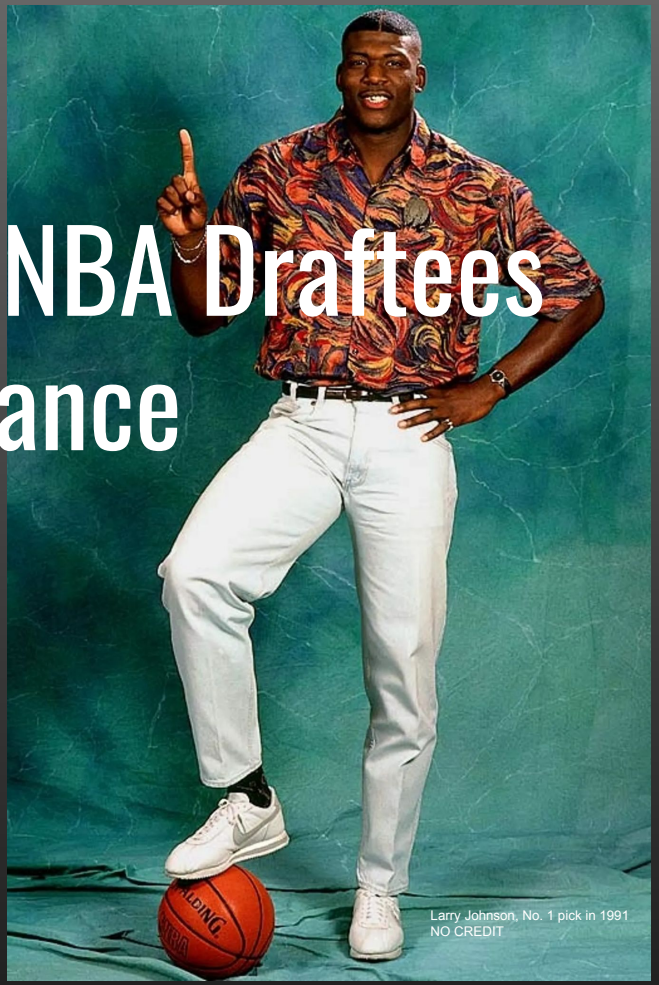
Predicting the Success of NBA Draftees Based on College Performance

...

By Joseph Aquino



The Problem

- Performance prediction of NBA prospects has considerable room for improvement, as shown by the lackluster drafting track records of many teams.
- Without relatively accurate performance forecasts of NBA prospects, the ownership and management of NBA teams cannot effectively determine which prospects to acquire or properly evaluate their young talent.
- Both the on-court and financial success of NBA teams are negatively impacted by inaccurate performance forecasts of prospects, since a single draft choice can completely change the outlook of an organization.

Measuring the Success of NBA Draftees

I have chosen to use 5-year peak performance (measured by maximum single-season positive Win Shares) to gauge the success of NBA draftees for the following reasons:

- The majority of an NBA player's development occurs during this timeframe.
- NBA teams are able to retain first-round Draft picks for 4 years on affordable contracts, after which expensive new contracts often need to be negotiated.
- Win Shares is a popular advanced statistic that assigns credit for team success to individual players on a team.
- WS handles the problem of assigning performance scores to draftees that never play in the NBA (zero WS) better than statistics like VORP or RAPTOR WAR.

NBA draftee success can be predicted with statistical and machine learning regression models

There are features (college stats and biographical data) that influence the 5-year peak performance of NBA draftees which can be identified by using a regression model.

We need a regression model that...

- Can predict the 5-year peak performance of NBA draftees using both NCAA statistics and biographical data.
- Can simultaneously relate multiple features to 5-year peak performance.
- And will only include features that have a significant effect on NBA success.

The highest-scoring regression model for our use is

Elastic Net

with Pearson's Rank Based
Feature Selection

Using my model

We can predict with relatively good accuracy (6.84 MSE)...

- The 5-year peak performance (measured by maximum single-season positive Win Shares) of any NBA draftee who played NCAA basketball and has data for each of the model features (2011 - present), since the performance of my model has shown to be generalizable.

