An Improved Approach to Algorithmic Draft Selection for Teams in the National Basketball Association

Noah Rowe, Queen's University

Journal of Global Sport Management

Abstract

This work presents an algorithm to automate the National Basketball Association (NBA) draft selection process for a given team by making unbiased draft choices. Development of the algorithm is achieved by replicating and extending previous work in this area. Increased performance is achieved through an extended data set, accounting for the impact of existing rosters on draftee performance, and extensive model optimization. When implemented on the appropriate subset, the automated algorithm is able to provide on average of 1.14 additional annual expected wins per draft pick. This translates to over 3.4 million dollars in added value per draft pick, or just under an average of 7 million dollars annually.

1 Introduction

The National Basketball Association (NBA) is the second most profitable professional sports league in North America, generating $10.7 billion dollars in revenue over the 2018-2019 season [1]. This is supported by dividing total player annual salaries by the number of wins in a typical season, revealing that NBA teams pay approximately 3.02 million USD per win (using 2020 figures) [2]. The League provides monetary incentive the winningest teams. For example, following their championship-winning year, the total worth of the Toronto Raptors increased by 25%, as compared to the League average of 14% [1]. This highlights a strong financial incentive for each private club to create a winning team. A notable component of team development and refinement is the annual NBA player draft, where prospective players are recruited through a structured process involving all teams. This is a competitive and zero-sum process, with teams attempting to secure the best players.

These draft selection choices often define a club for many years. That is, drafting teams hold privileged rights over their draft picks for up to four years [3]. Teams will also be able to leverage these contracts in trading scenarios, allowing for further team development. All of this supports the need for a reliable process to secure talented draft picks, and provides evidence towards the fact that player personnel decisions have been cited as the primary function of League general managers [4].

It is understood that data-driven, objective analysis is required by NBA general managers to optimize player personnel choices; however, it is often difficult to exclude emotion and biases from the decision making process. For example, Berri, Brook, and Fenn [5] find that front offices disproportionately focus on offensive performance, giving rise to suboptimal player evaluation. This practice highlights the need for a data-driven approach to player recruitment, and helps to explain an influential trend in the NBA today. Specifically, what can be considered a data analytics revolution is unfolding across the League, with almost every team now hosting a data analytics department focused on this type of analysis [6].

Published work related to optimizing player recruitment decisions is currently available. For example, Maymin outlines a system for automating all aspects of player acquisition, including draft selections, free agent signings, and player trades [7]. The draft selection model is used as a foundation to develop an improved approach, and Maymin's results will serve in this work as a useful baseline to assess final model effectiveness. Based on data availability, this paper focuses only on player who played a minimum of one season of college-level basketball and at least three seasons in the NBA. This represents approximately two thirds of all draftees.

1.1 Development of an Automated System

The final model described in this paper should provide managerial staff on an NBA team with the ability to indiscriminately assess player acquisition opportunities in the absence of bias and human emotion. This can be achieved through the elimination of biases driven by non-relevant features (i.e. former NBA-player parents, physical features/appearance, sexual orientation, religion, etc.). The final algorithm should be able to recommend the most suitable player for a given team, while considering their existing roster.

This work will replicate the results described by Maymin in [7], and incorporate existing roster data to improve upon these results. Success of the final model is measured based on additional wins anticipated with use of the model, as opposed to historical choices.

2 Theoretical Background

2.1 Player Personal Decision Inefficiencies

There is a growing acceptance of data analytics within the sports world, including the NBA. Teams are using statistical methods for player recruitment, contract evaluation, injury prevention, and more [8]. However, despite the significant monetary impacts resulting from poor choices and the recent proliferation of statistical information regarding players, NBA owners and managers still struggle to make efficient draft selections [5]. Examples of this include placing too much emphasis on irrelevant information (college team playoff performance), and over-valuing the impact of scoring [5].

BLANK explicitly explores the existence and impact of an anchoring bias when evaluating pre-NBA player talent [9]. It was found that expert rankings have a disproportionately large impact on a given player's draft ranking, despite still being correlated with player success. This implies that, even though decision makers are attempting a quantitative analysis of a player based on expert opinions, they are unable to correctly balance the impact of the many available statistics.

