Predicting Future NBA Success

For General Managers and Head Coaches

By Satvik Reddy

Project Overview

The best NBA players reach the distinction of making an All-Star, All-NBA, or All-Defense team during their careers

Once these honors are achieved, player values shoot up, and the competition to gain those players’ services increases for general managers

By predicting which players have the potential to achieve an honor after only their rookie season, head coaches will be able to focus development efforts better, and general managers will be able to focus their resources on the correct players.

Data Information

NBA player statistics for every season dating back to 1950 was made available on Kaggle by Omri Goldstein.

Dataset contains basic statistics like points, rebounds, and assists, and advanced statistics like WS (win shares) and VORP (Value over replacement player).

All-Star, All-NBA, and All-Defense data was obtained from RealGm.com.

Data Wrangling

Removed data from seasons before 1985 because that is how long the modern versions of All-Star, All-NBA, and All-Defense teams have existed

Removed all seasons from players except for their rookie seasons.

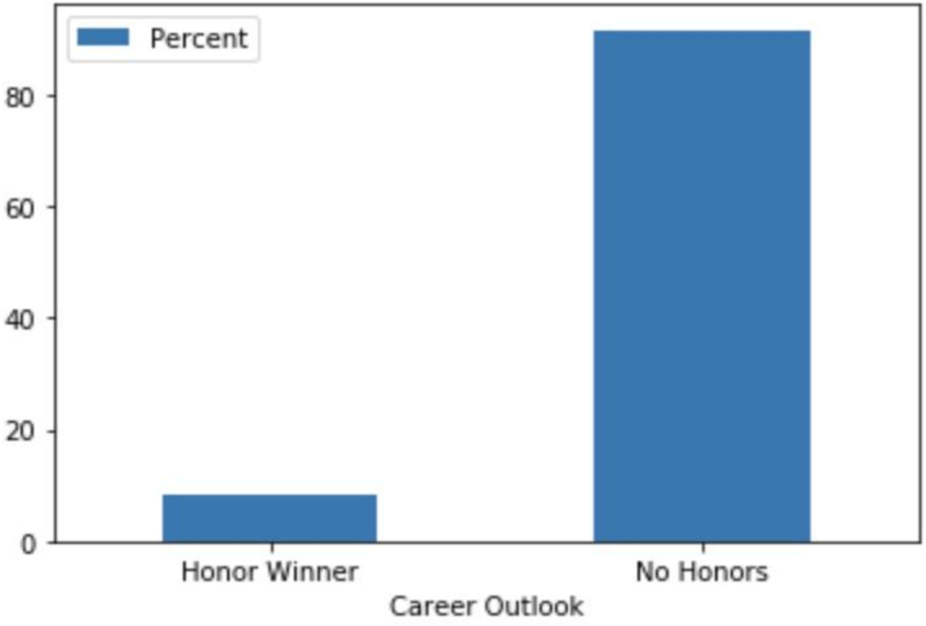
Removed features that had information from the draft; I did not want model to be biased on when a player got drafted.

Filled in missing data with statistics from Basketball Reference

Exploratory Data Analysis

Percentage of NBA players that are “Honor Winners” (All-Star, All-NBA, All-Defense), and percentage that win “No Honors”.

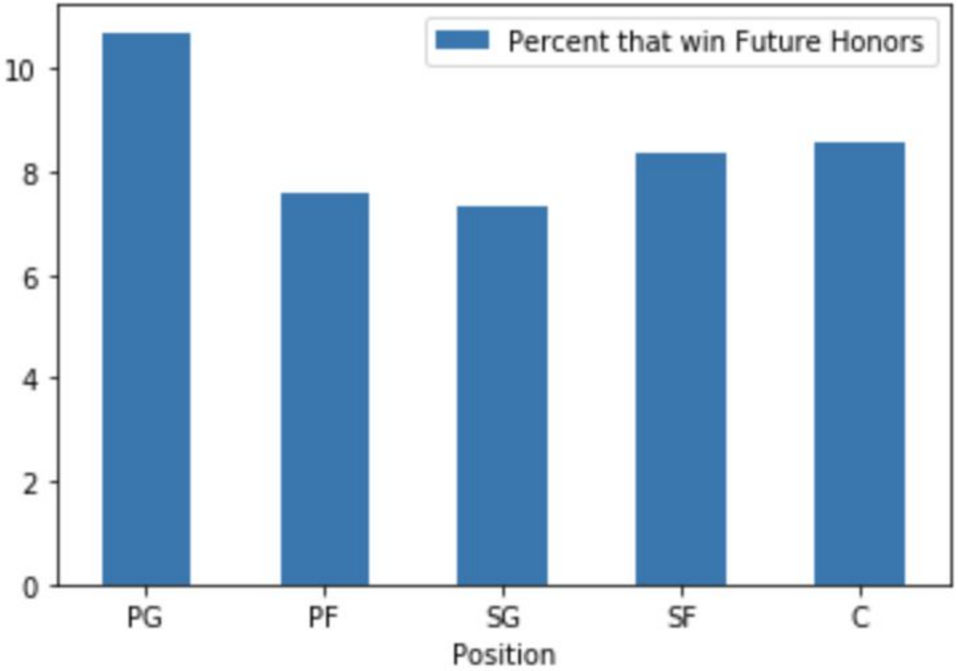
Less than 10% of players win a honor; this highlights the advantage teams can gain when they can predict accurately which players will win one.



