NBA Draft Analytics: Boom or Bust?

Isaiah Hartman, Walmart and Oklahoma State University Student

ABSTRACT

Each year, thousands of players from all around the world submit their name to the National Basketball Association (NBA), in hopes that their talent warrants an invitation to the upcoming draft. The draft a process that enables teams to select the best talent to add to their roster. One of the major obstacles for teams in this process is trying to weed out the bad players, so to speak, from the good. Traditional selection methods have no real empirical backing and currently possess no foolproof measure of ensuring that the players chosen will perform well in the league. The purpose of this research paper is to develop models that can assist scouts in selecting the optimal players and prevent the selection of players who are more prone to "bust" or exhibit poor performance. In this study, I have developed several predictive models, such as decision trees and neural networks, that aide in forecasting a player's first-year performance in the NBA. These models illustrate how pre-draft characteristics such as an individual's height, hand size, weight, wingspan, and other attributes serve as a predictor for their first-year performance. The models were developed using SAS® Enterprise Miner™, and the best model will be selected based on the validation data mean squared error.

INTRODUCTION

Determining a potential player's worth or benefit to the team is a pivotal decision across all sports that can truly determine an organization's success or failure for years to come. Currently, in the National Basketball Association, when trying to determine which players are better than others, teams use a rather subjective process entailing scouts sent out to watch individuals perform, and then based on certain metrics predetermined by the team gauge how much added benefit the player can give to their firm. While players are still subject to participating in a draft, where height, weight, hand size and other physical ability tests are recorded, there is no real methodology in place where data driven insights can aide this process. Instead, these measurements are just tertiary data to aid the eyeball test from the team scouts.

Throughout the duration of this paper, I build several predictive models that aid in determining a player's first year performance in the NBA using various data mining techniques and instruments. The data was collected from qualified rookies' performances and NBA Draft Combine statistics over ten

years ranging from 2006-2016. The data consists of information regarding their first year performance: points, minutes played, blocks, assists, steals, rebounds and shooting percentage, as well as physical attributes recorded from the Combine: body fat percentage, height, weight, agility time, bench press and information regarding their leap. Using information from the players' first year performance I have created a performance metric that will serve as a target variable for one of the main models in this research. In addition to this, the models illustrate the likelihood that based on these Combine metrics, a rookie will score X amount of points, Y amount of rebounds, Z amount of steals and other attributes which can be used to assist the valuation of a players' worth or added benefit to the team, dependent upon the individuals fit and organization's needs.

Each year, the NBA hosts a draft where the best basketball talent from all around the world has the opportunity to submit their name into the metaphorical hat in hopes that one of the thirty teams will value them high enough for selection. The way the draft works is teams are awarded a pick 1-30 based upon their prior year's performance, then these teams each take turns selecting the best player (or fit) left on the board to be the newest member of their squad. Once each of the thirty picks have been made, the second round commences, after all sixty picks have been performed, then the draft ceases until the following year. Due to the fact that each team has only two picks, just one error allows a team to be left with essentially a multi-million dollar deficit which can haunt them for years to follow, thus the need for good data driven insight in this selection method is imminent.

METHODOLOGY

I collected data containing information from over 420 different NBA rookies who were drafted or signed within the last ten seasons. This information consisted of players' attributes from their body fat percentage and hand size, to how many assists they recorded compared to their turnover ratio during their rookie campaign. I created two separate datasets using data scraped from espn.com/nba/statistics for information about rookies first year performance in the NBA and stats.nba.com/draft/combine for characteristics regarding rookies attendance at the NBA Draft Combine. Since this study aims to measure an individual's first year performance, I included athletes of all positions; point guard, shooting guard, small forward, power forward and center into one giant dataset, so the players are evaluated on a level playing field. In addition to this, data was collected for both datasets from rookie classes over the last ten years, so players are evaluated compared to other first year performers and can accurately depict historically speaking, what

characteristics lead to what type of performance during the first year in the NBA.

FIRST YEAR PERFORMANCE

Due to the fact that different teams call for different playing situations for the rookies, I attempted to normalize many of the variables according to a 48 minute standard, the length of a regulation NBA game. The first year performance dataset consisted of twenty independent variables and one dependent interval variable that assists as a holistic measure of a rookie's first year performance. I came up with this calculation myself and it attempts to add points for positive performance while subtracting from the measure for negative. The calculation for this metric is as follows: ((Points*((3*3P%)+(2*FG%)+(FT%))*AST/TO)+((RP48+STP48+BLKP48)/PFP48))*(MPG/48), complete information regarding the first year performance dataset is below.