In contrast to relying to heavily on a single statistic, drafting errors are often attributed to the decision maker thinking their domain knowledge allows them to ignore statistical trends [10]. Sailofsky explains that this can be related to Health and Tversky's (1991) [FIX CITE] competence hypothesis. Heath and Tversky's findings showed that decision makers choose to stake more on their own assessments when they consider themselves to have expert knowledge on the subject, often to their own detriment.

The goal of this work is not to highlight the specific inefficiencies currently present in NBA drafting decisions. The goal of this paper, similar to the intention of Maymin in [7], will be to provide an (improved) benchmark of which managerial decisions can be evaluated against. This model will not be able to indicate the specific areas where decision makers are lacking, but will act as a gauge to measure drafting productivity.

2.2 Machine Learning Techniques

Machine learning is an obvious direction to approach this problem from, due to the availability of data and the desire for a judgement-absent decision. This paper will use a neural network model for formulating draft predictions, as it was found to be the most effective model. Other models tested include k-nearest-neighbours [11], random forest models [12], and gradient boosting models [13].

2.2.1 Neural Networks

Neural networks are a type of supervised machine learning model designed to replicate the behaviour of a biological human brain [14]. These networks are initialized with neurons connected within various layers that perform mathematical operations, transforming an input vector to a final output. Neural networks have a complex parameter space and performance is highly dependent on the structure of the model. Model training can be computationally intensive, depending on model size. Neural networks are a type of black-box model that does not provide justification of its decisions. Due to the relatively small model size required for accurate predictions, and interpretability not being a required feature, we are not hindered by these negative traits in this application.

2.2.2 K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm that classifies data entries into k different categories, where k is selected by the user [15]. Within this application, k-means clustering is used to identify players similar to one another, according to their feature vectors. K-means clustering was chosen because of its computational speed and its capacity to specify the number of clusters in advance of implementation. It is also necessary to choose an algorithm which allows for the classification of new data without restructuring the original categorization model, ensuring that clustering remains consistent across all players.

2.2.3 Leave-one-out Training

When training a machine learning model, it is important that no entries in the testing data set are present in the training data; thus preventing the possibility of training bias, as all testing data points have never been seen by the model. Due to the relatively small amount of data used in this thesis, further limitation of the data size greatly reduces model accuracy. Leave-one-out (LOO) training is a type of training that seeks to maximize training data, while preventing testing exposure. For each entry in the training data, a unique model is trained on all other data entries. This model can generate predictions on the excluded data point with maximal training exposure, while preventing training bias. Figure 1 summarizes this process.

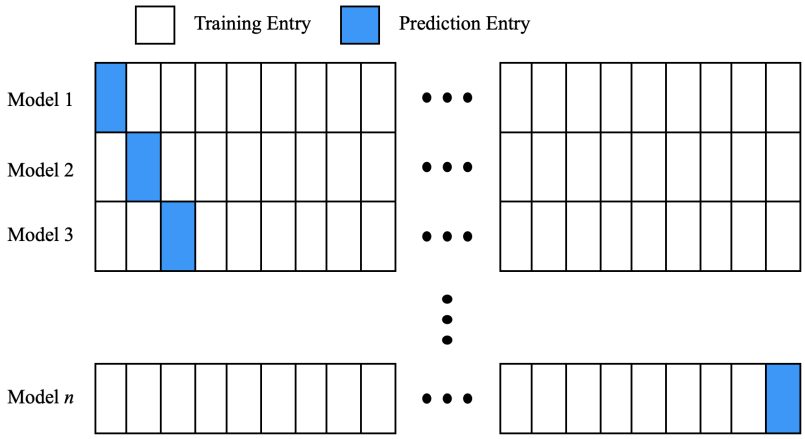

Figure 1: A diagram illustrating the leave-one-out training procedure. Each row represents an independent training session, resulting in n different models.

