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

A MACHINE LEARNING APPROACH

Agenda

- Introduction
- Methods
- EDA
- Results
- Conclusions

Introduction

- The NBA Draft Combine is an annual chance for the league's scouts and decision-makers to collect critical intel about players. It includes physical measurements, agility drills, 'pro day' workouts, scrimmages and interviews. (This study focuses only on anthropometric and strength and agility measurements)
- Box Plus/Minus (BPM), according to Basketball Reference, is a basketball box score-based metric that estimates a player’s contribution in points above league average per 100 possessions played based on box score information, position, and the team’s overall performance.
- **This study investigates the potential of using anthropometric and strength/agility measurements from draft combine to predict a player’s first-year and four-year average BPM using statistical modelling and modern machine learning methods.**

Methods: Data Processing

- 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.
- He combined data set is cleaned. Missing values are set to -1 and players that played less than 250 minutes are excluded from the data set.
- The resulting data set includes data on 543 players.

Methods: Modeling (Regression)

The following are performed both for first-year BPM and four-year average BPM as the target:

1. Separate models, one for each position: Lasso regression (grid search CV for best alpha) was performed as variable-selection method. The variables with non-zero coefficients were fed into a regular multiple linear regression model.

2. Separate models, one for each position: Ridge regression (grid search CV for best alpha) was performed

3. One model for all: multiple machine learning regression algorithms (decision tree, random forest, XG boost, support vector regressor, K-Nearst-Neighbor regressor) were performed. Multiple hyper parameters were tested using grid-search cross validation.

Methods: Modeling (Classification)

The following are performed both for first-year BPM and four-year average BPM as the target:

BPM is transformed into a binary variable with greater or equal to zero as the cutoff. According to definition of BPM on basketball-reference.com, a BPM of zero would indicate the player is at least a six-men type of player. Thus, the classification predicts whether a player is at least a sixth men.

Because most of the players have negative BPMs, the minority class is up-sampled to the same level as the majority class.

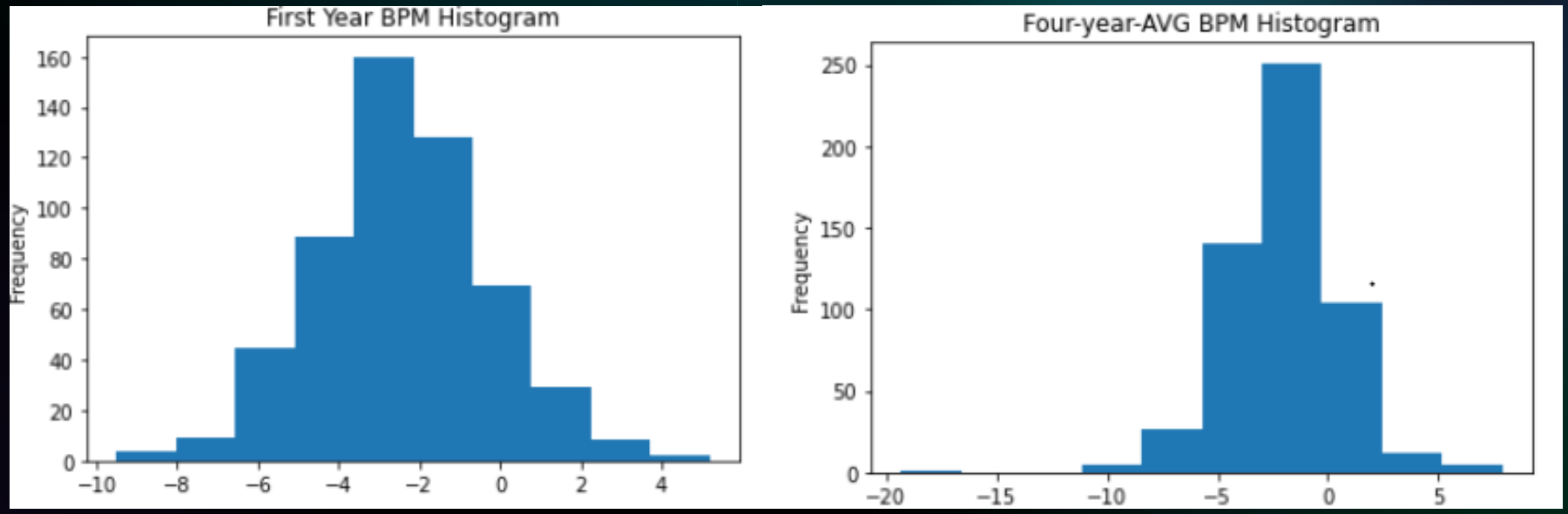
Multiple classification algorithms (naïve bayes, random forest, ada boost, support vector machine, light GBM) were performed on the binary target.