Exploratory Data Analysis

Percentage of players at each position that have won honors in their careers

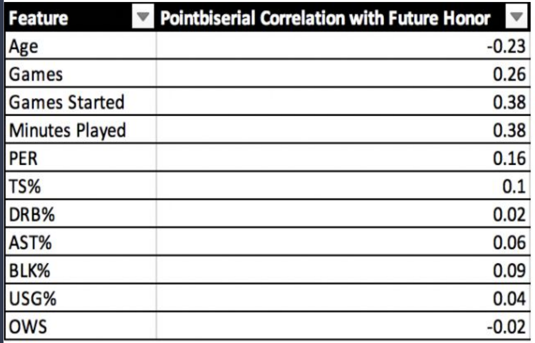
Point Guards and Centers have a higher than average likelihood of achieving an honor.



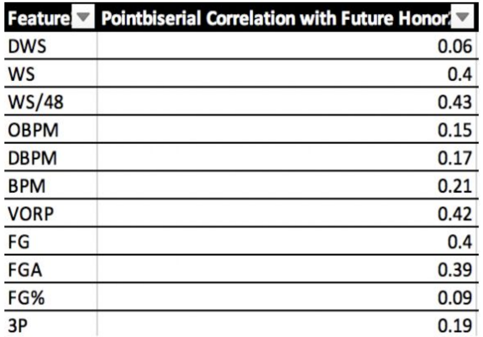
Exploratory Data Analysis

Significant (p < .05) Point Biserial Correlations between each feature and winning a “Future Honor”.

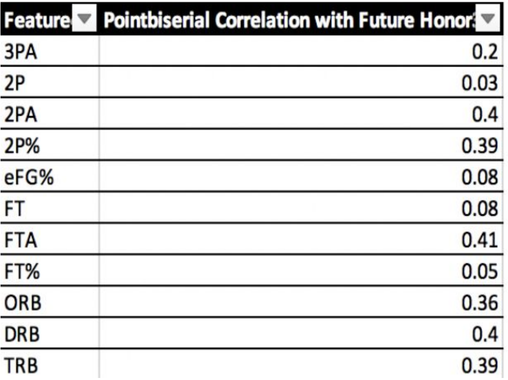
Advanced Stats (WS, WS/48, VORP) and Basic Stats (Pts, FGs) are among the highest correlations.



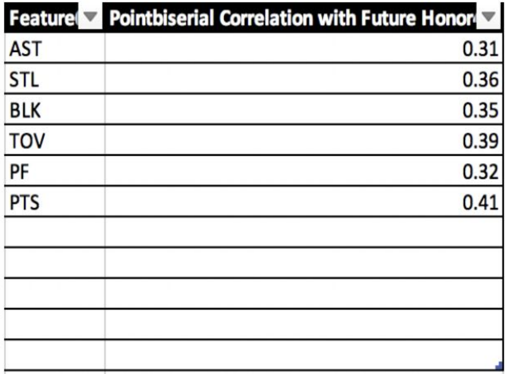
Feature	Pointbiserial Correlation with Future Honor
Age	-0.23
Games	0.26
Games Started	0.38
Minutes Played	0.38
PER	0.16
TS%	0.1
DRB%	0.02
AST%	0.06
BLK%	0.09
USG%	0.04
OWS	-0.02



