SMU Data Science Review

Volume 1 | Number 3 Article 7

2018

Predicting National Basketball Association Success: A Machine Learning Approach

Adarsh Kannan
Southern Methodist University, akannan@smu.edu

Brian Kolovich
Southern Methodist University, bkolovich@smu.edu

Brandon Lawrence
Southern Methodist University, blawrence@smu.edu

Sohail Rafiqi
Southern Methodist University, srafiqi@smu.edu

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

 Part of the Applied Statistics Commons

Recommended Citation
Kannan, Adarsh; Kolovich, Brian; Lawrence, Brandon; and Rafiqi, Sohail (2018) "Predicting National Basketball Association Success: A Machine Learning Approach," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 7.
Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/7

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

Predicting National Basketball Association Success: A Machine Learning Approach

Adarsh Kannan, Brian Kolovich, Brandon Lawrence, Sohail Rafiqi

Master of Science in Data Science
Southern Methodist University
6425 Boaz Lane, Dallas, TX 75205, USA
akannan@smu.edu, bkolovich@smu.edu,
blawrence@smu.edu, srafiqi@smu.edu

Abstract. In this paper, we present a machine learning based approach to projecting the success of National Basketball Association (NBA) draft prospects. With the proliferation of data, analytics have increasingly become a critical component in the assessment of professional and collegiate basketball players. We leverage player biometric data, college statistics, draft selection order, and positional breakdown as modelling features in our prediction algorithms. We found that a player's draft pick and their college statistics are the best predictors of their longevity in the National Basketball Association.

1 Introduction

The National Basketball Association (NBA) has been holding an annual draft for prospects since 1947. The draft is responsible for the highest influx of new players to the league. Teams look to develop a strategic edge over the competion in the Draft, but all teams have the same information. This information includes biometric measurements and past performance for college players[1]. This paper uses a machine learning approach to predict success using player information available on Draft Day.

Each year, many analysts rank the players in the Draft[2]. To aid teams in their decision, the NBA hosts a combine where players are put through a battery of physical tests and their physical characteristics are measured.This information is made available to the public and is used just as much by the media as the NBA teams themselves. Major publications typically attempt to predict the future performance of prospective players and provide in depth analysis for the fans to consume [3]. Despite the depth of data available, most commentary focuses on the intangible aspects of players [4].It is to be expected that external[5] and intangible[6] factors play a role in success, the aim of this paper is to determine the role of the quantifiable, puiblicly availabel data in success.

We evaluate classification techniques to determine how well the publicly available player data predicts success in the NBA. For this study, success is defined

as a player having played 174 games during the time frame of interest. The 174 game cutoff represents the average number of games played by the player in the data set. Our approach will be to evalute three binary classification models: logistic regression, random forest, and support vector machine. We select the best model based on a combination of error metrics. Once selected, we run the model on the reduced model, biometric data, and the full model, biometric data and collegiate data.

The random forest classifier outperformed logistic regression and support vector machine for this datra set and was used for the classification exercise and subsquent analysis. The reduced model shows little predictive power. The full model, however, does show that draft pick and collegiate statistics significantly boost the model's predictive power. We conclude that collegiate performance and pick number in the draft are the best predictors for a player's future success in the NBA.

The remainder of this paper is organized as follows: In Section 2 we discuss the dataset, its acquisiton, related challenges, and the definition of sucess. In Section 3, we observe the distribution of various player metrics, test for covariates, and explain adjustments to variables. We explain the models that the paper will evaluate, the reasoning behind them, and the evaluation metrics in Section 4. The model results are discussed in Section 4. Section 5 discusses feature performance and analysis. We discuss the ethics of data collection and biometric tracking in Section 6. Section 7 explores the implications of the analysis as well as the potential for future work.