3 Method

This section describes the approach used to rank players of the same draft class. Performance predictions are generated for each player, using various statistics related to player biometrics and college performance. Model results are then verified on historical data using a mock draft simulation that allows for the quantification of the benefit in using the model. Figure 2 explains how the two raw data sets are processed, combined, predicted upon, and evaluated using a flow-style diagram.

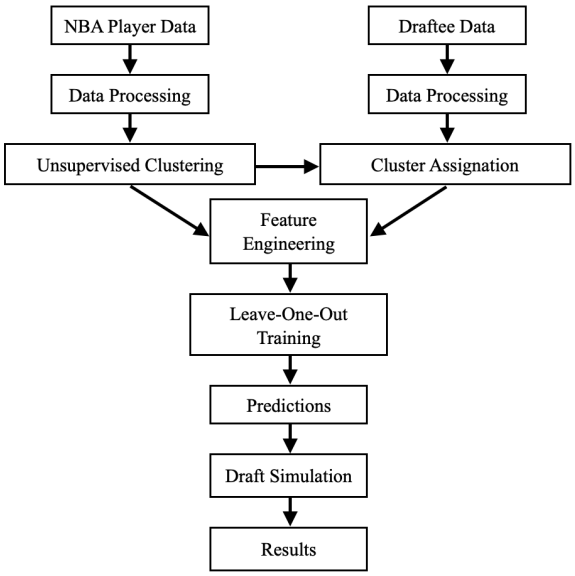

Figure 2: A visual summary of the steps taken to transform the raw data into feature vectors and generate model output.

3.1 Data Collection

Data for this project were collected using web-scraping software, as well as from published data sets. All data scraping was conducted using Octoparse, a commercial software [16]. Data related to draftees was collected from multiple sources [17, 18, 19, 20, 21], and historical NBA data was downloaded from the Kaggle data set "NBA Players stats since 1950" [22], containing player stats from 1950 to 2017. Table 1 shows the list of features used to create the input vectors, and Table 2 outlines the statistics used to develop a target variable. Draftee data spans from 2000 to 2017, and all NBA data prior to 2000 is excluded from the analysis, as it is deemed to be less relevant as compared to more recent data. Removing players with fewer than 3 years of data resulted in a data set containing 655 draftees. Mean imputation was used to address missing values.

Table 1: A summary of the predictive features collected. Values denoted with * are also contained within the Kaggle data set "NBA Player stats since 1950" and represent the same statistic, but applied to NBA games. All college basketball statistics are averaged over each season played.

Data Point	Description	Source
Mock	Drafting order from mock drafts performed by ESPN.com	ESPN.com
RSCI	Ranking of the player when they graduated high school	Draftexpress.com
Awards	The number of NCAA awards won by the player	Basketball.realgm.com

Table 1 – continued from previous page

Statistic	Description	Source
SOS	A measure of how difficult the player's college schedule was	Basketball.realgm.com
Position*	Player position	NBA.com
Wingspan	Player wingspan (NBA combine metric)	NBA.com
Reach	Player standing reach (NBA combine metric)	NBA.com
Body fat %	Player body fat percentage	NBA.com
Hand width	Player hand width	NBA.com
Hand length	Player hand length	NBA.com
Height*	Player height	NBA.com
Weight*	Player weight	NBA.com
Leap	Maximum vertical leap (NBA combine metric)	NBA.com
Speed	Sprint speed performance (NBA combine metric)	NBA.com
Agility	Lane agility performance (NBA combine metric)	NBA.com
BPR	Bench press repetitions performed (NBA combine metric)	NBA.com
GP	Total games	Basketball-reference.com
MP	Total minutes	Basketball-reference.com
P*	Total points	Basketball-reference.com
A*	Total assists	Basketball-reference.com
STL*	Total steals	Basketball-reference.com
BLK*	Total blocks	Basketball-reference.com
TOV*	Total turnovers	Basketball-reference.com
OBR*	Total offensive rebounds	Basketball-reference.com
DBR*	Total defensive rebounds	Basketball-reference.com
eFG%*	Effective field goal percentage	Basketball-reference.com
ORB%*	Offensive rebound percentage	Basketball-reference.com
DRB%*	Defensive rebound percentage	Basketball-reference.com
AST%*	Assist percentage	Basketball-reference.com
TOV%*	Turnover percentage	Basketball-reference.com
STL%*	Steal percentage	Basketball-reference.com
BLK%*	Block percentage	Basketball-reference.com
USG%*	Usage percentage	Basketball-reference.com
PER	Player efficiency rating	Basketball-reference.com
OR	Dean Oliver's offensive rating	Basketball-reference.com
DR	Dean Oliver's defensive rating	Basketball-reference.com
OWS*	Offensive win shares	Basketball-reference.com
DWS*	Defensive win shares	Basketball-reference.com
3PA*	Total three point shot attempts	Basketball-reference.com
3P%*	Three point shot percentage	Basketball-reference.com
FTA*	Total free throw attempts	Basketball-reference.com
FT%*	Free throw percentage	Basketball-reference.com
FGA*	Total field goals attempted	Basketball-reference.com
PF*	Total personal fouls	Basketball-reference.com
Combine	Whether the player completed the NBA combine	Basketball-reference.com