Feature	Pointbiserial Correlation with Future Honor
DWS	0.06
WS	0.4
WS/48	0.43
OBPM	0.15
DBPM	0.17
BPM	0.21
VORP	0.42
FG	0.4
FGA	0.39
FG%	0.09
3P	0.19



Feature	Pointbiserial Correlation with Future Honor
3PA	0.2
2P	0.03
2PA	0.4
2P%	0.39
eFG%	0.08
FT	0.08
FTA	0.41
FT%	0.05
ORB	0.36
DRB	0.4
TRB	0.39

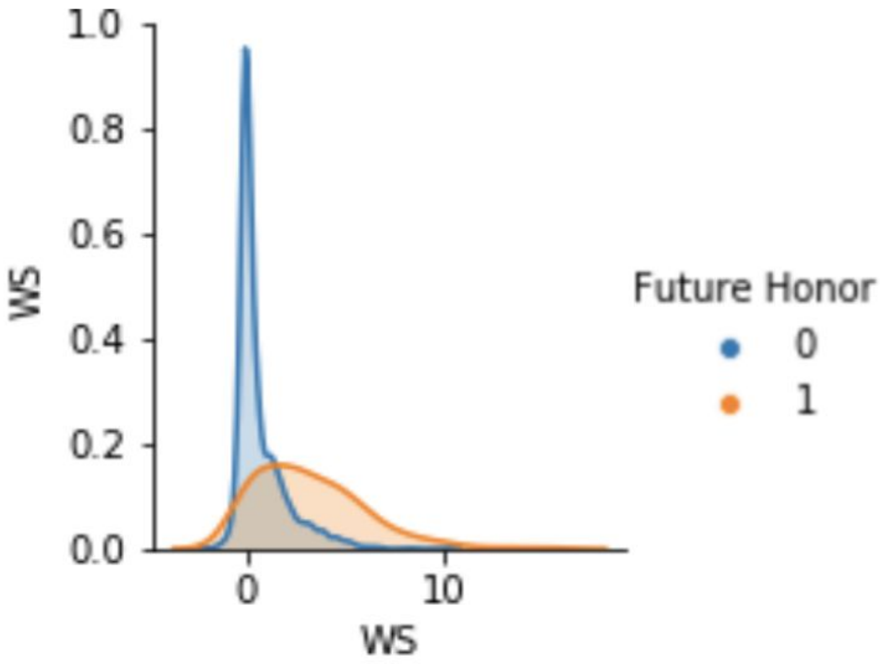


Feature	Pointbiserial Correlation with Future Honor
AST	0.31
STL	0.36
BLK	0.35
TOV	0.39
PF	0.32
PTS	0.41