2 Dataset Selection

2.1 Dataset Overview

Our data is predicated on our ability to design a model that could predict the probability of 'success' of NBA draft combine participants. Due to sparseness in historical data, we restrained the data to the 2009-2014 NBA draft combines and to the players who participated in all biometric measurements. 2014 was chosen as the cut-off year since we believe anything less than that was not enough time for a player to play in the required 174 games, which is our threshold for determining success/no success. The final dataset in the analysis consists of 194 records across 30 columns.

Table 1. Data Set Description

Variable	Description	Additional Details	Variable Type
player	PlayerID	n/a	Categorical
college	Player college	n/a	Categorical
draft_yr	Year drafted	n/a	Ordinal
fnl_coll_rpi	Final Ratings Percentage Index of player's final college season	n/a	Ordinal
success	Dependent variable in the analysis for success	1 = success, 0 = no success	Numeric
age_first_yr	Age at the start of 2017 NBA season	n/a	Ordinal
draft_pick	Order of selection in player's respective draft	1-60 = drafted, 61 = undrafted	Numeric
hght_noshoes	Height w/o shoes, (inches)	n/a	Numeric
hght_wtshoes	Height w/ shoes, (inches)	n/a	Numeric
wingspan	Wingspan (inches)	n/a	Numeric
Standing_reach	Standing reach (inches)	n/a	Numeric
vert_max	Max vertical leap (inches)	n/a	Numeric
vert_maxreach	Max reach from vertical (inches)	n/a	Numeric
vert_nostep	Vertical w/t no steps (inches)	n/a	Numeric
clg_gms_plyed	Total number of games played in college	n/a	Numeric
pts_ppg	Average points per game from college career	n/a	Numeric
trb	Average rebounds per game from college career	n/a	Numeric
ast	Average assists per game from college career	n/a	Numeric
fg2_pct	Average 2 point field goal percentage from college career	n/a	Numeric
fg3_pct	Average 3 point field goal percentage from college career	n/a	Numeric
ft_pct	Average free throw percentage from college career	n/a	Numeric
guards	Binary variable indicating guard position	1=guards, 0=not guards	Numeric
forwards	Binary variable indicating forward position	1=forwards, 0=not forwards	Numeric
centers	Binary variable indicating center position	1=centers, 0=not centers	Numeric
drafted	Binary variable indicating drafted or undrafted	1=drafted, 0=undrafted	Numeric
nba_gms_plyed	Total number of games played in NBA	n/a	Numeric

The biometric data, player name, and draft year were pulled and aggregated directly from a user uploaded dataset from data.world[1], a data distribution platform. It consists of a basketball player's biometric statistics from their respective NBA draft combine, which is held annually prior to the actual draft. The biometric statistics consist of wingspan, hand size, vertical jump, etc. Additionally, the college statistics and college attended data were scraped from theSports Reference website[2] and combined with the biometric data. For college statistics, we used the averages for all statistics along with total games, which is just the total number of games played. The total number of NBA games played data was scraped from Basketball Reference[3], a website that provides aggregated NBA data. Our dependent variable, Success, was coded and determined by the total number of NBA games played by the player.To derive the final ratings percentage index variable, we downloaded the 2009-2014 ratings percentage index data from the NCAA website[4] and cross-referenced it against the player's college and draft year. Lastly, we created custom features by referencing existing variables or through manual reconciliation from Web research

Table 2. Player Feature Data Set

Feature	Description
Age_first_year	age of player at the starting of the NBA season
Guards	coded as 1 if the player is either a point guard or shooting guard
Forwards	coded as 1 if the player is either a small forward or power forward
Centers	coded as 1 if the player is a center
Drafted	coded as 1 if the player is drafted in his respective draft

3 Exploratory Data Anlaysis

3.1 Pre-Modelling

Due to the multivariate nature of the model comprised of several data on different scales, e.g. average assists per game ranges from 0.1-8.0 while final RPI ranges from 1-206, we standardized the feature values across the dataset. By stabilizing the range and variability of the data, we reduce the risk of certain

[1] https://data.world/achou/nba-draft-combine-measurements
[2] https://www.sports-reference.com/cbb/players
[3] https://www.basketball-reference.com/players/
[4] https://www.ncaa.com/rankings/basketball-men/d1/ncaa-mens-basketball-rpi

features exhibiting unequal contributions to the model predictions.