Observations from Exploratory Data Analysis

5-year peak performance rankings

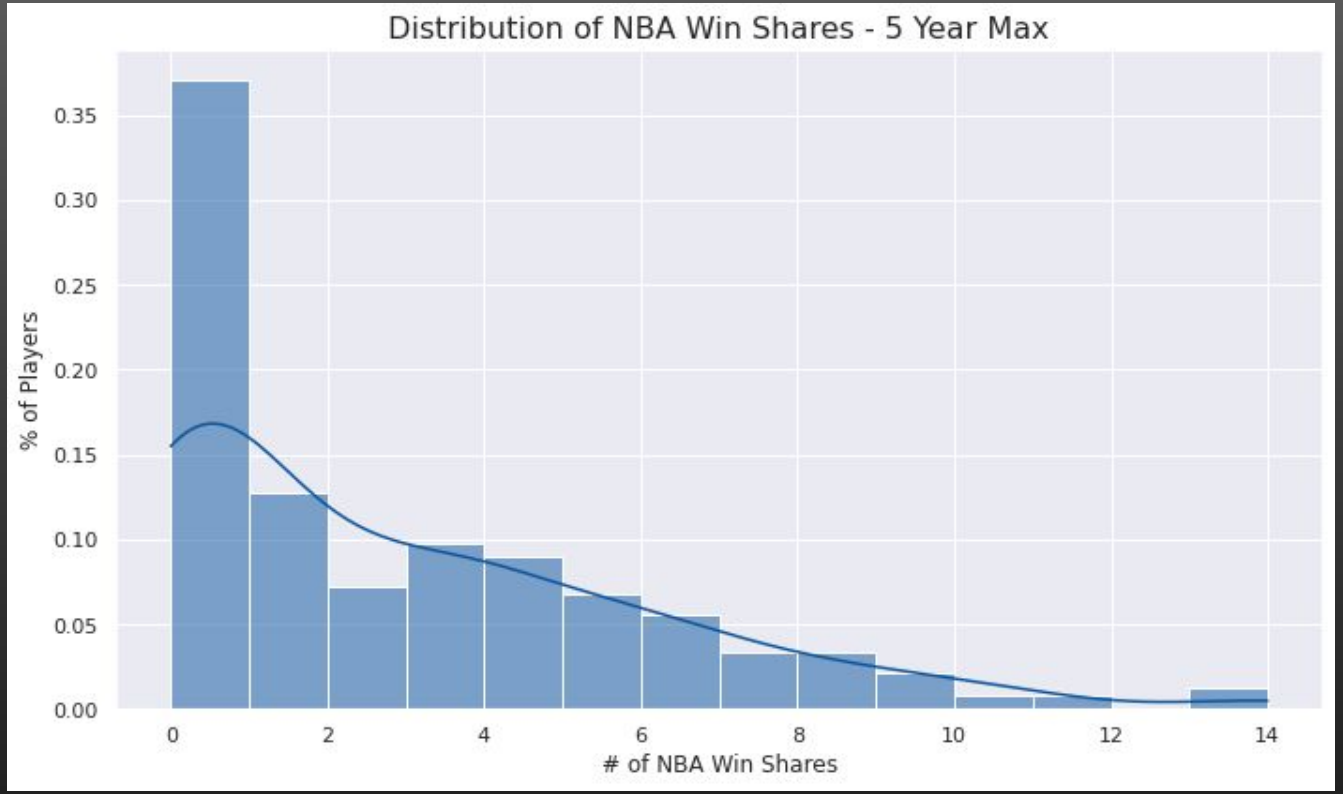
Max Single-Season WS of Players Drafted from 2011 to 2015

1.	Anthony Davis	14.0
2.	Karl-Anthony Towns	14.0
3.	Kawhi Leonard	13.7
4.	Jimmy Butler	11.2
5.	Draymond Green	11.1
6.	Damian Lillard	10.6
7.	Kyrie Irving	10.4
8.	Andre Drummond	9.9
9.	Kemba Walker	9.9
10.	Isaiah Thomas	9.7