Table 2: Statistics used to generate the model target variable. See source locations for variable derivations.

Statistic	Description	Source
Win shares	Win shares earned by the player per NBA season	Basketball-reference.com
Wins produced	Estimated wins produced by the player per NBA season	boxscoregeeks.com
Estimated wins added	Estimated wins added by the player per NBA season	insider.espn.com

3.2 Data Manipulation

Draftee specific data and historic NBA data are separately used to create player-specific features. The raw draftee statistics were used to build each feature vector and new features were derived to help a model identify important measures. The position feature was separated into three columns: Center, Forward, and Guard, indicating what positions a draftee can play. The mock draft feature represents 6 different columns, each denoting different mock evaluations. The statistics shown in Table 1, along with the engineered features described in Table 3, comprise a length 55 base feature vector unique for each draftee in the data set.

Table 3: A summary of the derived statistics, generated from the original features. These features were generated for both the NBA data set and the draftee data.

Derived Statistic	Description
Points per field goal attempt	Total points divided by total field goal attempts
Points per game	Total points divided by total games played
Points per minute played	Total points divided by total minutes played
Minutes played per personal foul	Total minutes played divided by total personal fouls
Free throw attempt per field goal attempt	Total free throw attempts divided by total field goal attempts
Minutes played per three point attempt	Total minutes played divided by total three point shot attempts
Assist per turnover	Total assists divided by total turnovers

3.2.1 Drafting Team Features

In addition to the base feature vector, two sets of drafting team-specific features were created for each player, both focused on comparing the draft candidate to the current roster of the drafting team. One set focused on comparing the draft candidate to the starting roster of the drafting team, and the other set focused on comparing the draftee to similar-styled players on the drafting roster. Starting players for each roster were assumed to be the five players with the greatest number of minutes played. These features are summarized in Table 4.

Table 4: A description of the features specific for each drafting team. Two versions of these features (excluding label count) were created for each draftee: one calculated over the five players with the most minutes (starting players) on the drafting team, and one over the subset of players belonging to the same label as the draftee, resulting in 11 new features. Distance metrics are passed as a model parameter.

Feature	**Description**
Average distance	The average distance value between the draftee and each player in the subset
Distance dot minutes	The dot product of the distance value between the draftee and a given player in the subset, and that players respective total minutes played
Distance dot win shares	The dot product of the distance value between the draftee and a given player in the subset, and that players respective win shares
Minimum distance	The minimum distance metric value between the draftee and each player in the subset
Minimum distance win shares	The win shares of the player from the subset with the minimum distance value between them and the draftee
Label count	The number of players on the drafting teams roster belonging to the same cluster as the draftee

The subset of similar styled players was determined from the results of k-means clustering. The algorithm was trained on NBA players and applied on the draftee data set to generate a label for each draft candidate. Draftee labels can be compared to the labels of the drafting roster to identify players who accomplish similar roles. Figure 3 shows the results of k-means clustering applied to NBA player statistics with three clustering groups, summarized on a subset of the features.