As an initial exploration of the data, we analyzed the distributions from the categorical variables to identify any potential pitfalls or sparseness in the population data. Figure 1 represents the distribution of the players by basketball position. From this, we see that the highest concentration is from the guard position, which comprises 47% of the population. Forwards comprise 38% and centers encompass 15% of the data, respectively. While the position breakdown is skewed towards guards and forwards, it makes sense given that we consolidated the point/shooting guards as guards, small/power forwards as forwards, and centers as a stand-alone position.

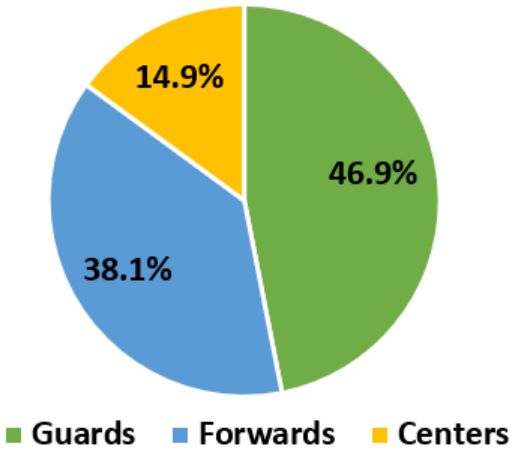

Fig. 1. Distribution of Players by Position

We then assessed the distribution of players by draft year, as seen in Figure 2. As noted in previous sections, biometric and data in general was inconsistent and sparse in the early to late 2000s. While the percentage of 2009 participants is significantly lower than the other drafts, there are no fundamental differences in how the biometric data was measured, or how college statistics were tabulated that could potentially induce confounding factors in the analysis.

Lastly, to assess correlation, we isolated the variables with the highest correlation (Figure 3). We considered any metric with a value above 0.70 or below -0.70 highly correlated. We observe certain features exhibiting high correlation, which could be a sign of redundancy. For example, "Height No shoes" and "Height

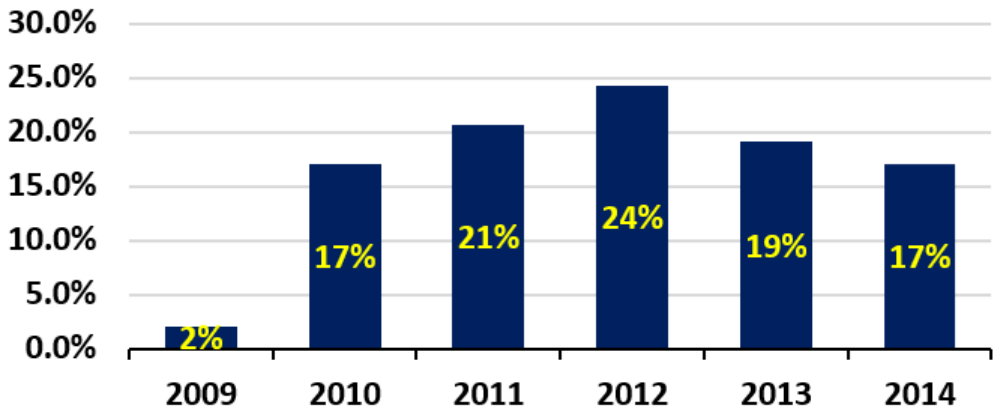

Fig. 2. Distribution of Players by Draft Year

With shoes" have near perfect positive correlation. This makes sense, as the two matrics are effectively measuring the same thing. We will inspect in this features in greater detail during the modelling phase.