➢ *All players became All-Stars*

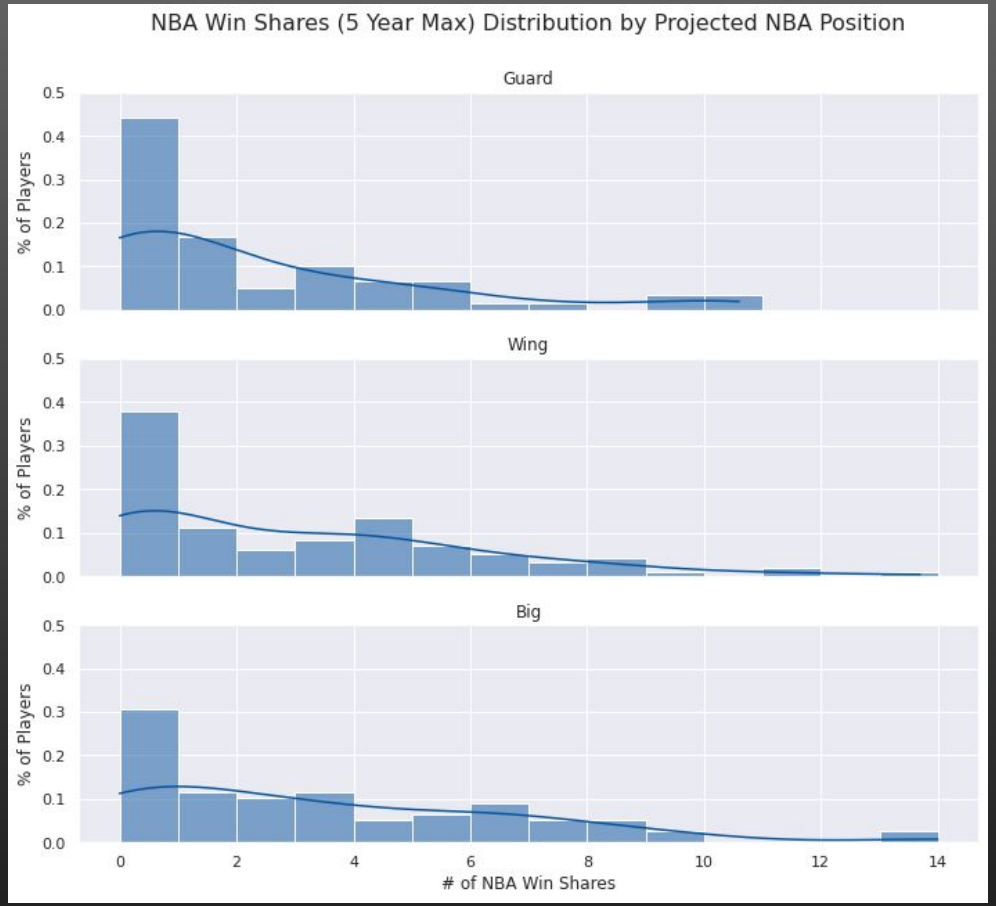
5-year peak performance

- Nearly half the players have a 5-year peak performance of less than 2 Win Shares.
- The number of players decreases exponentially as WS increase, demonstrating that few NBA draftees have successful careers.