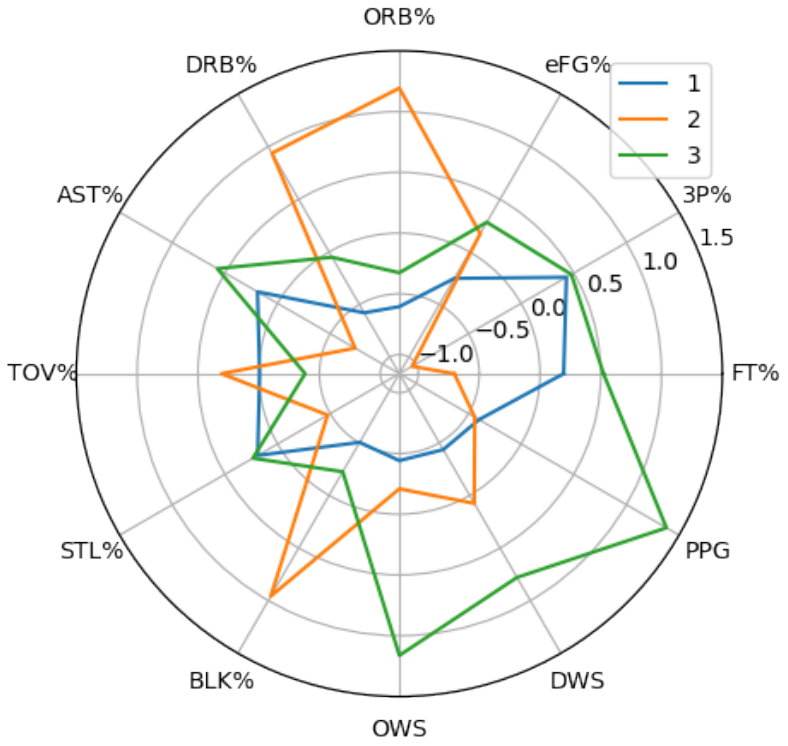

Figure 3: A plot demonstrating the differences between the three clusters generated by k-means clustering. Each line shows the average values of a cluster, normalized using standard scaling. See Table 1 for variable descriptions.

Figure 3 demonstrates three distinct clusters containing players of varying skill sets and abilities. Cluster 1 contains below-average players (presumably bench players) with no dominant skill set. Cluster 2 contains players with a focus on rebounding and blocks, while lacking in shooting ability (most likely centers). Cluster 3 shows scoring-focused players that are talented shooters and passers. The cluster sizes were 3203, 1586, and 2061 respectively. This is important, as clusters should be well distributed in order to have sufficient samples on NBA rosters. Rosters missing a cluster type have features set to the data set minimum for that particular feature.

3.2.2 Model Target

As the goal of this thesis is to develop an algorithm to identify the highest potential draftees, it is necessary to quantify the contribution a player provides to their team. This is done using the wins made metric, as defined by Maymin in [7] using Equation 1:

$$WinsMade = \frac{WS + WP + EWA}{3} \tag{1}$$

where WS is win shares, WP is wins produced, and EWA is earned win average. These variables and their sources are detailed in Table 2. Wins made is a measure of the number of wins a player is responsible, calculated by averaging the output of three metrics designed to quantify the same. Wins made is the target variable each machine learning model is designed to predict.

3.3 Model

The core model used to generate predictions for each draftee is a densely connected artificial neural network (ANN). The core model used to generate predictions for each draftee was trained using LOO training. This resulted in 655 unique models, each one excluding a specific player from the training data. These models could then be used to predict on the excluded entry using the most training data possible. Therefore, after training, each draftee has an associated model prediction.

3.3.1 Model Validation

In order to validate the model, a simulation was designed to replicate an actual draft with a single team utilizing model predictions to make their choices. A drafting simulation allows for the conversion of model predictions to actual team benefit. This is important, as model error (in the form of mean squared error) is not representative of model effectiveness, as the purpose of this thesis is not to predict draftee performance, but rather to develop a drafting algorithm that maximizes drafting efficiency for the controlling team. That is, the model presented here is designed to identify the highest potential draft candidate, rather than to predict the exact performance of a given candidate. This work uses the results of a drafting simulation (added wins made per draft pick) as the dominant validation metric.