Exploratory Data Analysis

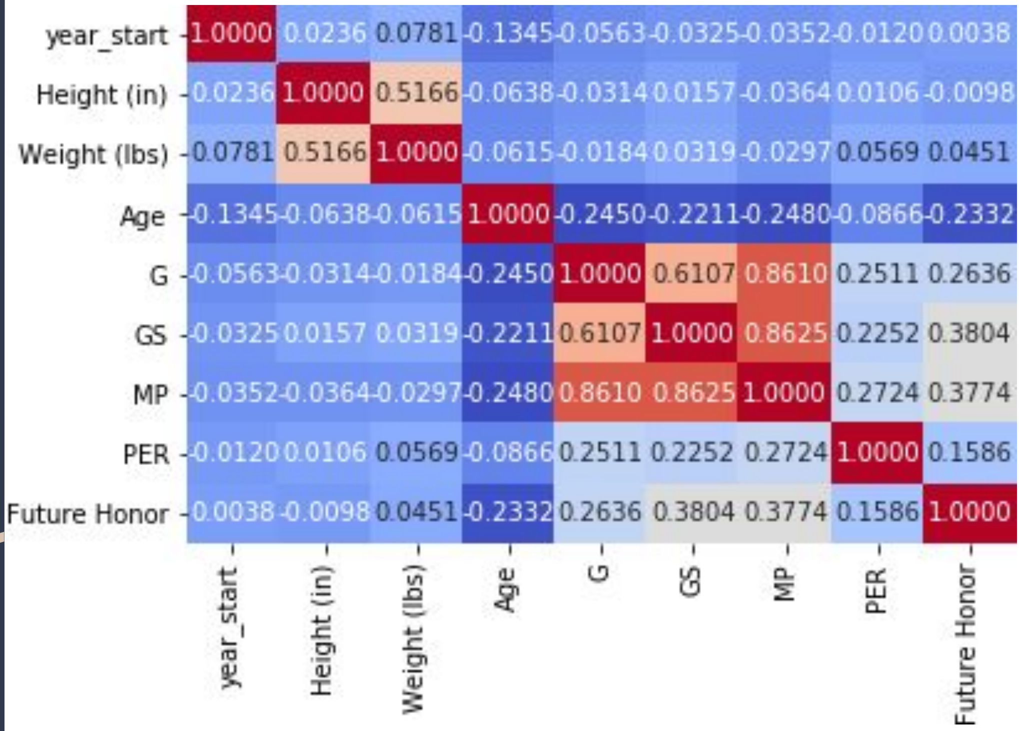
Pairplot between WS and Future Honor

Players who win honors have a WS curve wider and right-shifted compared to non-honor winners



Exploratory Data Analysis

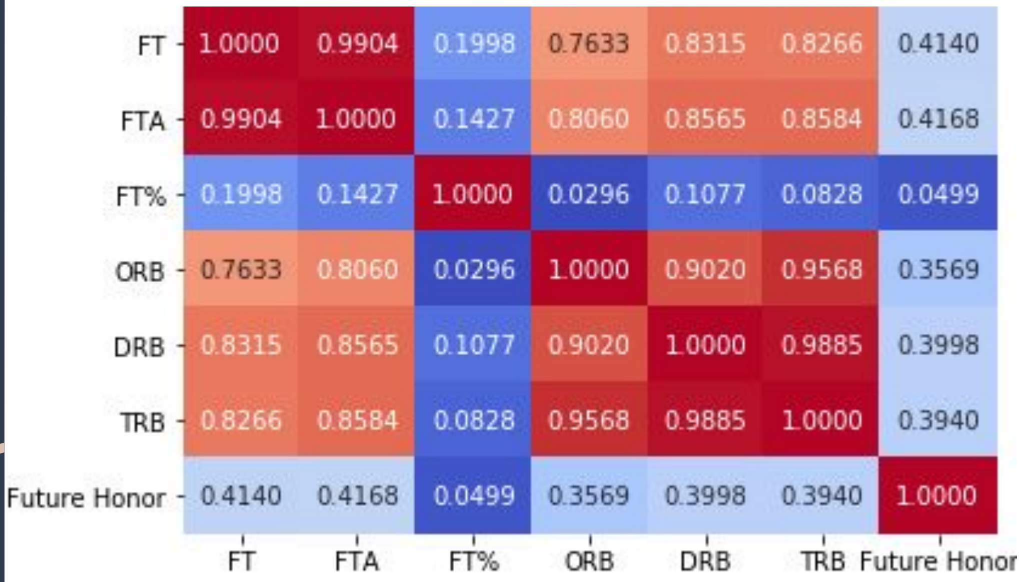
Correlation Matrices of Player Statistics to investigate Collinearity



	year_start	Height (in)	Weight (lbs)	Age	G	GS	MP	PER	Future Honor
year_start	1.0000	0.0236	0.0781	-0.1345	-0.0563	-0.0325	-0.0352	-0.0120	0.0038
Height (in)	0.0236	1.0000	0.5166	-0.0638	-0.0314	0.0157	-0.0364	0.0106	-0.0098
Weight (lbs)	0.0781	0.5166	1.0000	-0.0615	-0.0184	0.0319	-0.0297	0.0569	0.0451
Age	-0.1345	-0.0638	-0.0615	1.0000	-0.2450	-0.2211	-0.2480	-0.0866	-0.2332
G	-0.0563	-0.0314	-0.0184	-0.2450	1.0000	0.6107	0.8610	0.2511	0.2636
GS	-0.0325	0.0157	0.0319	-0.2211	0.6107	1.0000	0.8625	0.2252	0.3804
MP	-0.0352	-0.0364	-0.0297	-0.2480	0.8610	0.8625	1.0000	0.2724	0.3774
PER	-0.0120	0.0106	0.0569	-0.0866	0.2511	0.2252	0.2724	1.0000	0.1586
Future Honor	0.0038	-0.0098	0.0451	-0.2332	0.2636	0.3804	0.3774	0.1586	1.0000

