NBA Draft Analytics: Boom or Bust?

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.

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.

First Year Performance

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

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.

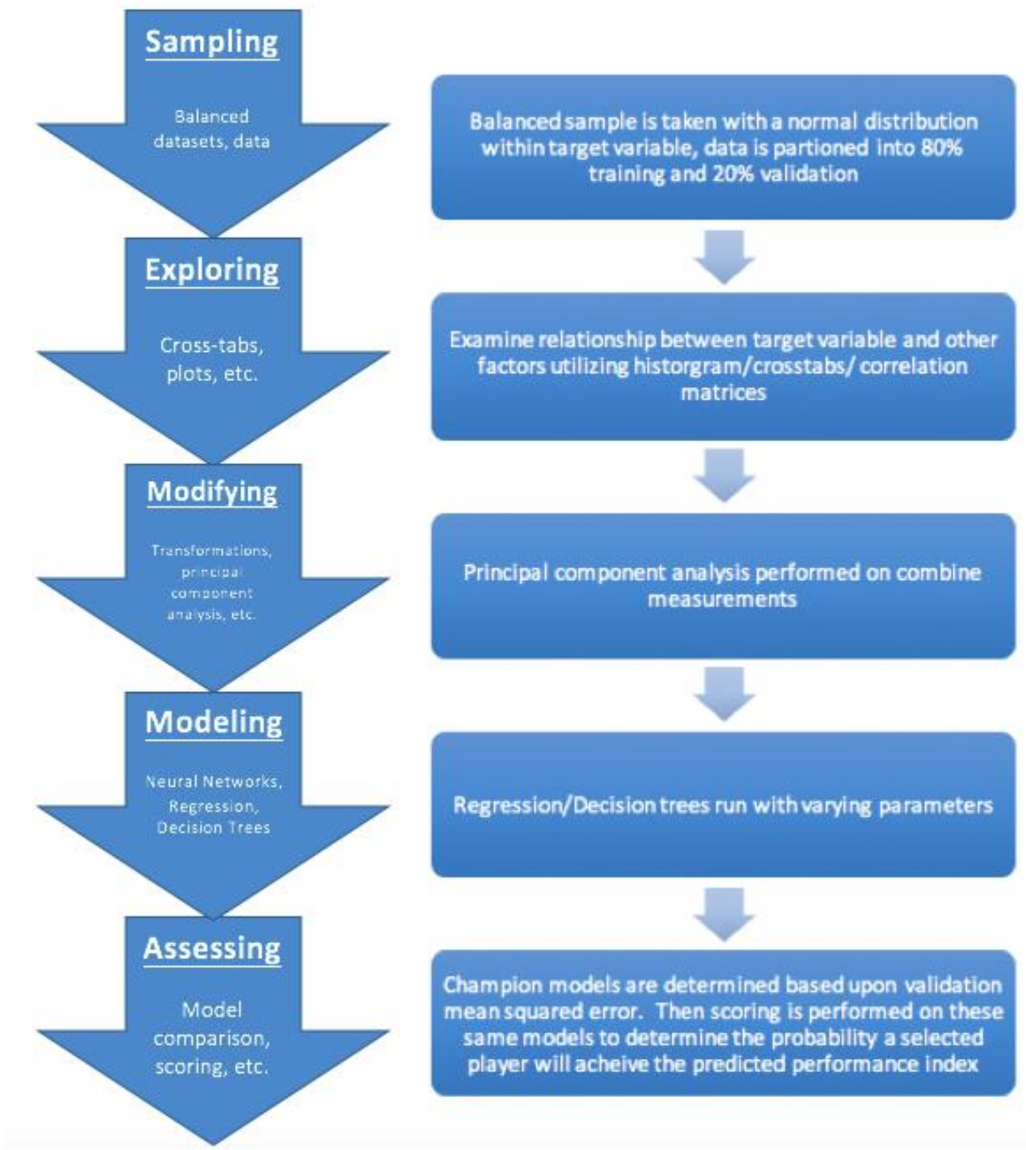
NBA Rookie Combine

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 rookies 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.