	Height No Shoes	Height With Shoes	Wingspan	Standing Reach	Vertical No Step	Weight	Assists	Guards	Forwards
Hght_noshoes		0.996	0.845	0.914	0.787	0.75	-0.708	-0.778	0.388
Hght_wtshoes	0.996		0.844	0.915	0.782	0.756	-0.704	-0.779	0.38
Wingspan	0.845	0.844		0.902	0.822	0.737	-0.688	-0.711	0.366
Standing_reach	0.914	0.915	0.902		0.796	0.713	-0.681	-0.724	0.359
Vert_nostep_rch	0.787	0.782	0.822	0.796		0.593	-0.688	-0.664	0.393
Weight	0.75	0.756	0.737	0.713	0.593		-0.605	-0.72	0.374
Assists	-0.708	-0.704	-0.688	-0.681	-0.688	-0.605		0.647	-0.397
Guards	-0.778	-0.779	-0.711	-0.724	-0.664	-0.72	0.647		-0.738
Forwards	0.388	0.38	0.366	0.359	0.393	0.374	-0.397	-0.738	

Fig. 3. Correlation Results

4 Model Determination

4.1 Model Types

We determined the best binary classification model for answering our question of interest - "How do we predict future success and non-success based off biometric, college statistics, draft order, and position data as described above?" Since the answer to our question is a dichotomous non-numeric outcome, it stands to reason that we will focus on supervised classification algorithms as a roadmap

for determining the appropriate model type.

We evaluated three distinct supervised classification models as part of our testing plan, which are (1) Logistic Regression, (2) Random Forest, and (3) Support Vector Machine. We discuss the advantages of each below.

The Logistic Regression (LR) model performs well with features that exhibit linear relationships and with binary responses that can be mathematically separated through linear equations. In instances where the linearity assumption is violated, feature sets may benefit from various types of data transformations, such as a log or higher order transformation. The predictions may also be evaluated as probabilities or as odds ratios, depending on the preferred measurement of the analyst.

A LR model may be interpreted by the following equation:

$$logit(p) = ba + b_1X_1 + b_2X_2 + b_3X_3 + ...b_kX_k$$

where p is the probability of presence of a particular response. A transformation of the logit may be interpreted as the logged odds, as expressed below://

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presense\ of\ response}{probability\ of\ absence\ of\ response}$$

Conversely, a Support Vector Machine (SVM) model builds a function (or hyperplane) that aims at separating binary classes of data[7]. The linear SVM algorithm is represented by points in a plot, where it attempts to divide the points into the highest possible separation, also referred to as the margin. While LR models are linear based, a SVM is capable of performing non-linear classification, by using a kernel trick such as a radial basis function.

Figure 4 highlights the hyperplane (solid line), the two support vectors (dashed lines), and the maximum margin between the two support vectors.

Lastly, the Random Forest (RF) approach to classification offers considerable differences between both the LR and SVM models. The RF is a decision tree based model that does not require linear relationships in the data. Additionally, since the model is a combination of decision trees, it is suitable for our problem since binary "decisions" are made based on the thresholds computed in each respective branch in the tree.

4.2 Model Selection Criteria

As stated above, we will be testing three classification algorithms in an attempt to find the best fit for our model predictions. As part of the two-pronged model

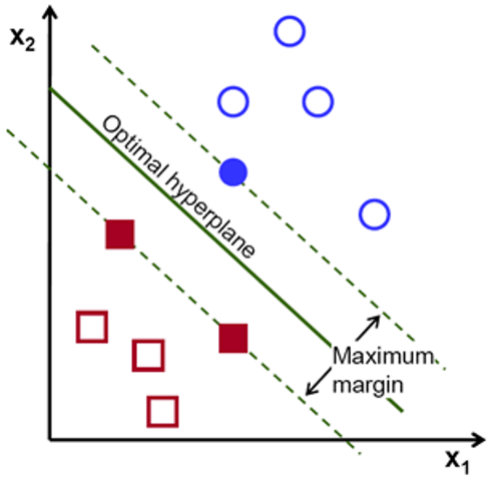

Fig. 4. Support Vector Machine Visualized

Fig. 5. Random Forest Decision Tree

approach, our first classification task may be interpreted as a "reduced" model, since it only incorporates biometric data and the age of the player prior to his first year of NBA experience. Our intention is to isolate the biometric and age data to assess the feasibility of a model comprised solely of biometrics. For our second classification task, we wish to assess the impact of incorporating additional features, notably the college statistics, RPI, draft order, and positional breakdown. Our assumption is that the model predictions should improve, but that assumption will remain undetermined until the final results are assessed.