Exploratory Data Analysis

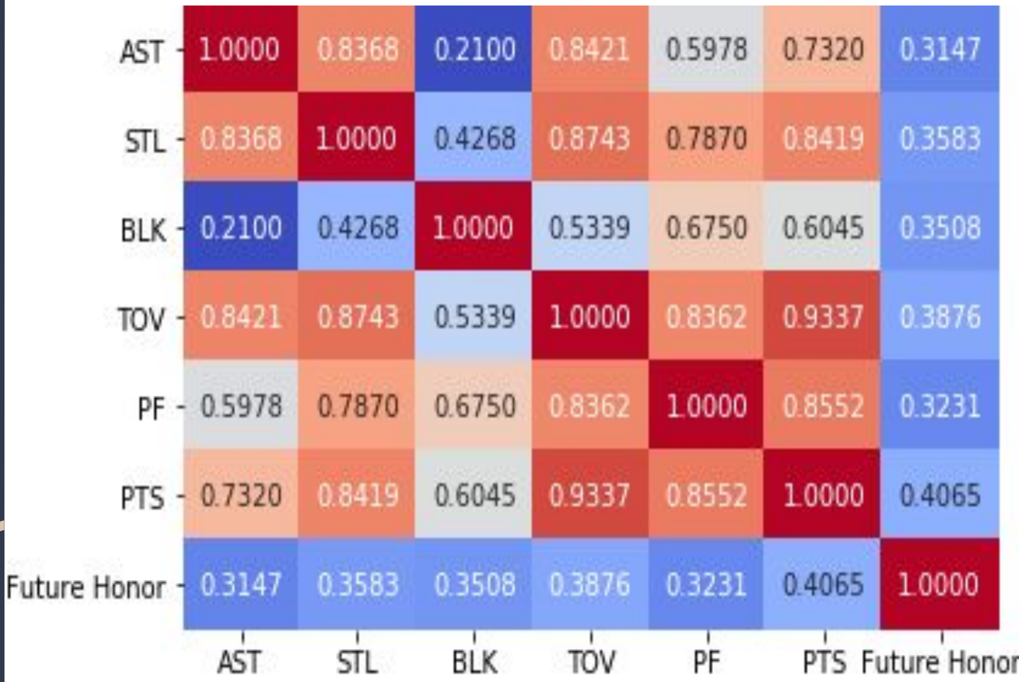
Correlation Matrices of Player Statistics to investigate Collinearity



	FT	FTA	FT%	ORB	DRB	TRB	Future Honor
FT	1.0000	0.9904	0.1998	0.7633	0.8315	0.8266	0.4140
FTA	0.9904	1.0000	0.1427	0.8060	0.8565	0.8584	0.4168
FT%	0.1998	0.1427	1.0000	0.0296	0.1077	0.0828	0.0499
ORB	0.7633	0.8060	0.0296	1.0000	0.9020	0.9568	0.3569
DRB	0.8315	0.8565	0.1077	0.9020	1.0000	0.9885	0.3998
TRB	0.8266	0.8584	0.0828	0.9568	0.9885	1.0000	0.3940
Future Honor	0.4140	0.4168	0.0499	0.3569	0.3998	0.3940	1.0000

Exploratory Data Analysis

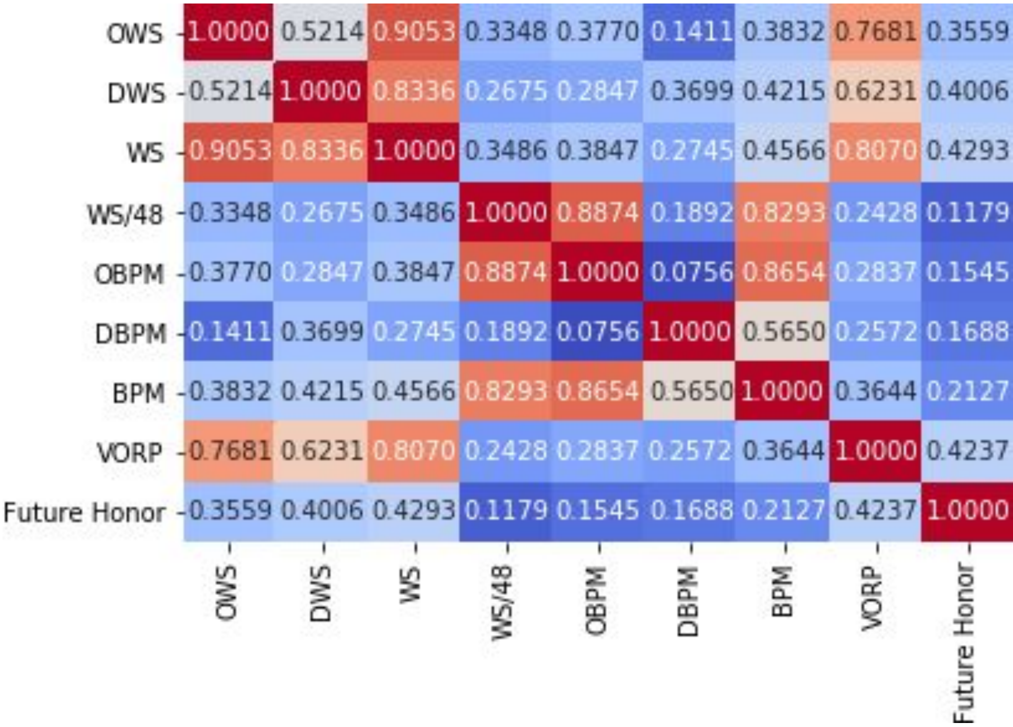
Correlation Matrices of Player Statistics to investigate Collinearity



