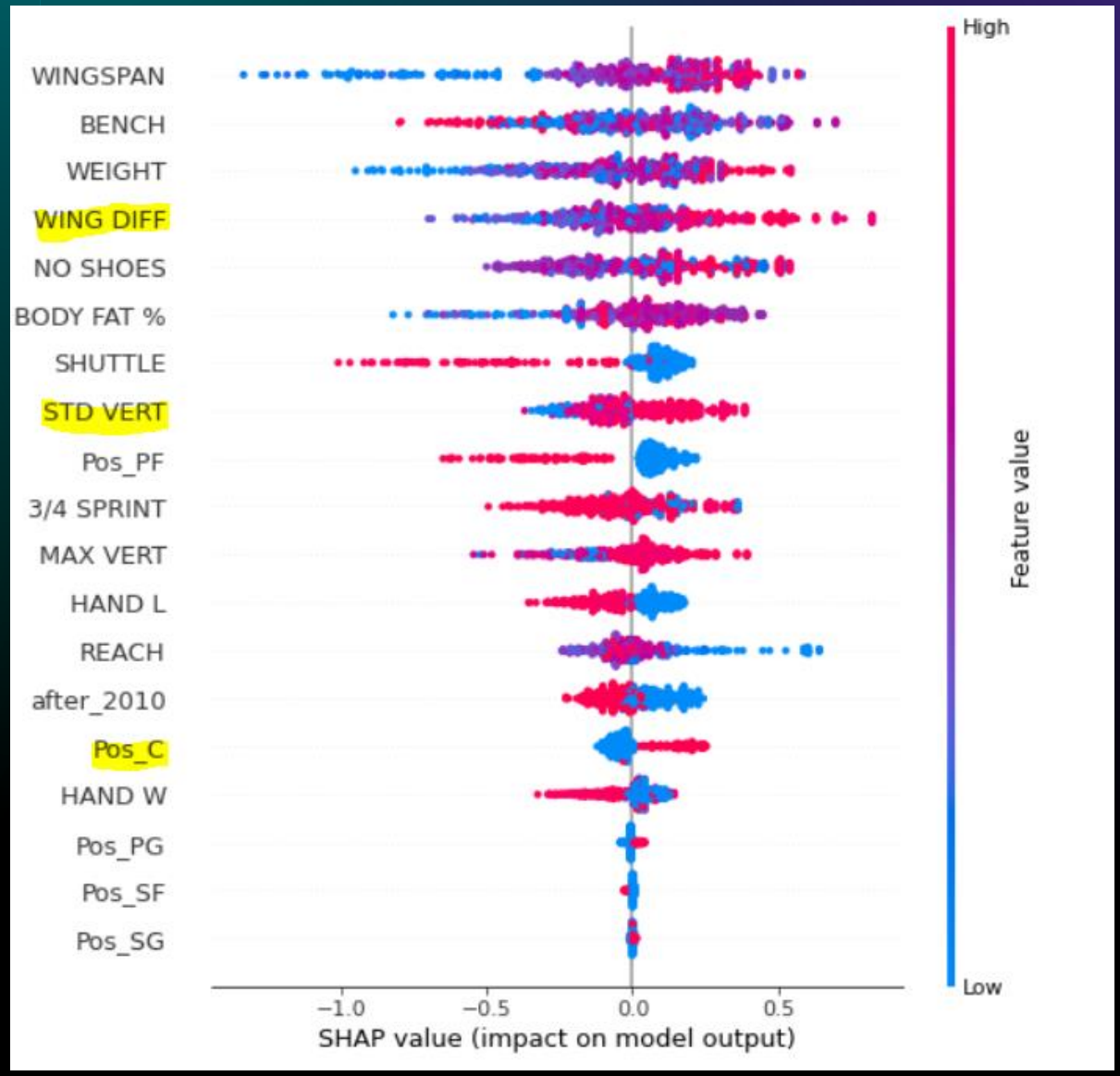
Classification results are as good as flipping a coin.



	precision	recall	f1-score	support
0	0.90	0.66	0.77	143
1	0.17	0.50	0.26	20
accuracy			0.64	163
macro avg	0.54	0.58	0.51	163
weighted avg	0.81	0.64	0.70	163

Results: SHAP Scores from LightGBM





Conclusion

- **Anthropometric and strength/agility measurement alone cannot be used to robustly predict BPM.**
- Four-average BPM is more difficult to predict than first-year BPM.
- Future studies could include predictors such as combine scrimmage game statistics and combine shooting drill statistics.

Is Berkson's Paradox the Culprit?

" If you're invited to the combine, you have some minimum combination of measurable athletic ability and non-athletic basketball ability. The lack of relationship is more to do with this bias of the population of who is invited to the combine than the ability of the players themselves." ---- *Paul Sabin, Data Scientist @ESPN*