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

EDA

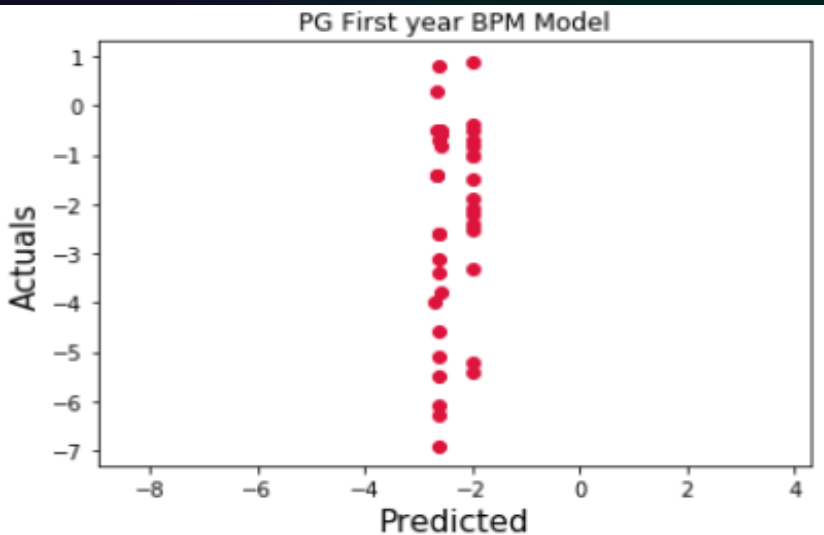
BPM	Avg_BPM
-0.007803	0.027175
-0.000646	0.035829
0.013904	0.014807
-0.005542	0.046498
0.033978	0.048577
-0.023454	-0.031933
-0.011898	-0.023983
-0.015006	-0.029097
-0.018192	-0.033933
0.015353	-0.011755
0.025157	-0.001800
-0.035951	-0.018980
-0.041762	-0.022603
-0.010835	0.006380

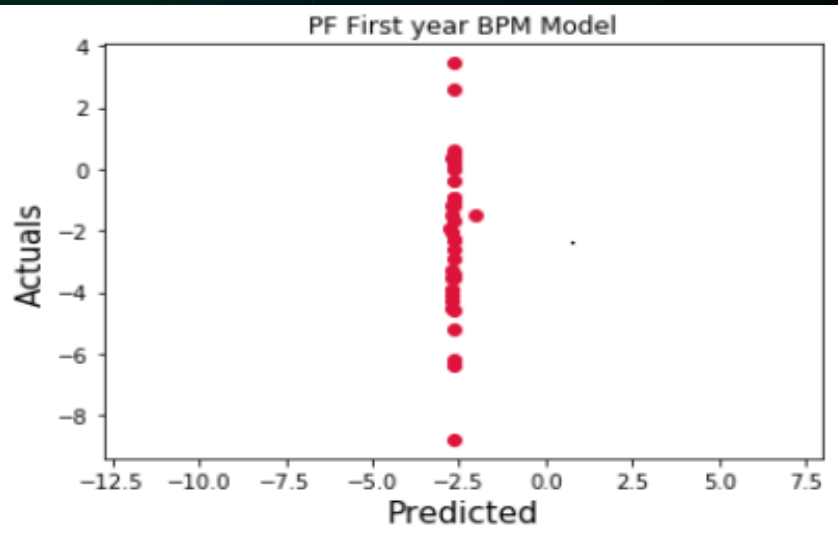
- NO anthropometric and strength/agility measurements are linearly correlated with BPM
- Most first year BPM values are negative.
- BPM values are approximately normally distributed