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

Performance Index

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	2312.320346	8.54	<.0001
Error	258	69821	270.623807		
Corrected Total	266	88320			

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	4528.8277	2.79	0.0120
BODY_FAT__*MAX_VERTICAL_LEAP___INCHES_*MAX_VERTICAL_LEAP___INCHES_	1	1135.2233	4.19	0.0416
HEIGHT_W_O_SHOES*THREE_QUARTER_SPRINT___SECONDS_*TRUE_WINGSPAN	1	31.3756	0.12	0.7338

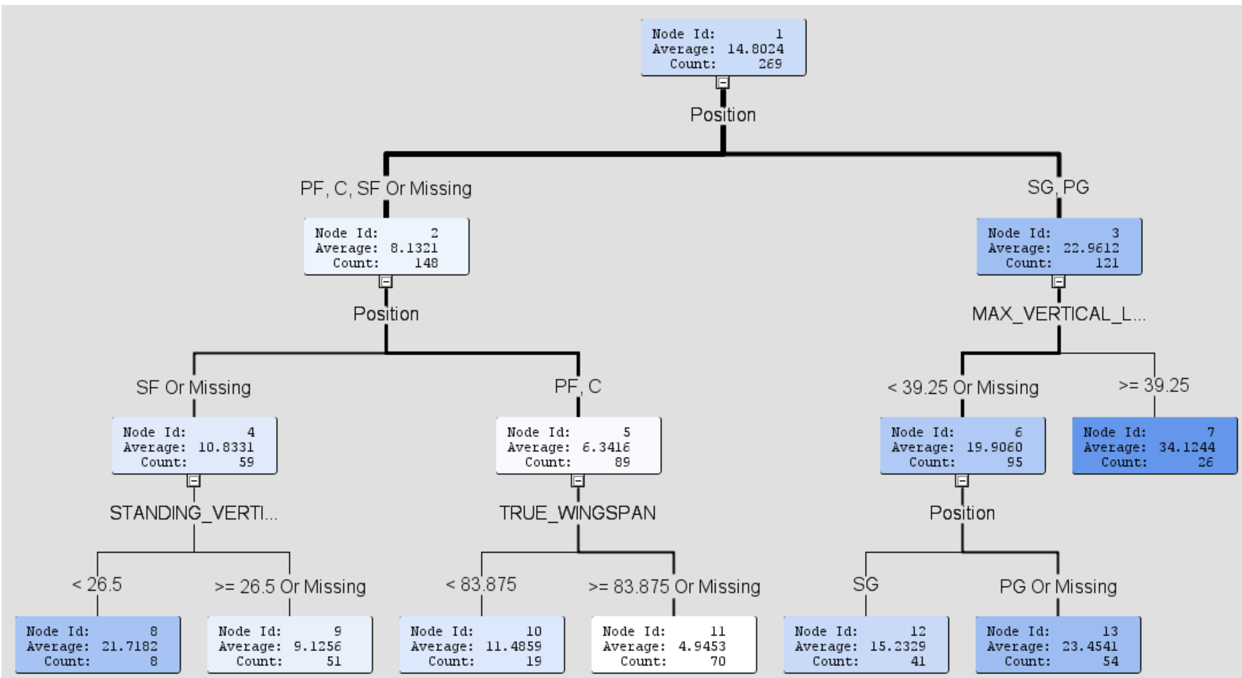
Finally, the estimate table allows you to clearly depict the influence that each corresponding value for the respective variable has on the prediction of the players first year performance index. The variable position is broken down so franchises can adjust their formula accordingly when attempting to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

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	11.7006	19.1927	0.61	0.5426
Position	C	1	-6.0975	5.0212	-1.21	0.2257
Position	F	1	-7.3603	14.3310	-0.51	0.6080
Position	G	1	1.4998	14.4617	0.10	0.9175
Position	PF	1	-5.7428	4.2394	-1.35	0.1767
Position	PG	1	13.2582	4.4573	2.97	0.0032
Position	SF	1	-1.2829	3.9052	-0.33	0.7428
BODY_FAT__*MAX_VERTICAL_LEAP___INCHES_*MAX_VERTICAL_LEAP___INCHES_		1	0.0822	0.0401	2.05	0.0416
HEIGHT_W_O_SHOES*THREE_QUARTER_SPRINT___SECONDS_*TRUE_WINGSPAN		1	-0.00029	0.000865	-0.34	0.7338

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.