Variable	Level	Variable Description
Name	**Nominal**	A unique identifier for all athletes
Position	**Nominal**	Listed position on team roster
GP	**Interval**	The number of games participated in during rookie season
MPG	**Interval**	The average amount of minutes played for games players are involved in
PTS	**Interval**	The total amount of points scored during rookie campaign
FG%	**Interval**	An individual's overall shooting percentage during their first year
3P%	**Interval**	The historical likelihood of an individual making a 3 point shot
FT%	**Interval**	How likely each individual is to make a free-throw
OFF	**Interval**	The amount of offensive rebounds gathered during first year

DEF	**Interval**	The number of defensive rebounds obtained during first year
REB	**Interval**	The total number of rebounds during rookie campaign
RP48	**Interval**	The expected number of rebounds an individual can gather per 48 minutes played
AP48	**Interval**	The average number of assists distributed by players per 48 minutes played
AST/TO	**Interval**	A players individual assist to turnover ratio
STP48	**Interval**	The mean number of respective steals per 48 minutes by players
ST/TO	**Interval**	An individual's steals to turnover ratio
ST/PF	**Interval**	The ratio of steals to points made by a player
BLKP48	**Interval**	The expected number of blocks by a respective individual over 48 minutes
BLK/PF	**Interval**	The ratio of blocks to points for by an individual over 48 minutes
PFP48	**Interval**	The number of points scored by a player every 48 minutes during their rookie campaign
Performance Index	**Interval**	A holistic metric used as a summation of all the other indices to serve as a gauge of a players performance during their first year, higher number Corresponds to better performance

NBA ROOKIE COMBINE

The NBA Rookie Combine dataset consisted of fourteen input variables, and the one first year performance index which served as the target variable. Information in this dataset regards individual performance and personal characteristics which were measured during their Combine day. These attributes are all taken by professionals, so the accuracy of their reflection upon each individual is extremely high. For a complete record of information concerning this dataset, reference the table below.

Variable	Level	Variable Description
Body Fat %	**Interval**	The level of body fat recorded during players medical
Hand Length (inches)	**Interval**	The length of players hand from bottom of palm to middle finger in inches
Hand Width (inches)	**Interval**	The width in inches of players hand
Height W/Shoes	**Interval**	Individuals height with their basketball shoes on
Height W/O Shoes	**Interval**	Individuals height in their socks
Weight (lbs.)	**Interval**	The amount each individual weighs
Lane Agility Time (seconds)	**Interval**	The time in seconds recorded during the lane agility drill
Shuttle Run (seconds)	**Interval**	The time in seconds recorded during the shuttle run drill
Three Quarter Sprint (seconds)	**Interval**	The time in seconds recorded during the three quarter sprint drill
Standing Vertical Leap (inches)	**Interval**	The height a player jumped from a standing position
Max Vertical Leap (inches)	**Interval**	The height recorded allowing the player to jump in any fashion they see fit
Max Bench Press (repetitions)	**Interval**	The number of times each player benched 135lbs
True Wingspan	**Interval**	The length in inches measured from a players fingertip on one hand to the fingertip on the other
Standing	**Interval**	The distance in inches a player jumped from a standing position
GP	**Interval**	The number of games played in last competitive season
Performance Index	**Interval**	Used as the target variable for this dataset