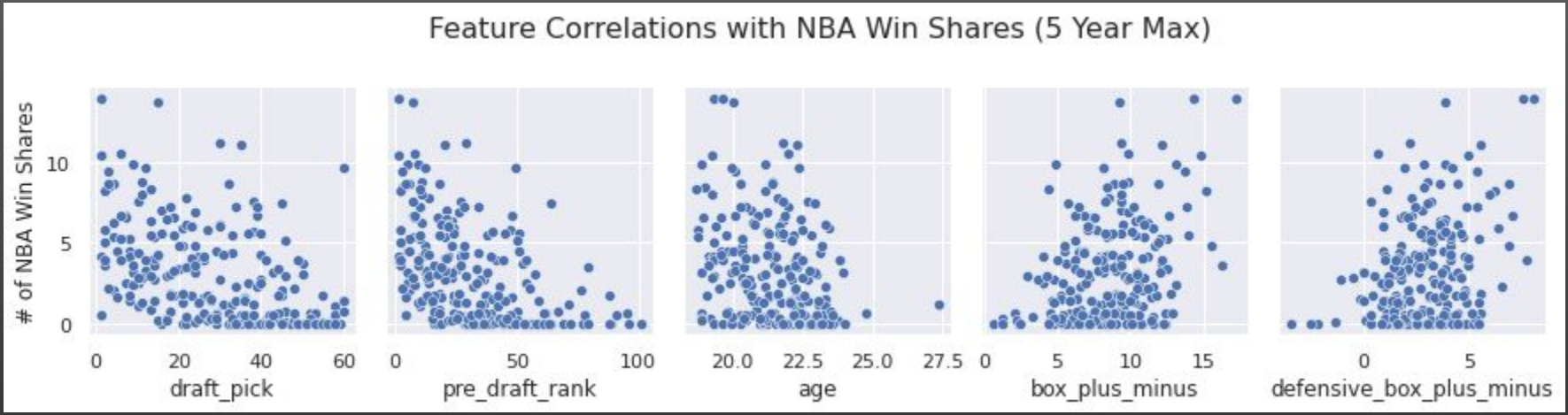


5-year peak performance
by projected NBA position

- The 5-year peak performance distributions of each projected NBA position (guard, wing, and big) look very similar: an observation supported by Kolmogorov-Smirnov tests.
- Consequently, projected NBA position likely has little effect on 5-year peak performance.



Features correlated with 5-year peak performance



- Of the 35 features of the dataset, NBA Draft position (r = -0.51), ESPN pre-draft ranking (r = -0.49), age (r = -0.34), Box Plus/Minus (r = 0.33), and Defensive Box Plus/Minus (r = 0.32) have the strongest correlation with 5-year peak performance.
- The correlations of these features with the target variable are shown in the plots above.

Features highly correlated with each other

- NBA Draft position is highly correlated with ESPN pre-draft ranking ($r = 0.82$); prospects are generally ranked close to their eventual draft position.
- Total rebound percentage is highly correlated with offensive rebound percentage ($r = 0.90$); good rebounders are generally also good offensive rebounders.
- Several advanced performance statistics are also highly correlated with each other (i.e. WS/40, PER, BPM), since they tend to rate players similarly.

Summary of observations

- 5-year peak Win Shares is a good indicator of draftee success, as each player ranked in the top-10 (2011 - 2015) became an NBA All-Star.
- As 5-year peak performance increases, the number of players who attain that level decreases exponentially, demonstrating that very few NBA draftees have successful careers.
- 5-year peak performance does not seem to be influenced by projected NBA position.
- NBA Draft position, ESPN pre-draft ranking, age, Box Plus/Minus, and Defensive Box Plus/Minus have the strongest correlation with 5-year peak performance.
- Lastly, the strong correlations among several features indicate that the dataset has a high level of multicollinearity, which must be taken into consideration when modeling.