	AST	STL	BLK	TOV	PF	PTS	Future Honor
AST	1.0000	0.8368	0.2100	0.8421	0.5978	0.7320	0.3147
STL	0.8368	1.0000	0.4268	0.8743	0.7870	0.8419	0.3583
BLK	0.2100	0.4268	1.0000	0.5339	0.6750	0.6045	0.3508
TOV	0.8421	0.8743	0.5339	1.0000	0.8362	0.9337	0.3876
PF	0.5978	0.7870	0.6750	0.8362	1.0000	0.8552	0.3231
PTS	0.7320	0.8419	0.6045	0.9337	0.8552	1.0000	0.4065
Future Honor	0.3147	0.3583	0.3508	0.3876	0.3231	0.4065	1.0000

Exploratory Data Analysis

Correlation Matrices of Player Statistics to investigate Collinearity



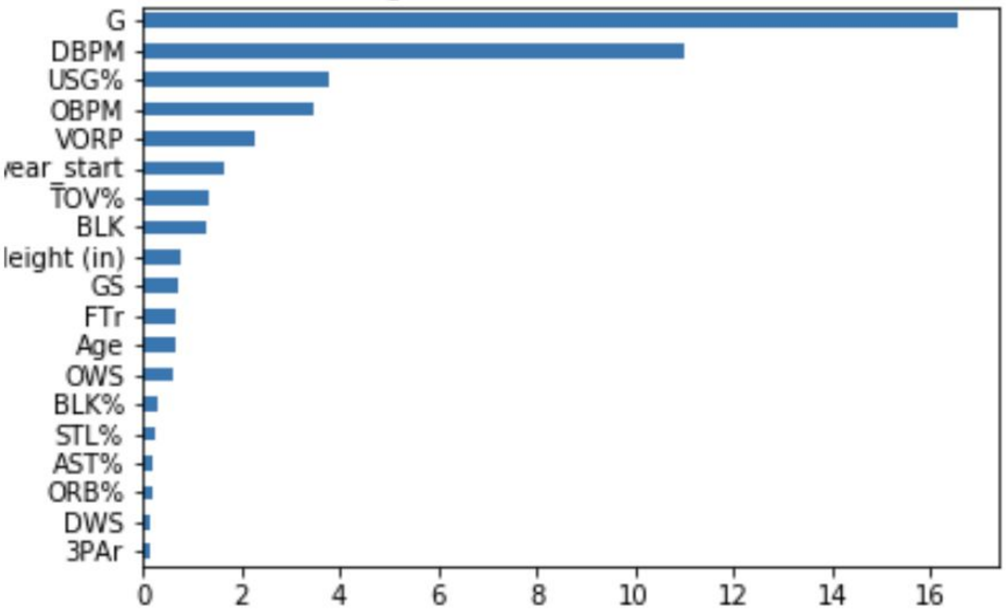
	OWS	DWS	WS	WS/48	OBPM	DBPM	BPM	VORP	Future Honor
OWS	1.0000	0.5214	0.9053	0.3348	0.3770	0.1411	0.3832	0.7681	0.3559
DWS	0.5214	1.0000	0.8336	0.2675	0.2847	0.3699	0.4215	0.6231	0.4006
WS	0.9053	0.8336	1.0000	0.3486	0.3847	0.2745	0.4566	0.8070	0.4293
WS/48	0.3348	0.2675	0.3486	1.0000	0.8874	0.1892	0.8293	0.2428	0.1179
OBPM	0.3770	0.2847	0.3847	0.8874	1.0000	0.0756	0.8654	0.2837	0.1545
DBPM	0.1411	0.3699	0.2745	0.1892	0.0756	1.0000	0.5650	0.2572	0.1688
BPM	0.3832	0.4215	0.4566	0.8293	0.8654	0.5650	1.0000	0.3644	0.2127
VORP	0.7681	0.6231	0.8070	0.2428	0.2837	0.2572	0.3644	1.0000	0.4237
Future Honor	0.3559	0.4006	0.4293	0.1179	0.1545	0.1688	0.2127	0.4237	1.0000