Using the data, several models were created including a decision tree and linear regression that can aid in predicting future rookie's performance. Predictive modeling for this project was performed in accordance with the steps outlined in the SEMMA (Sample, Explore, Modify, Model and Assess) process, which was developed by SAS Institute Inc., a complete list of the steps is portrayed by the figure below.
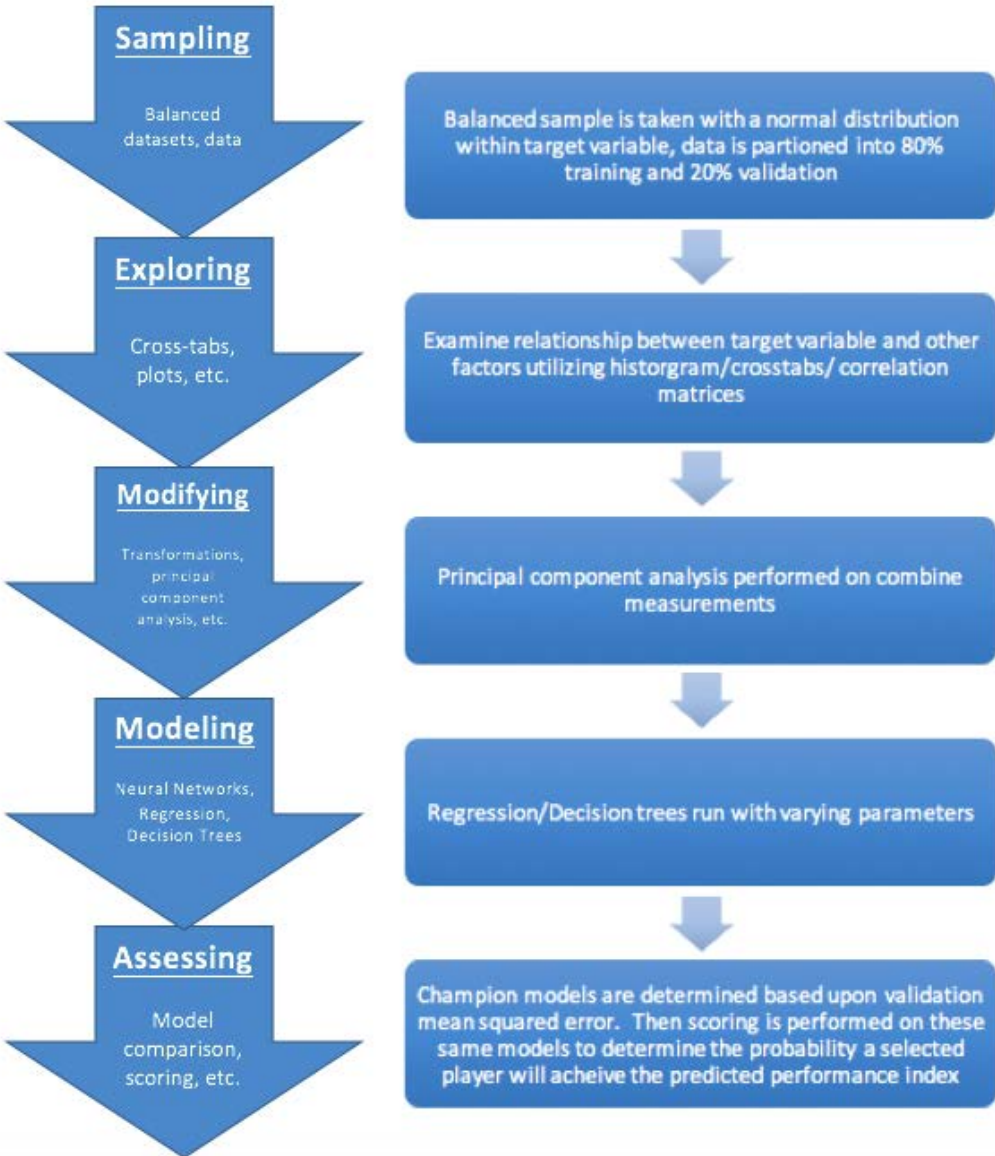
Sampling

Balanced datasets, data

Balanced sample is taken with a normal distribution within target variable, data is partioned into 80% training and 20% validation

Exploring

Cross-tabs, plots, etc.

Examine relationship between target variable and other factors utilizing historgram/crosstabs/ correlation matrices

Modifying

Transformations, principal component analysis, etc.

Principal component analysis performed on combine measurements

Modeling

Neural Networks, Regression, Decision Trees

Regression/Decision trees run with varying parameters

Assessing

Model comparison, scoring, etc.

Champion models are determined based upon validation mean squared error. Then scoring is performed on these same models to determine the probability a selected player will acheive the predicted performance index

Figure 1. SEMMA Process

In accordance with the above data mining technique, these steps were executed on both of the datasets, but specifically for the Combine dataset prior to predictive modeling.

After the data collection phase was completed, the next step was to ensure I had a balanced sample of all five draft-able positions which would assist in eliminating bias from the model and results. In an attempt to adjust for the oversampling, the prior probabilities were set with respect to the percentage of positive results in the Combine dataset which was gauged according the mean performance index score. Data was then partitioned for modeling purposes into 80% training and 20% validation using the stratified sampling method. Next, I utilized several exploratory tools such as box plots, histograms, scatter plots and odds ratio to clearly illustrate the relationship between the performance index and my input variables. Initially, I found wingspan and a few other variables to have a relatively high association with the target variable, bringing forth a multicollinearity issue. To combat this, I engaged in a Principle Component Analysis with which I selected PCs with eigenvalues greater than one then renamed them accordingly to utilize for the remainder of my analysis.