Table 3 gives an overview of the statistical measures that we will use to assess the quality of each classifier.

Table 3. Model Quality Measures

Measure	Description
Average Precision score	the number of true positives over the number of true positives plus the number of false positives.
Average Recall	the number of true positives over the number of true positives plus the number of false negatives.
Average F1 score	the harmonic mean of precision and recall.

By assessing all scores as opposed to just one, we inspect the model's ability (or lack of ability) to accurately classify success and to reduce the number of false positives and false negatives. In general, the higher the precision, recall, and F1 scores, the better the model.

Additionally, a weakness of the accuracy statistic is that it ignores the negatives costs associated with misclassification. As a counter, we will use a cost matrix that is predicated on penalties for misclassified predictions (i.e. false positive and false negatives) and credits for correct predictions (i.e. true positives and true negatives). For consistency, we will use the following cost methodology for all of models. A lower score indicates a better model.

$$CostScore = \sum((TruePositives * -1) + (FalsePositives * 10) + (FalseNegatives * 10) + (TrueNegatives * -1) \tag{1}$$

Prior to fitting our model and making predictions, we will initiate a training/testing data split of 80/20. The testing data serves as the true acid test of how the model will perform post production, since it does not have the luxury of

the 'expected' outcome data that is allocated to the training data. To reduce the variance introduced from performing training/testing splits, we will implement a 10-fold cross validation that removes the possibility of the split only having one set. Both measures ensure a properly validated and robust approach prior to model implementation.

5 Analysis

5.1 Reduced Model

The scikit-learn module via Python was the primary method used to generate the classification results. For each classifier, we implemented a custom grid search that computes a score of the estimator from an optimal set of parameters used in the cross-validation. Additionally, we retained the same random sample generator for each classifier to ensure the same sample(s) were using in determining optimal fit.

Table 4 is a compilation of the computed Precision, Recall, and F1 Scores from each of the three classifiers. The RF model clearly trumps the LR and SVM methods, as it outperforms both models in every precision, recall, and F1 metric.

Table 4. Classifier Results

Logistic Regression				
Classifer	Precision	Recall	F1	Support
No Success	0.5	0.61	0.55	18
Success	0.59	0.48	0.53	21
Average/Total	0.55	0.54	0.54	39

Support Vector Machine				
Classifer	Precision	Recall	F1	Support
No Success	0.5	0.61	0.55	18
Success	0.59	0.48	0.53	21
Average/Total	0.55	0.54	0.54	39

Random Forest				
Classifer	Precision	Recall	F1	Support
No Success	0.71	0.67	0.69	18
Success	0.73	0.76	0.74	21
Average/Total	0.72	0.72	0.72	39

As a secondary level of measure effectiveness, we also computed the Cost Scores associated with each classifier. Due to its strong ability to classify false

negatives, the RF again was the best model based on its total cost score of 82, compared to 159 and 159 from SVM and LR. The computed cost score from confusion matrices are show below.

Logistic Regression

```
Confusion Matrix
[[11  7]
 [11 10]]
```

Cost Score = (11)(-1) + (7)(10) + (11)(10) + (10)(-1) = 159

Support Vector Machine

```
Confusion Matrix
[[11  7]
 [11 10]]
```

Cost Score = (11)(-1) + (7)(10) + (11)(10) + (10)(-1) = 159

Random Forest

```
Confusion Matrix
[[12  6]
 [ 5 16]]
```

Cost Score = (12)(-1) + (6)(10) + (5)(10) + (16)(-1) = 82

Fig. 6. Computed Cost Scores

Since the evaluation criteria is heavily in favor of RF, we refrain from using the LR and SVM methods and focus solely on the RF method for evaluation.