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	Label	Number of Splitting Rules	Importance
Position	Position	3	1.0000
MAX_VERTICAL_LEAP___INCHES_		1	0.4937
STANDING_VERTICAL_LEAP___INCHES_		1	0.2545
TRUE_WINGSPAN		1	0.1943

Discussion and Future Scope

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.

Appendix

Performance Index Regression Fit Statistics

Fit Statistics

Target=Performance_Index Target Label=Performance Index

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	1514.15
ASE	Average Squared Error	260.32
AVERR	Average Error Function	260.32
DFE	Degrees of Freedom for Error	260.00
DFM	Model Degrees of Freedom	9.00
DFT	Total Degrees of Freedom	269.00
DIV	Divisor for ASE	269.00
ERR	Error Function	70025.31
FPE	Final Prediction Error	278.34
MAX	Maximum Absolute Error	78.96
MSE	Mean Square Error	269.33
NOBS	Sum of Frequencies	269.00
_NW_	Number of Estimate Weights	9.00
RASE	Root Average Sum of Squares	16.13
RFPE	Root Final Prediction Error	16.68
RMSE	Root Mean Squared Error	16.41
SBC	Schwarz's Bayesian Criterion	1546.50
SSE	Sum of Squared Errors	70025.31
SUMW	Sum of Case Weights Times Freq	269.00

Performance Index Decision Tree Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train
Performance_Index	Performance Index	_NOBS_	Sum of Frequencies	269
Performance_Index	Performance Index	_MAX_	Maximum Absolute Error	70.56649
Performance_Index	Performance Index	_SSE_	Sum of Squared Errors	65797.99
Performance_Index	Performance Index	_ASE_	Average Squared Error	244.6022
Performance_Index	Performance Index	_RASE_	Root Average Squared Error	15.63976
Performance_Index	Performance Index	_DIV_	Divisor for ASE	269
Performance_Index	Performance Index	_DFT_	Total Degrees of Freedom	269

Points

Regression

Analysis of Variance

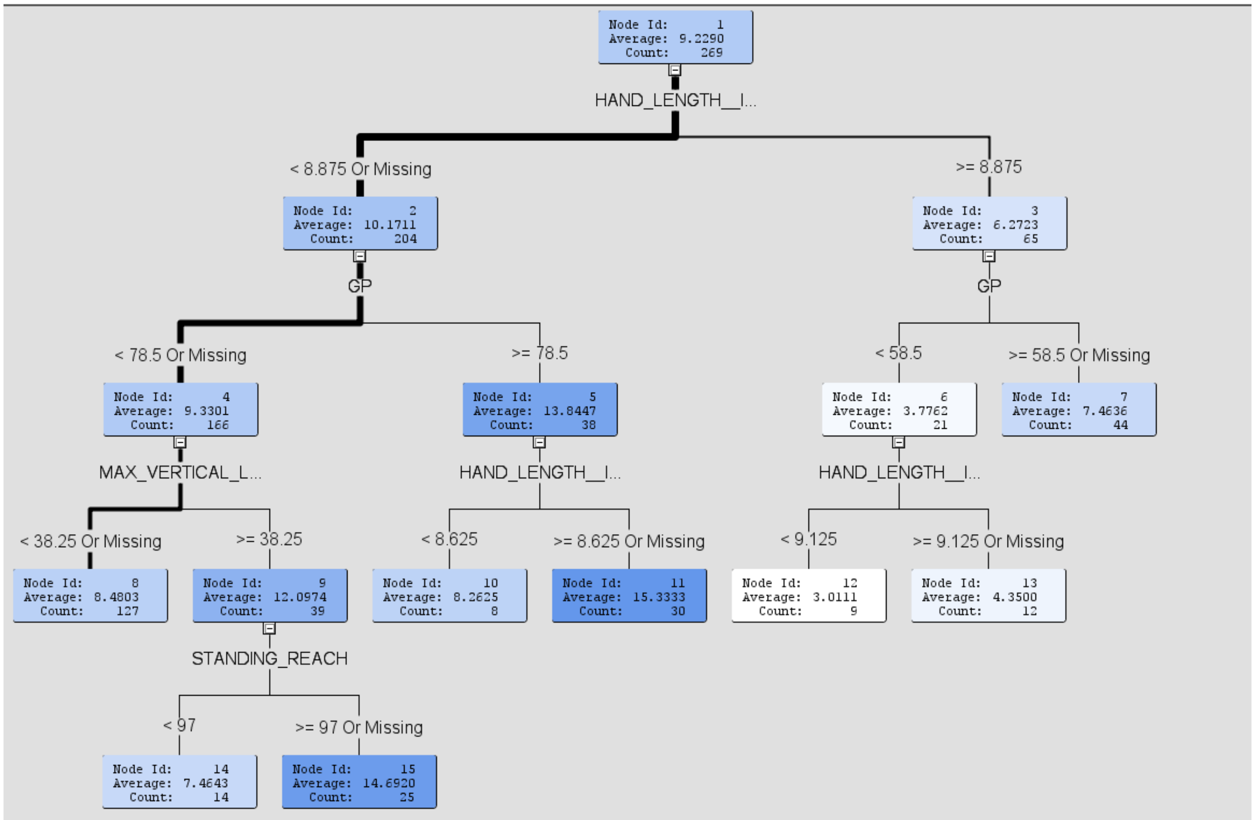
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1449.616233	362.404058	12.56	<.0001
Error	262	7559.325115	28.852386		
Corrected Total	266	9008.941348			