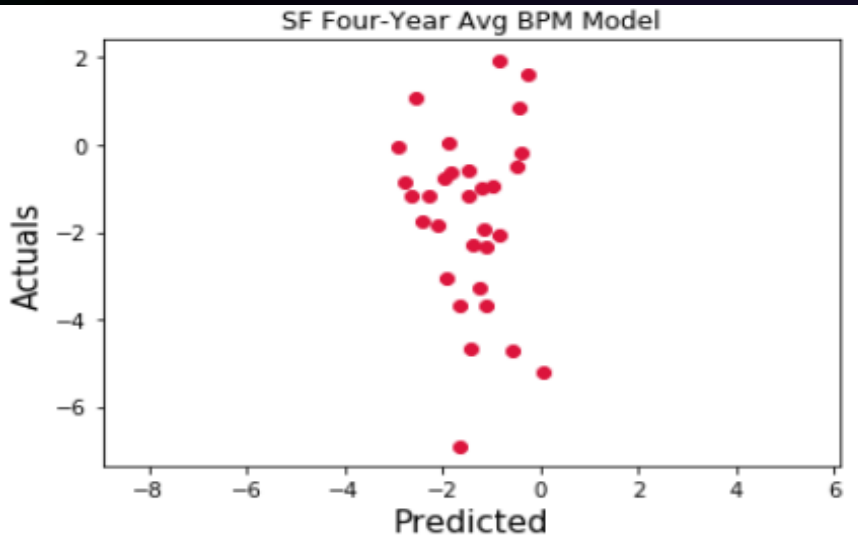


Results: Lasso and Multi-Linear Regression

- 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.