After implementing the custom grid search and cross validation, an optimal model was fit using 10 estimators and a max tree depth of 10, with an accuracy score of 0.638 as seen below.

```
In [31]:
1 #print the best paramters found and best score
2 print(rfclf_Bio.best_score_)
3 print(rfclf_Bio.best_params_)

0.6388350699646016
{'max_features': 'sqrt', 'n_estimators': 10, 'criterion': 'entropy', 'max_depth': 10}
```

Fig. 7. Reduced Model Parameters and Score

We plotted the features importance values from the RF Reduced model. as shown in Figure 8 we can conclude that vertical jump and height of the player had the highest predictive power of player performance, followed by the players' wingspan and age prior to the first season. Hand width and hand length had less noticeable importance while the remaining features had roughly equal importance to the model predictions.

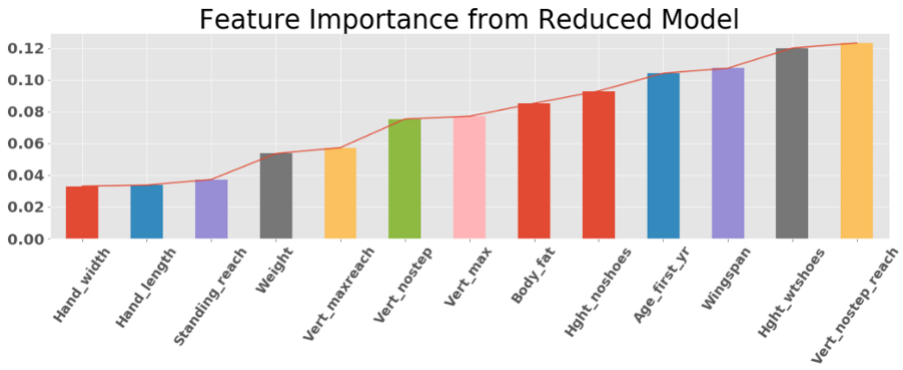

Fig. 8. Reduced Model Feature Importance Values

To attempt to obtain higher classification metrics, we will run a separate "Full Model" that incorporates all of the initial features into the predictions. These features include college statistics, draft selection order, and position breakdown.

5.2 Full Model

Outside of the additional features, the process for generating the Full Model results remained unchanged from the Reduced Model. As we mentioned in our earlier discussion, the genesis behind the Full Model is to test the hypothesis of observing a boost in predictive power, accuracy, and precision.

Overall, the Full Model demonstrates better precision and recall on averaged compared to the Reduced Model. While the positive recall statistic decreased from 0.76 to 0.67, we observe a significant increase in the model's ability to predict no success, as evidenced by the 0.67 to 0.94 increase in recall. Lastly, the model's ability to accurately predict success was quite significant with a precision score of 0.93.

Like the Reduced Model, we plotted the feature significance and its impact on the model predictions. As seen in Figure 10, we can conclude that the order

	Reduced Model			Full Model			Delta		
Classifer	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
No Success	0.71	0.67	0.69	0.71	0.94	0.81	0.00	0.27	0.12
Success	0.73	0.76	0.74	0.93	0.67	0.78	0.20	-0.09	0.04
Average/Total	0.72	0.72	0.72	0.83	0.79	0.79	**0.11**	**0.07**	**0.07**

Fig. 9. Reduced Model & Full Model Comparison

of selection in the draft had by far the highest impact, followed by total average rebounds in college. Additional college statistics such as average points per game and average three point field goal percentage also had significant importance to the model results.