Machine Learning

Feature Importance and Selection

Isolated features with VIF scores below 10 (to avoid collinearity)

Ran logistic regression with remaining features and “Future Honor”, obtained Log-Odds for features with p<.05



Machine Learning

Feature Importance and Selection

Other Methods of Obtaining Feature Importances:

StatsModel Logit function to obtain coefficients for each features' influence on winning a future honor.

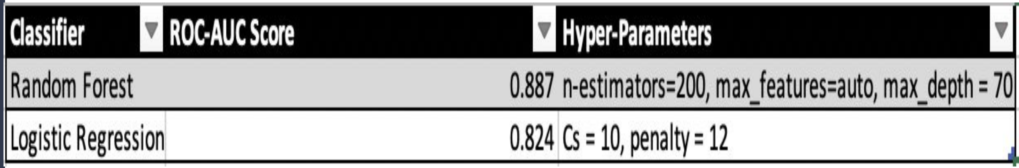
Ski-Kit learn package feature importance scores for each feature

Machine Learning

Model Selection

Hyper-Parameter Tuning and ROC-AUC scores for Random Forest and Logistic Regression

Random Forest was used for my final model, given better ROC-AUC score.

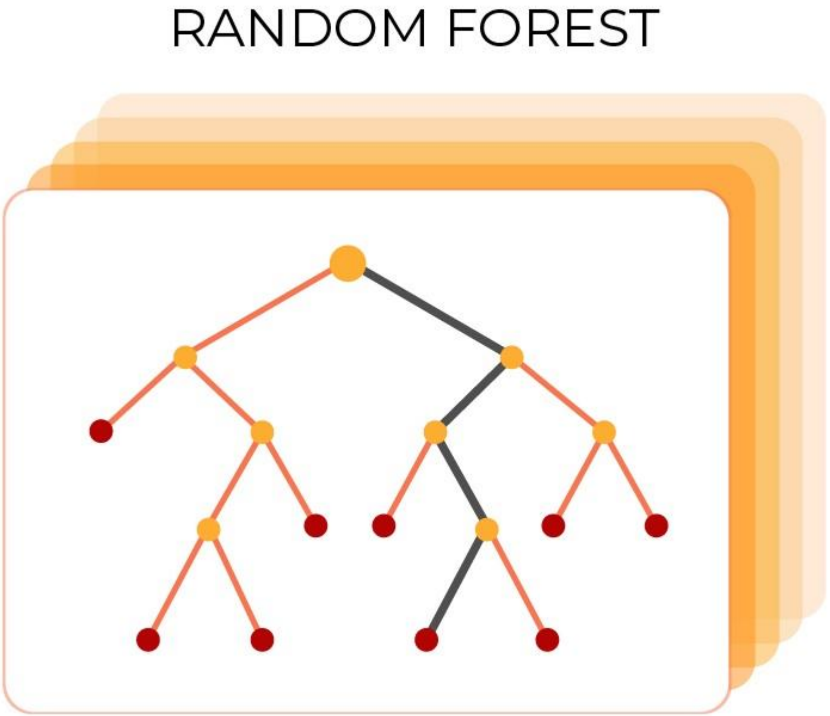


Classifier	ROC-AUC Score	Hyper-Parameters
Random Forest	0.887	n-estimators=200, max_features=auto, max_depth = 70
Logistic Regression	0.824	Cs = 10, penalty = 12