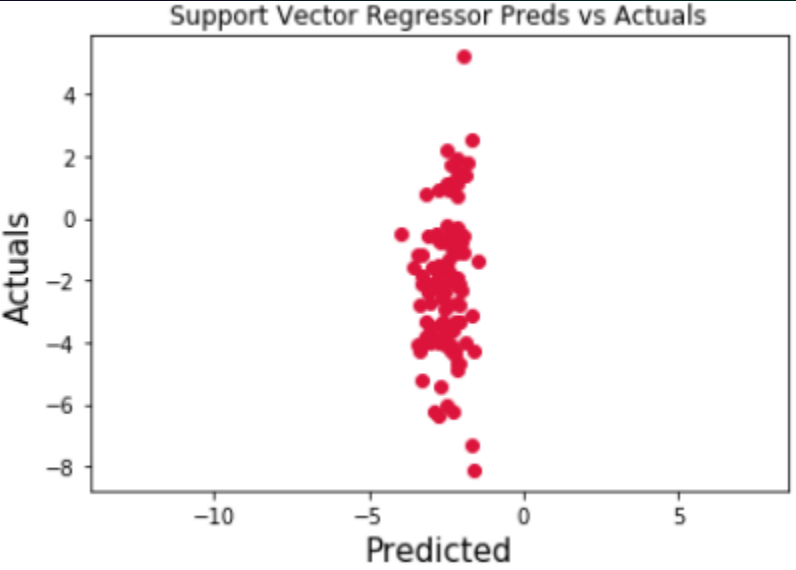


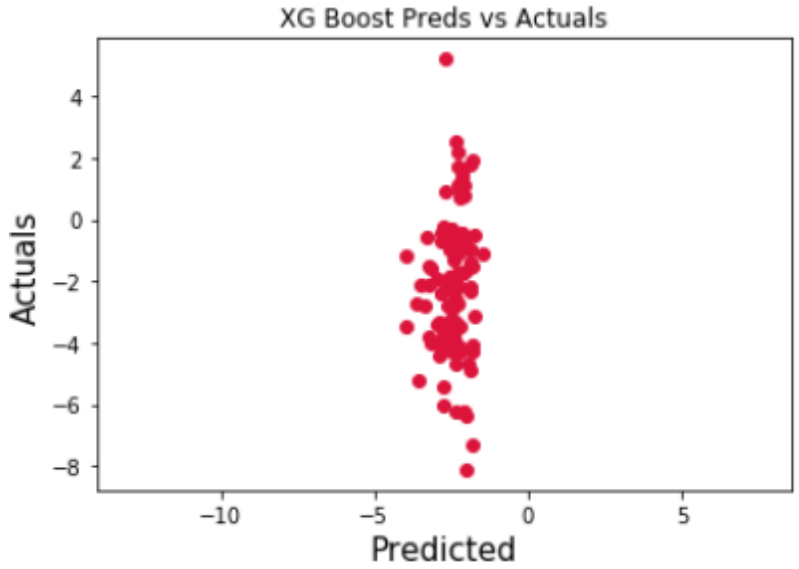


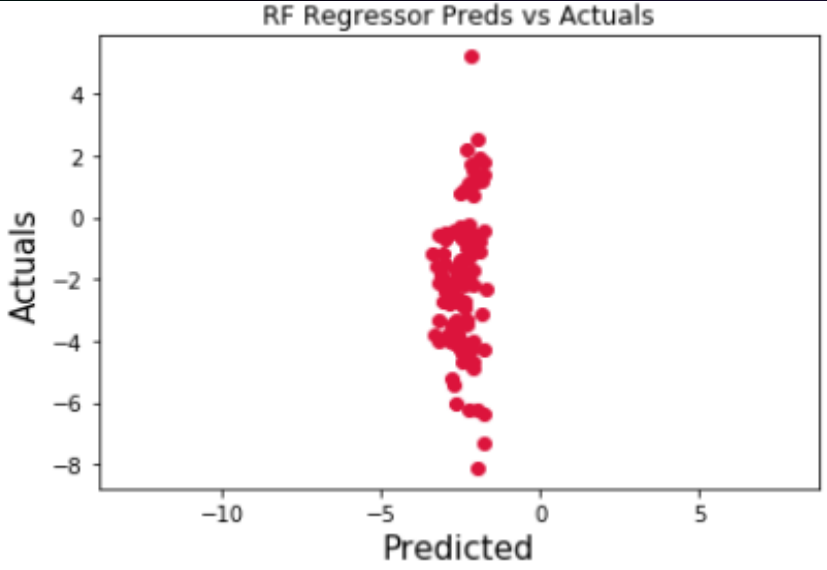


Results: Machine Learning Regression

- 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.