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.3681	4.3244	2.86	0.0046
GP*MAX_VERTICAL_LEAP___INCHES_	1	0.00255	0.000819	3.11	0.0020
BODY_FAT__*GP*GP	1	0.00570	0.00261	2.19	0.0295
HAND_LENGTH__INCHES_*HEIGHT_W_O_SHOES*THREE_QUARTER_SPRINT___SECONDS_	1	-0.00969	0.00275	-3.53	0.0005
LANE_AGILITY_TIME___SECONDS_*STANDING_REACH*TRUE_WINGSPAN	1	0.000114	0.000048	2.39	0.0178

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	916.08
ASE	Average Squared Error	29.03
AVERR	Average Error Function	29.03
DFE	Degrees of Freedom for Error	264.00
DFM	Model Degrees of Freedom	5.00
DFT	Total Degrees of Freedom	269.00
DIV	Divisor for ASE	269.00
ERR	Error Function	7809.10
FPE	Final Prediction Error	30.13
MAX	Maximum Absolute Error	14.83
MSE	Mean Square Error	29.58
NOBS	Sum of Frequencies	269.00
_NW_	Number of Estimate Weights	5.00
RASE	Root Average Sum of Squares	5.39
RFPE	Root Final Prediction Error	5.49
RMSE	Root Mean Squared Error	5.44
SBC	Schwarz's Bayesian Criterion	934.06
SSE	Sum of Squared Errors	7809.10
SUMW	Sum of Case Weights Times Freq	269.00

Decision Tree



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
HAND_LENGTH__INCHES_		3	1.0000
GP	GP	2	0.8756
STANDING_REACH		1	0.6606
MAX_VERTICAL_LEAP___INCHES_		1	0.6028

Fit Statistics

Target=PTS Target Label=PTS

Fit Statistics	Statistics Label	Train
NOBS	Sum of Frequencies	269.00
MAX	Maximum Absolute Error	12.92
SSE	Sum of Squared Errors	6334.28
ASE	Average Squared Error	23.55
RASE	Root Average Squared Error	4.85
DIV	Divisor for ASE	269.00
DFT	Total Degrees of Freedom	269.00

Steals to Turnover Ratio

Regression

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3.828702	0.425411	5.09	<.0001
Error	257	21.471865	0.083548		
Corrected Total	266	25.300567			

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Position	6	2.6669	5.32	<.0001
STANDING_REACH*THREE_QUARTER_SPRINT___SECONDS_	1	1.5180	18.17	<.0001
HAND_LENGTH__INCHES_*HEIGHT_W_O_SHOES*SHUTTLE_RUN___SECONDS_	1	0.4743	5.68	0.0179
STANDING_REACH*STANDING_VERTICAL_LEAP___INCHES_*WEIGHT__LBS_	1	0.3410	4.08	0.0444