For a given draft, a simulation is performed by controlling a single team's choices with selections generated by the model. For teams not utilizing the model, it is assumed that they select the draftee that was chosen earliest, out of the remaining players. This is not entirely realistic, as different teams prioritize players differently, but it is a necessary assumption in the context of this work. Teams controlled by the model select the player with the highest predicted wins made, and drafting order in the simulation reflects that which is found in the actual draft process. When the simulation is complete, added wins is calculated as the difference between the original and new wins made of a teams draft choices.

4 Results

The predictive model was implemented on the data set 30 times to assess its effectiveness, and the results are summarized in Figures 4 and 5. Figure 4 illustrates the comparison between current draft selections and model outputs using added wins per draft pick. Figure 5 shows the average benefit for each team.

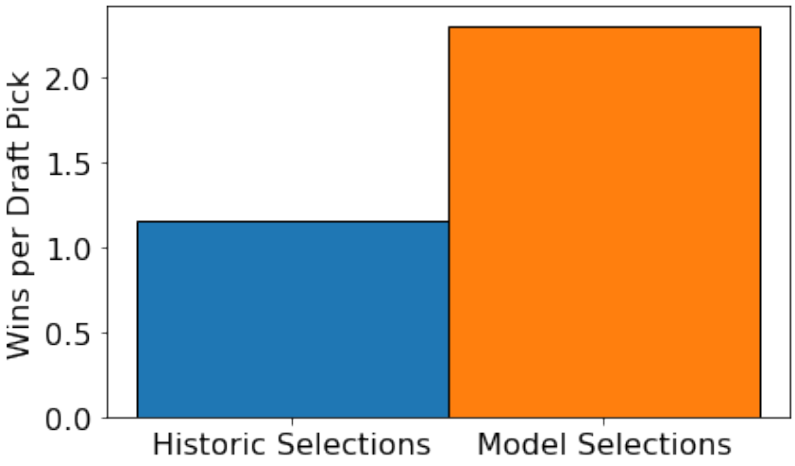

Figure 4: A bar graph comparing the average wins per draft pick from current selection habits (left) and the models draft choices (right).

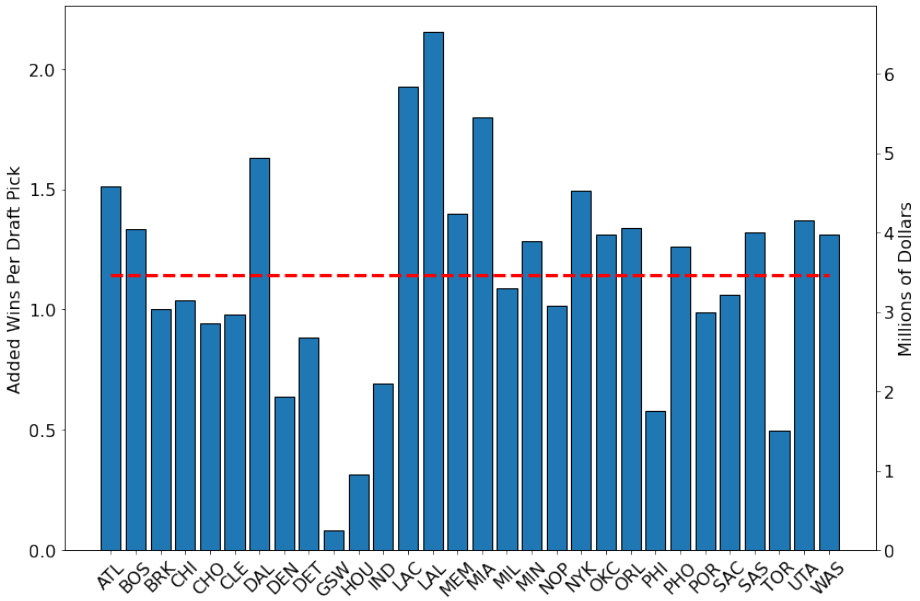

Figure 5: A bar graph showing the average added wins per draft pick for each team in the NBA. The left y-axis shows results in terms of average added wins. The right y-axis shows results in terms of the average added monetary value per draft pick, in millions of dollars. The red dotted line represents the average value across all teams: 1.1 +/- 0.5 in average added wins per pick or 3 +/- 1 in added millions of dollars per draft pick.