Input variables were then run through classification models and several unique regression and decision tree models were assembled to determine which model had the highest accuracy and most useful inputs towards predicting the players first year performance. Some of the variation amongst the regression model consists of the selection method - forward, backward and stepwise, as well as different levels for the polynomial degree and two-factor interaction. For the decision tree models, the variation stemmed from adjustments to the maximum depth, significance level for splits and interval target criterion either ProbF or Variance. These models were all then evaluated based upon validation mean squared error rate, then once the best model of each type was determined, it was run again with the entire dataset. The champion model from each test provides a system of predicting a player's first year performance in the NBA based upon the performance metric outlined above which can be utilized within franchises as a means of evaluating a player's potential benefit to the organization.

RESULTS

The first analysis conducted concerned the prediction of a players first year performance. I ran a linear regression with the performance index as a target variable and all other variables from the Combine table as inputs. The champion regression model allowed for multi-factor interaction and a forward selection method.

REGRESSION

The analysis of variance table illustrates that the overall model is statistically significant as p<alpha at a .05 level of significance so we reject the null hypothesis that the mean performance score is the same across the variables.

	Analysis of Variance					
Source	DF		Sum of Squares	Mean Square	F Value	PR > F
Model	8		18499	2312.32	8.54	<.0001
Error	258		69821	270.62		
Corrected Total	266		88320			

Output 1. Output from a Regression ANOVA

The analysis of effects for this model portray the level of significance for each of these variables and show that position as well as the multi-factor variable are still statistically significant. This table clearly depicts position as the most significant input, perhaps this could lead to a need to separate analyses according to individuals position.

	Type 3 Analysis of Effects			
Effect		DF	Sum of Squares	F Value	PR > F
Position		6	4528.83	2.79	.012

		DF	Estimate	Std. Error	t Value	PR > |t|
Body Fat * Max Vertical Leap^2		1	1135.22		4.19	0.0416
Height W/O Shoes * Three Quarter Sprint * True Wingspan		1	31.38		0.12	0.7338

**Output 2. Output from a Regression Type 3 Analysis of Effects**

Finally, the estimate table allows you to clearly depict the influence that each corresponding value for the respective variable has on the prediction of the players first year performance index. The variable position is broken down so franchises can adjust their formula accordingly when attempting to predict the players score.

Parameter		DF	Estimate	Std. Error	t Value	PR > |t|
	Analysis of Maximum Likelihood Estimates					
Intercept		1	11.7	19.2	0.61	0.54
Position	C	1	-6.1	5.02	-1.21	0.23
Position	F	1	-7.36	14.33	-0.51	0.61
Position	G	1	1.5	14.46	0.1	0.92
Position	PF	1	-5.74	4.34	-1.35	0.18
Position	PG	1	13.3	4.46	2.97	0.0032
Position	SF	1	-1.28	3.9	-0.33	0.74
Body Fat * Max Vertical Leap^2		1	0.08	0.04	2.05	0.04
Height W/O Shoes * Three Quarter Sprint * True Wingspan		1	-0.0003	0.0009	-0.34	0.73