Machine Learning

Thresholding the Model

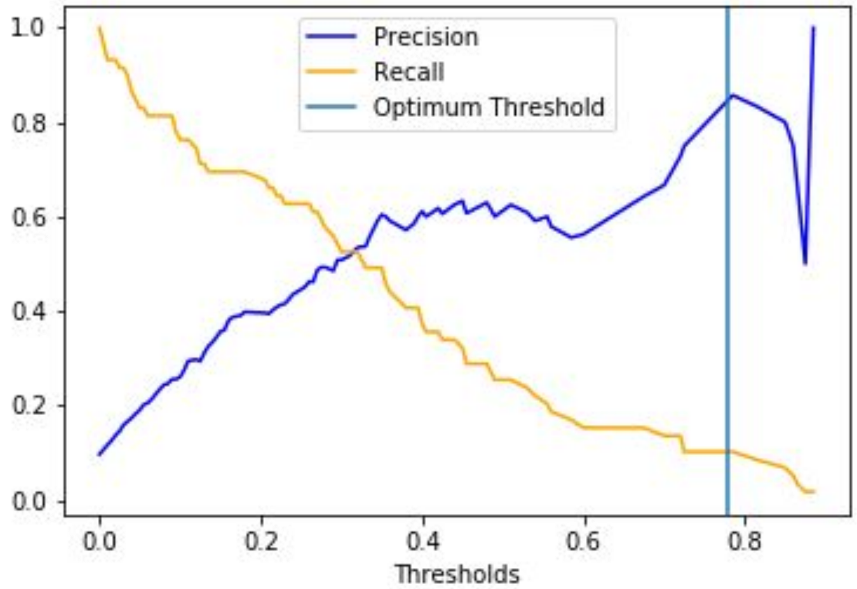
Model predicted 116 distinct probabilities for players likelihood of winning a “Future Honor”. I had to decide on a threshold probability value for which I could designate probabilities above that value as a “Future Honoree”.



Machine Learning

Thresholding the Model: Business Case 1

NBA Head Coaches deciding on their starting five. Precision is important over recall here; the players picked should truly have Future Honoree potential or the coach is wasting valuable development time on them.

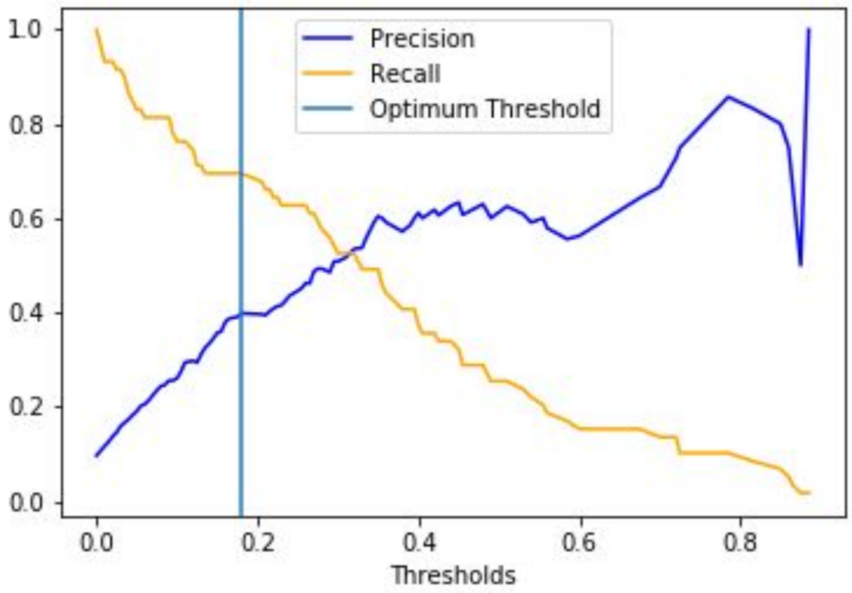


Machine Learning

Thresholding the Model: Business Case 2

NBA General Managers must decide which third and fourth year players to extend contracts to.

Recall is important here, want to keep highest proportion of future honorees as possible.



Conclusion

Very small % of players ascend to Future Honoree status, big advantage if they can be identified after rookie season

Advanced Statistics are as important as basic statistics for my model

Point Guards and Centers have an easier time becoming Future Honorees

Random Forest Model gave me highest accuracy levels, but the model should be used with different threshold values based on precision/recall trade-off.

Conclusion

Weaknesses

Model does not account for external variables impacting player success (team quality). Creating new features that account for team quality could help negate this.

Next Steps

Evaluate model's success as current crop of young players ascend to honoree status

Incorporate players' first 2-3 seasons into model rather than just the first season

Make individual models for All-Star, All-NBA, and All-Defense because different skill sets help for each honor.