5 Discussion

The results shown in Figure 5 demonstrate that current draft selection methods implemented by NBA clubs are inefficient, with each selection costing teams millions of dollars annually. It is estimated that utilization of this model can provide an extra 3.4 million dollars per draft pick when compared to current methods. Therefore, as each club has an average of 2 draft picks per season, current annual losses are estimated at almost 7 million dollars annually.

The model introduced in this work outlines a strategy to remove human bias (that can interfere with sound decision making) from the NBA draft selection process, while doubling the efficiency of draftee selections by NBA teams within the simulation timeline (2000-2017). The model discussed in this paper provides an additional 1.14 wins per draft pick, translating to an added 3.4 million dollars of value per pick. This represents an improvement on work in the existing literature, which is estimated to provide approximately 1 additional win per draft pick. Because the current model is unable to assess certain groups of draftees, it should be deployed in situations similar to those in which it was evaluated. Overall, the results prove that this model can provide significant benefit to potential users. Further, these results can serve a benchmark for future work in this area.

References

[1] The business of basketball: Forbes releases 22nd annual nba team valuations. https://www.forbes.com/sites/forbespr/2020/02/11/the-business-of-basketball-forbes-releases-22nd-annual-nba-team-valuations/?sh=4525618675ff, 2020. Accessed: 2021-02-28.

[2] Nba salaries. https://hoopshype.com/salaries/. Accessed: 2020-03-20.

[3] A. Jessop. The structure of nba rookie contracts. https://www.forbes.com/sites/aliciajessop/2012/06/28/the-structure-of-nba-rookie-contracts/?sh=5c029b347299, 2012. Accessed: 2021-01-10.

[4] G. M. Wong and C. Deubert. National basketball association general managers: An analysis of the responsibilities, qualifications and characteristics. *Villanova Sports and Entertainment Law Journal*, 18, 2011.

[5] S. L. Brook D. J. Berri and A. J. Fenn. From college to the pros: Predicting the nba amateur player draft. *Journal of Productivity Analysis*, 35:25–35, 2011.

[6] M. Mudric. How the nba data and analytics revolution has changed the game. https://www.smartdatacollective.com/how-nba-data-analytics-revolution-has-changed-game/. Accessed: Fall 2020.

[7] P. Z. Maymin. The automated general manager: Can an algorithmic system for drafts, trades, and free agency outperform human front offices? *Journal of Global Sport Management*, 2:234–249, 2017.

[8] University of Pennsylvania. The nba's adam silver: How analytics is transforming basketball. https://knowledge.wharton.upenn.edu/article/nbas-adam-silver-analytics-transforming-basketball/, 2017. Accessed: 2021-02-28.

[9] T. Berger and F. Daumann. Anchoring bias in the evaluation of basketball players: A closer look at nba draft decision-making. *Managerial and Decision Economics*, pages 1248–1262, 2021.

[10] D. Sailofsky. Drafting errors and decision making theory in the nba draft. 2018.

[11] Z. Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4.

[12] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[13] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38:367–378, 2002.

[14] D. J. Livingstone, D. T. Manallack, and I. V. Tetko. Data modelling with neural networks: Advantages and limitations. *Journal of Computer-Aided Molecular Design*, 11, 1997.

[15] T. Kanungo et al. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:881–892.

[16] Octoparse web scraping software. https://www.octoparse.com/, 2020. Accessed: Fall 2020.

[17] National basketball association website. NBA.com.

[18] Real general manager website. Basketball.realgm.com.

[19] Draftexpress website. Draftexpress.com.

[20] Basketball reference website. Basketball-reference.com.

[21] Entertainment and sports programming network website. ESPN.com.

[22] O. Goldstein. Nba players stats since 1950. https://www.kaggle.com/drgilermo/nba-players-stats, 2018. Accessed: 2020-01-10.