Modeling

Nested cross-validation

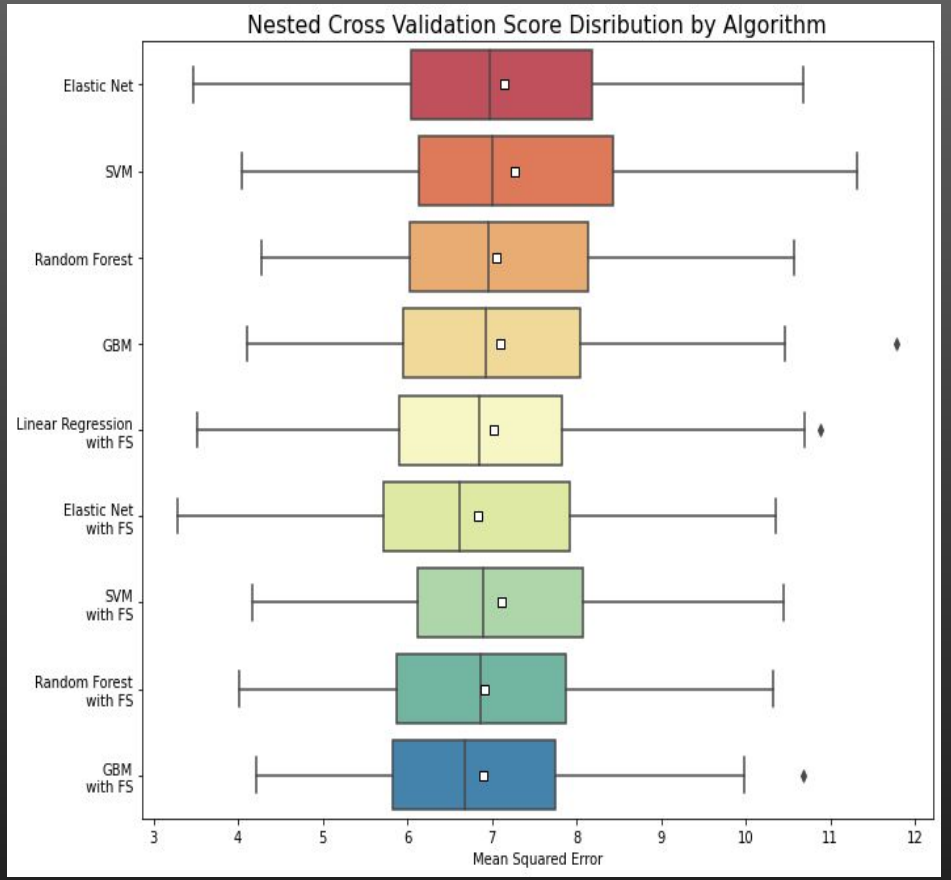
- Normally when modeling, the dataset is split into a train and test set, cross-validation is performed on the train set, and the performance of the best model is assessed on the test set.
- However, nested cross-validation allows you to perform both model selection and assessment without the need to set aside a test set, which is extremely useful when your dataset is small like mine (only 235 samples).
- The cross-validation procedure for model selection is nested inside the cross-validation procedure for model assessment. Instead of one tuned model being selected, several are selected, each is assessed on a different part of the dataset, and the scores are averaged to give a performance score.
- Nested cross-validation assesses the performance of the model obtained by applying the inner cross-validation procedure to the entire dataset.

Modeling process

- I compared the performance of 9 different optimized modeling algorithms using nested cross-validation with repetition and stratification.
- The following regression models were used: elastic net, support vector, random forest, gradient boosting, and linear regression.
- I also incorporated Pearson's rank based feature selection to reduce multicollinearity and model complexity.
- For each model but linear regression, 2 modeling algorithms were made: one with feature selection and one without. Due to high multicollinearity, linear regression could only be used with feature selection.

Nested cross-validation results

- The plot shows the nested cross-validation score distributions of each algorithm.
- Each algorithm performs significantly better than the target mean baseline model (9.75 MSE), and feature selection improves the performance of each algorithm.
- Elastic net regression with feature selection performs the best with an MSE of 6.84 ± 1.47.



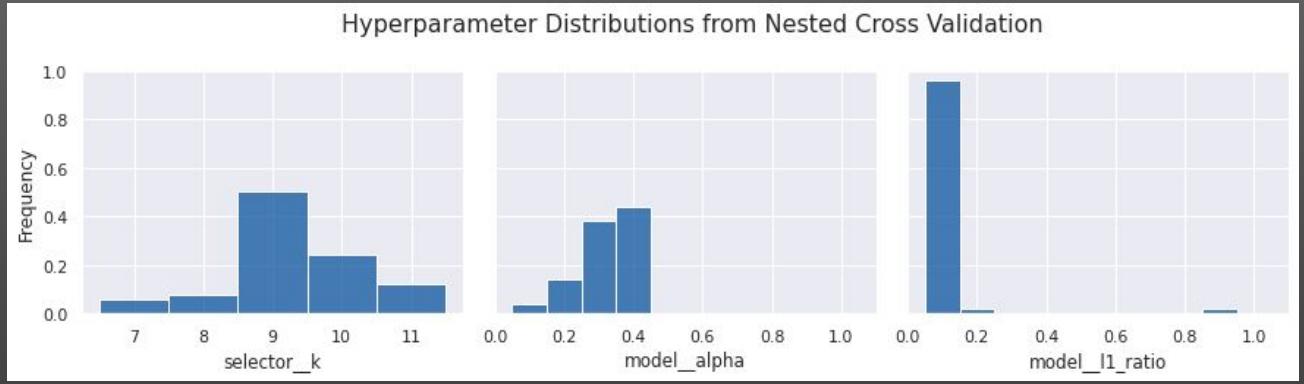
Model results

- Elastic net regression with Pearson's rank based feature selection is used to create my final model, since it attained the best nested cross-validation score.
- The inner-cross validation procedure of the nested cross-validation consists of 5 folds, 10 repeats, stratification, and the following hyperparameter grid:

  {'alpha': [0.1, 0.2, ... , 1.0], 'l1_ratio': [0.1, 0.2, ... , 1.0], 'k': [7, 8, ... , 11]}.
- When applied to the entire dataset, this procedure yields a model with the following hyperparameters: {'alpha': 0.3, 'l1_ratio': 0.1, 'k': 9}, which I will call the optimal cross-validatory chosen model.
- The nested cross-validation score (6.84 MSE) is the performance assessment of this optimal cross-validatory chosen model, my final model.

Model results

- Now I more closely analyze the nested cross-validation results of elastic net regression with feature selection.



- The plots show the distributions of hyperparameters from the models selected within the inner cross-validation.
- The hyperparameters are relatively stable across the models. This suggests that the optimal cross-validatory model (my final model) possesses stability and therefore is likely generalizable.
- The nested cross-validation score for this algorithm is 6.84, while the average inner cross-validation score is 6.79. The small difference between the inner and outer scores indicates the unlikelihood of overfitting within the inner cross-validation, which also suggests model generalizability.

Model results

- The dataset initially contained 35 features, but they were reduced to 9 for my final model via feature selection.
- My final elastic net regression model has an alpha value of 0.3 and an L1 ratio of 0.1, making it very similar to ridge regression.
- In order to make predictions using new data and view the importance of the features, the final model is fit on the entire dataset.
- The final elastic net regression model works on new data as follows: first the data is standardized, then 9 features are selected, and lastly the model multiplies each selected feature by a given coefficient and sums the results to make predictions.
- The model performs significantly better (2.62 RMSE) than the baseline (3.08 RMSE), but I had hoped that my extensive modeling process would lead to a more accurate model. This just demonstrates how difficult it is to accurately predict the success of NBA draftees.

Feature importance/coefficients

Ranked by absolute value

1. NBA Draft position: -0.58
2. ESPN pre-draft ranking: -0.48
3. Total rebound percentage: 0.34
4. Defensive Box Plus/Minus: 0.31
5. Age at NBA Draft: -0.29
6. Box Plus/Minus: 0.22
7. Win Shares per 40 minutes: 0.17
8. Player Efficiency Rating: 0.10
9. Defensive WS/40: 0.00

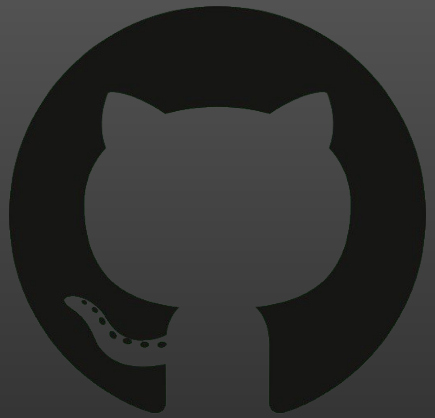
- My model selects 9 features, but when fitted on the training dataset, Defensive WS/40 is given a zero coefficient, so my final model effectively only has 8 features.
- Since my model incorporates correlation rank based feature selection, the 8 features of the model are those most correlated with 5-year peak performance.
- Also, since the model first standardizes the features, the importance of each feature is proportional to the absolute value of its coefficient.
- The 5 most important features of the model: draft position, pre-draft ranking, rebound percentage, Defensive BPM, and age, provide 80% of the influence for its prediction.

➢ *Going forward, I'll provide predictions for all the draftees of each NBA Draft starting with 2016!*

Potential uses for model

Personal Profile links

www.linkedin.com/in/joseph-aquino-a2644612



github.com/joe-aquino