Output 3. Output from a Regression Estimate Table

DECISION TREE

The next modeling technique I used for prediction was a decision tree, the champion decision tree's interval target criterion was ProbF, the maximum depth was 6 and the significance level for splits was 0.2, the model is shown below.
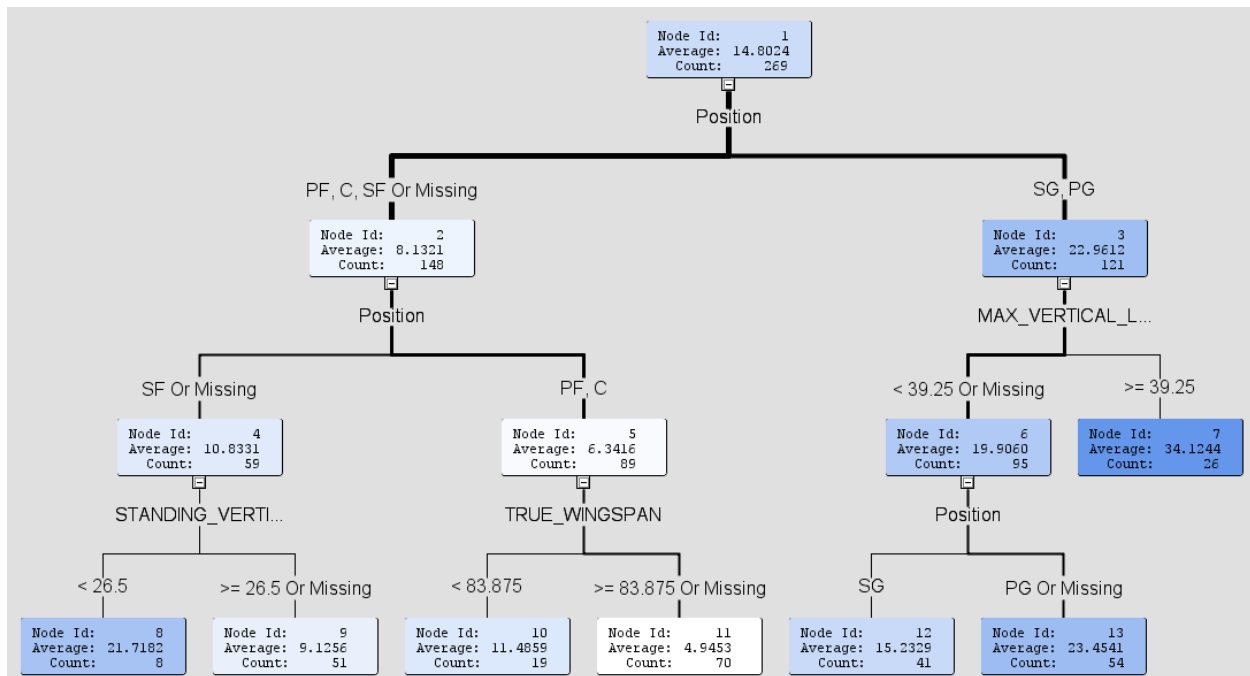
Display 1. Screenshot of Champion Decision Tree

This table illustrates the variable importance of each input selected by the model which clearly depicts position again as one of the most significant indicators of a player's first year performance. In addition to this, max vertical leap serves as a relatively good predictor, an insight that could absolutely benefit franchises for years to come.

Variable Importance			
Variable Name		**Number of Splitting Rules**	**Importance**
Position		3	1
Max Vertical Leap		1	0.4937
Standing Vertical Leap		1	0.2545
True Wingspan		1	0.1943

Output 4. Output from a Decision Tree Variable Importance

CONCLUSION

The results clearly illustrate that these models can accurately predict a NBA rookies' first year performance based upon several factors and interactions amongst those factors from the NBA Combine. Several unique predictive models were comprised for this prediction and the Champion Model was determined according to validation mean squared error. Once the specifications from both Champions were run on the entire dataset, according to average squared error, the model that performed the best is the decision tree. The performance of these models were not as high as I would have liked, so that leaves room for fine tuning/training in the future. Nonetheless, these models are still more reliable compared to the baseline and can still aid executives in making data driven decisions during the draft. Both the regression and decision tree provide a baseline metric of a player's probable first year performance, which can be used as a direct comparison between multiple individuals. I suggest using the performance metric as a valuation factor in addition to the information gathered by the teams' scouts to determine a players' overall impact during their first season.

This study is the first attempt to predict a rookies' first year performance using the performance index I comprised as a target variable and Combine performance/characteristics as inputs for the NBA. That being said, this research, in its current state does present a few shortcomings. I did not separate the players according to their listed position, certain positions, are more prone to certain statistics which influence the direction of the player's performance index. Creating separate datasets for each respective position would, in theory, allow for a more accurate prediction of players performance.

A player's contribution cannot be measured by this metric alone, as certain teams require players to take on more specific roles, and there are other factors such as defensive presence which certainly have an influence but are not recorded by this formula. The future scope of this research includes fine-tuning these models to predict more specific characteristics about first year performance dependent upon franchise needs. For instance, a franchise who lacks in scoring could use these model's to predict total points or a defensively lacking team can utilize them to determine the steals to turnover ratio as portrayed in the appendix.

REFERENCES

NBA Combine Statistics, Available at stats.nba.com/draft/combine.

NBA Player Statistics, Available at espn.com/nba/statistics

ACKNOWLEDGMENTS

This section is not required. If you include this section, do not change the heading style or the text "ACKNOWLEDGMENTS" of the preceding heading.

This is the text for the acknowledgments. This paragraph uses the PaperBody style.

CONTACT INFORMATION <HEADING 1>

Your comments and questions are valued and encouraged. Contact the author at:

    Isaiah Hartman
     (910) 467-7174
    isaiaho.hartman@gmail.com
    https://www.linkedin.com/in/isaiah-hartman-217119110/

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.