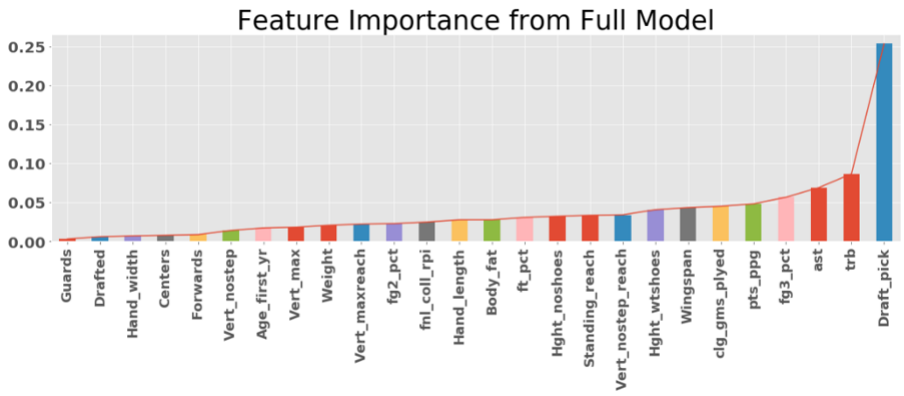

Fig. 10. Full Model Feature Importance Values

6 Ethics

This study was conducted using publicly available data sets. We recognize that our subjects, the players in the National Basketball Association, did not expressly consent to their performance data being used in this study. A player's performance data is inherently public and is constantly gathered and analyzed by teams, fans, gamblers, and other third parties[8]. As technology progresses, more metrics will become available to track ever more comprehensively[9]. Tracking biometric data in real time is a reality in professional sports. This has caused concern for some researchers as personally identifiable biometric data may overlap with protected medical data and may cross ethical lines if the data is used

inappropriately[10].

In this paper, we replaced player names with a unique PlayerID variable before running our models, but the player names are inherent to the source data as performance tracking of an individual requires identifiaction of the individual.

7 Conclusions and Future Work

Success in the National Basketball Association is best predicted by past performance and the position into which the player was drafted. College playing statistics greatly improves the predictive power of the model over biometric data alone. Wingspan, height, and reach were the most important, purely biometric indicators. We conclude that collegiate performance is the best predictor of success in the NBA. Biometric data collected at the NBA Combine does not have strong predictive power for future NBA player performance.

For future work, the exploration of additional biometric features that could be assessed at the NBA Combine (and are not currently measured) may yield additional insights into the biometric impact on player success in the NBA. Additionally, a comparison of the distribution of biometric parameters of successful NBA players versus the general population might yield insights into the distinction in biometric profiles of elite basketball players. Such insights would give early career recruiters, such as college recruiters, an edge in their recruiting strategy since they are not dealing with established, elite athletes.

References

1. e. a. Berri, David J., "The short supply of tall people: Competitive imbalance and the national basketball association," *Journal of Economic Issues*, pp. 1029–1041, 2005.
2. e. a. Berri, David J., "From college to the pros: predicting the nba amateur player draft," *Journal of Productivity Analysis*, pp. 25–35, 2010.
3. N. Paine, "Projecting the top 50 players in the 2015 nba draft class," Apr 2017.
4. J. Treme and R. Burrus, "Ncaa basketball: when does recruiting talent translate into wins for power conferences?," *Journal of Economics and Finance*, 2015.
5. J. R. Radzevick, "Does transition experience improve newcomer performance? evidence from the national basketball association," *Small Group Research*, p. 207–235, 2016.
6. O. release and NBA.com, "Nba salary cap set for 2017-18 season at 99.093 million," Jul 2017.
7. "Introduction to support vector machines."
8. K. Karkazis and J. R. Fishman, "Tracking u.s. professional athletes: The ethics of biometric technologies.," *American Journal of Bioethics*, vol. 17, no. 1, pp. 45 – 60, 2017.
9. M. Zimmer, ""but the data is already public": on the ethics of research in facebook," *Ethics and Information Technology*, no. 12(4), pp. 313–325, 2010.
10. R. Herschel and V. Miori, "Ethics & big data," *Technology in Society*, vol. 49, pp. 31 – 36, 2017.