Analysis of Maximum Likelihood Estimates

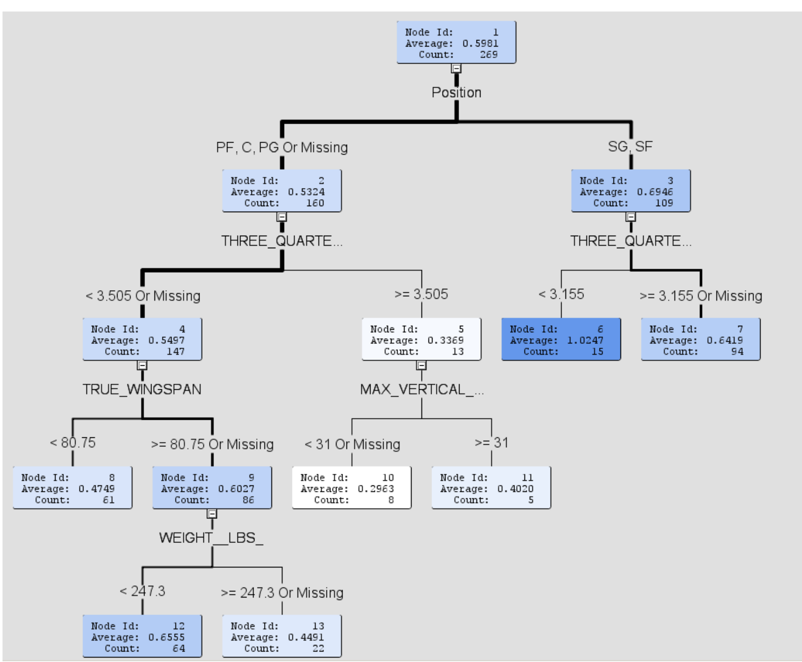
Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	1.9262	0.5174	3.72	0.0002
Position	C	1	0.1041	0.0921	1.13	0.2597
Position	F	1	-0.0579	0.2516	-0.23	0.8181
Position	G	1	-0.1866	0.2539	-0.73	0.4631
Position	PF	1	0.1155	0.0771	1.50	0.1355
Position	PG	1	-0.1996	0.0815	-2.45	0.0150
Position	SF	1	0.1813	0.0685	2.65	0.0087
STANDING_REACH*THREE_QUARTER_SPRINT___SECONDS_		1	-0.00573	0.00134	-4.26	<.0001
HAND_LENGTH__INCHES_*HEIGHT_W_O_SHOES*SHUTTLE_RUN___SECONDS_		1	0.000446	0.000187	2.38	0.0179
STANDING_REACH*STANDING_VERTICAL_LEAP___INCHES_*WEIGHT__LBS_		1	-5.16E-7	2.553E-7	-2.02	0.0444

Fit Statistics

Target=ST_TO Target Label=ST/TO

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	-657.268
ASE	Average Squared Error	0.081
AVERR	Average Error Function	0.081
DFE	Degrees of Freedom for Error	259.000
DFM	Model Degrees of Freedom	10.000
DFT	Total Degrees of Freedom	269.000
DIV	Divisor for ASE	269.000
ERR	Error Function	21.693
FPE	Final Prediction Error	0.087
MAX	Maximum Absolute Error	1.885
MSE	Mean Square Error	0.084
NOBS	Sum of Frequencies	269.000
_NW_	Number of Estimate Weights	10.000
RASE	Root Average Sum of Squares	0.284
RFPE	Root Final Prediction Error	0.295
RMSE	Root Mean Squared Error	0.289
SBC	Schwarz's Bayesian Criterion	-621.321
SSE	Sum of Squared Errors	21.693
SUMW	Sum of Case Weights Times Freq	269.000

Decision Tree



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
THREE_QUARTER_SPRINT___SECONDS_		2	1.0000
Position	Position	1	0.8369
WEIGHT__LBS_		1	0.5351
TRUE_WINGSPAN		1	0.4890
MAX_VERTICAL_LEAP___INCHES_		1	0.1189

Fit Statistics

Target=ST_TO Target Label=ST/TO

Fit Statistics	Statistics Label	Train
NOBS	Sum of Frequencies	269.000
MAX	Maximum Absolute Error	1.655
SSE	Sum of Squared Errors	20.125
ASE	Average Squared Error	0.075
RASE	Root Average Squared Error	0.274
DIV	Divisor for ASE	269.000
DFT	Total Degrees of Freedom	269.000