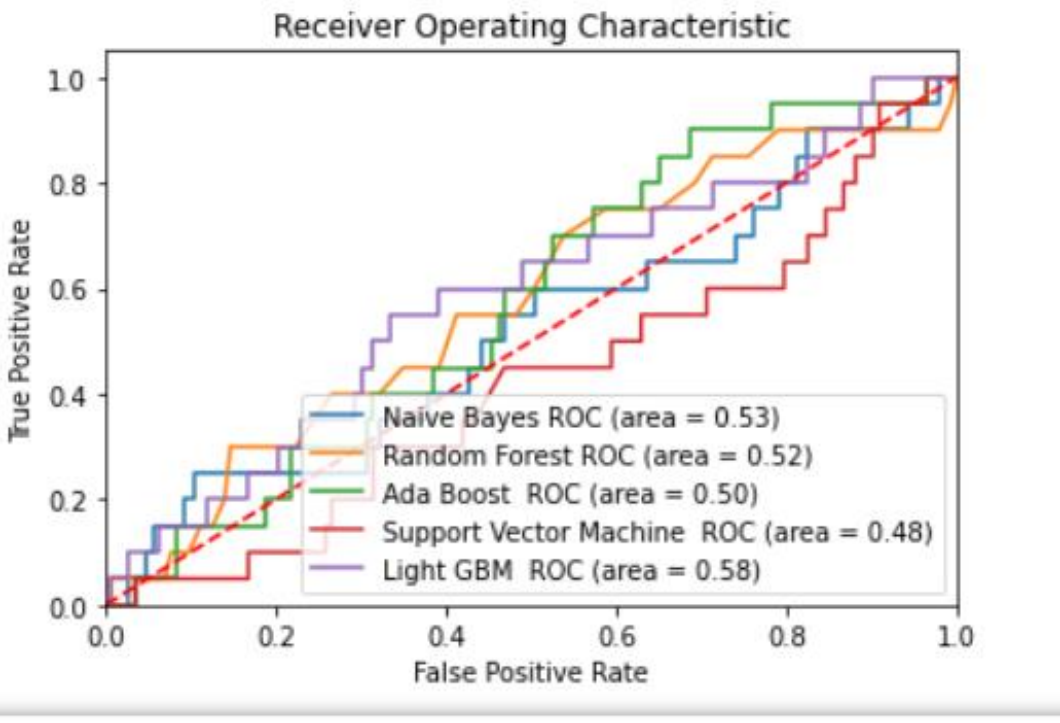






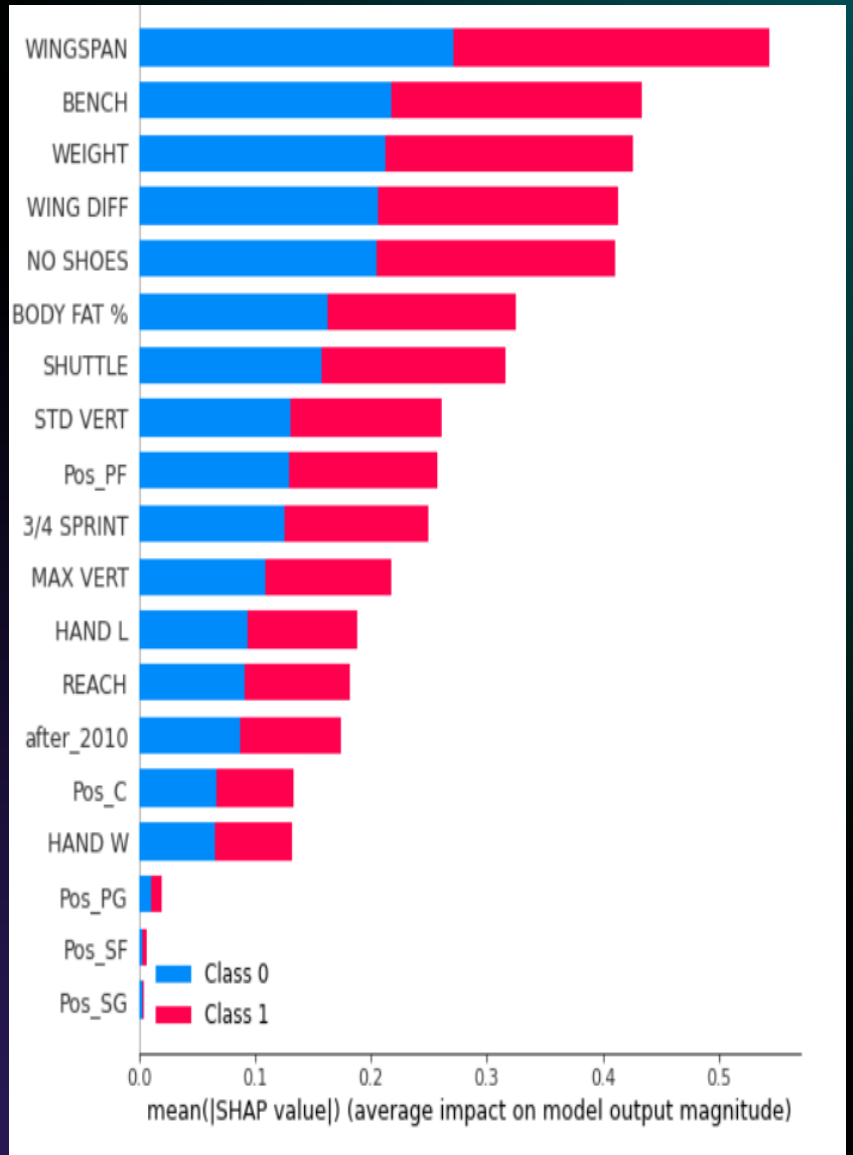
Results: Classifying a BPM of at least Zero

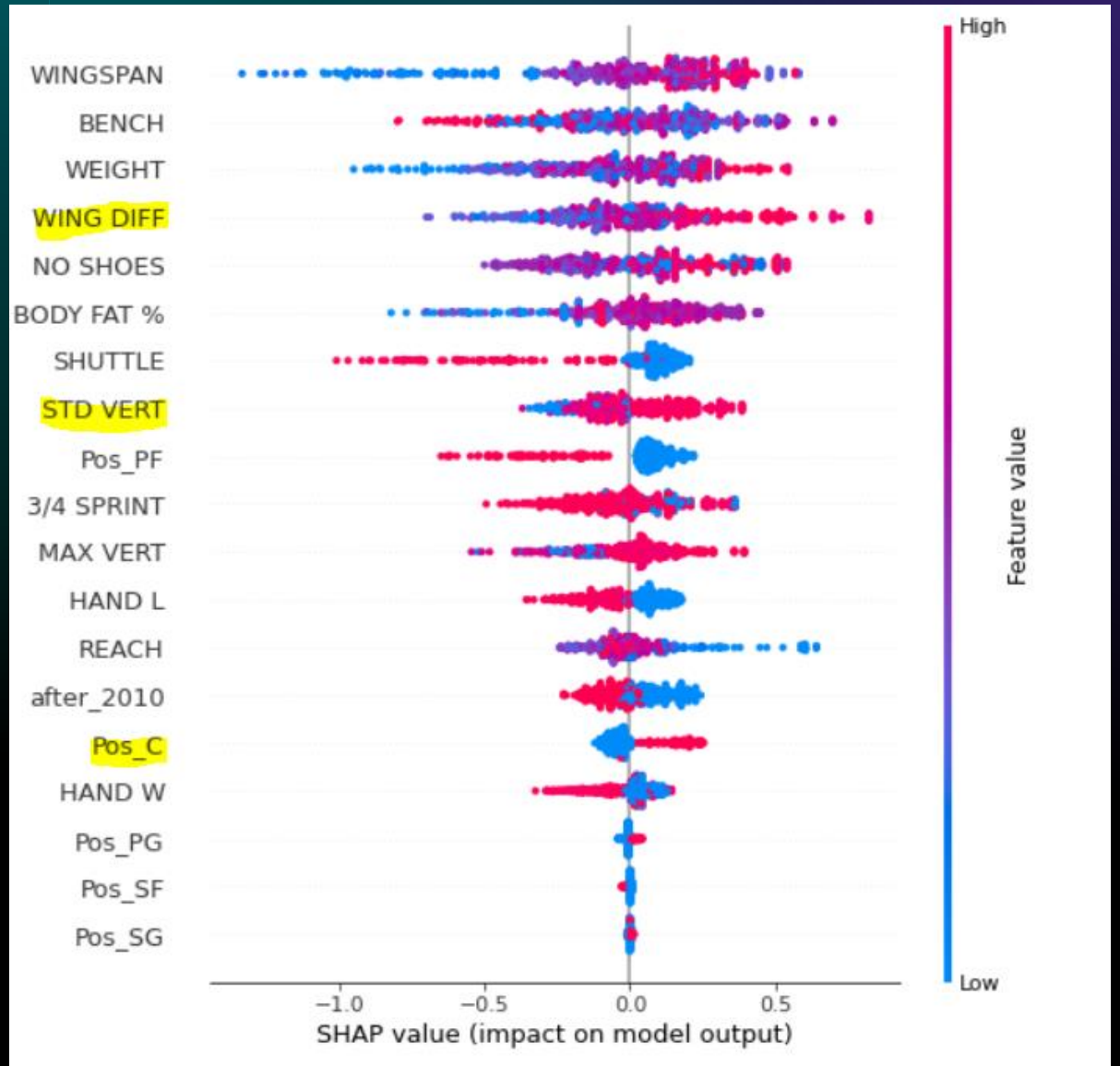
- Classification results are slightly better than just flipping a coin.



	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.