UPPSALA
UNIVERSITET

Predicting basketball performance based on draft pick
A classification analysis

Fredrik Harmén

Bachelor's thesis in Statistics

Advisor

Johan Lyhagen

2022

Abstract

In this thesis, we will look to predict the performance of a basketball player coming into the NBA depending on where the player was picked in the NBA draft. This will be done by testing different machine learning models on data from the previous 35 NBA drafts and then comparing the models in order to see which model had the highest accuracy of classification. The machine learning methods used are Linear Discriminant Analysis, K-Nearest Neighbors, Support Vector Machines and Random Forests. The results show that the method with the highest accuracy of classification was Random Forests, with an accuracy of 42%.

Table of content

1. Introduction

In the year 1946 the National Basketball association (NBA) was founded through a merger of two rivaling basketball leagues, the national basketball league and the basketball association of america.[1] At this time, the NBA consisted of a total of 17 different organizations with a team each.[2] In the early 80's, the NBA was troubled by money-losing franchises with low attendance and low tv-ratings. However this changed under the leadership of David Stern, who became NBA commissioner 1984. David Stern is credited to have helped the NBA transform into an international entertainment company through his marketing strategies and innovations.[3] The NBA has grown ever since, expanding into a 30 team league[4], accumulating fans all over the world and also becoming increasingly popular domestically in the United States to this day.[5]

At the end of each NBA season the NBA draft process takes place. The NBA draft process consists of different events such as, the NBA draft lottery, the NBA draft combine followed by the NBA draft. Players that enter the NBA draft are players that have not previously played in the NBA, such as former college players or international players from other basketball leagues around the world. The NBA draft is structured in a specific way, in order to bring balance to the NBA in terms of talent accumulation. Thus, at the end of each NBA season the worst teams are most likely to pick higher in the draft, that is, the worst team has the best opportunity to choose the best incoming talent for the next season.[6]

---

[1] Britannica, The Editors of Encyclopedia, *National Basketball Association*. *Encyclopedia Britannica*, 22 July 2021, https://www.britannica.com/topic/National-Basketball-Association, Accessed 2 May 2022.

[2] NBA, *A Chronology of the Teams in the NBA,* https://www.nba.com/celtics/history/nba-teams-chronology#:~:text=On%20June%206%2C%201946%2C%20the,Louis%20the%20West%20Division, Accessed 2 May 2022.

[3] See note 1.

[4] NBA, bhussey, *Charlotte Bobcats Become NBA´s 30th Team, 6 May* 2004, https://www.nba.com/hornets/news/official_team.html, Accessed 2 May 2022

[5] wbhsbullseye, D. Rudd & N. Solomon, *The Rise of the NBA,* 14 March 2019, https://wbhsbullseye.com/808/sports/the-rise-of-the-nba/, Accessed 2 May 2022.

[6] BasketballNoise, James, *How does the NBA draft work?*, 23 October 2021, https://basketballnoise.com/how-does-the-nba-draft-work/, Accessed 2 May 2022.

The first part of the draft process is the NBA draft lottery. The NBA draft lottery is an event where the selection order of the NBA draft is decided. The lottery consists of teams that did not make the playoffs, accumulating to 14 out of 30 total teams. The lottery system was first introduced in 1985 and has been continuously modified, with the latest modification being implemented in 2019. The lottery is currently constructed in such a way that the worst three teams have a 14% chance each at the number one overall pick, while giving the fourth worst team a 12.5% chance at the number one overall pick.[7] The odds of getting the number one overall pick gets lower and lower the further down in the lottery a team is.[8] The main reason for the latest modifications of the NBA draft lottery is to prevent teams from losing on purpose in order to get a higher in the draft, losing on purpose is considered to be bad for professional sports organizations in general. Reasons include, being dishonest to the fans as well as being embarrassing for the cities funding their local sports teams.[9]

After the NBA draft lottery is an event called the NBA draft combine. The combine is an event where the teams get to see the draft prospects perform different types of tests to demonstrate their skills live, take different types of body measurement tests as well as conducting interviews for the upcoming draft. Examples of tests include shooting drills, lane agility drills and bench press competitions.[10] Even though a prospect might have had a great college career and done really well to prove himself at the NBA draft combine, there is still a whole lot of uncertainty of how good of an NBA player a prospect might become. Even with the best scouting tools in the world, there are so many variables that go into a prospect becoming a successful NBA player.[11] A common way to get a sense of the risk / reward for a

---

[7] Sports Illustrated, H.Beck, *The Tanking Era As We Know It Is Over*, 26 April 2021, https://www.si.com/nba/2021/04/26/tanking-era-is-dead-play-in-tournament, Accessed 2 May 2022.

[8] NBA, NBA.com Staff, *NBA Draft Lottery: Odds, history and how it works,* 15 April 2022, https://www.nba.com/news/nba-draft-lottery-explainer, Accessed 2 May.

[9] The Colgate Maroon-News, J. Adams & Maroon-News Staff, *Tanking in the NBA: Good or Bad For the League?,* 11 February 2016, https://thecolgatemaroonnews.com/4167/sports/nattysports/tanking-in-the-nba-good-or-bad-for-the-league/, Accessed 2 May 2022.

[10] See note 6.

[11] Insider, F. Zaccaneli, *Former Mavericks president and GM explains the risks and rewards of evaluating NBA talent*, 20 June 2017, https://www.businessinsider.com/evaluating-nba-players-draft-2017-6?r=US&IR=T, Accessed 2 May 2022.

certain player is to do comparisons to former players, in terms of floor / ceiling. The floor being the expected performance of a prospect if he does not develop properly and ceiling being the best possible player a prospect might become.[12]

The final part of the NBA draft process is the actual NBA draft, where each team has an opportunity to pick a player at their given post-lottery position. The NBA draft consists of two rounds with 30 picks in each round. Where the position in the second round is the same as the 1st round, that is, the team who gets the first overall pick also gets the 31st pick.[13] Which prospect a specific team chooses might depend on a lot of different variables.
A common way teams evaluate a prospect is in terms of potential fit and/or best player available (bpa). If a team already has a good player at a certain position, they might try to fill another position on their roster with a prospect through the draft. Or teams might simply decide to pick the prospect that they believe has the most talent, regardless of their own team's current roster structure.[14]

As one might be able to tell, there are a lot of variables that go into the evaluation of an NBA draft pick. And the weight one might put on different variables might differ significantly depending on what team is making the decision. In order to simplify the evaluation of a prospect, I have created a model that attempts to quantify the value of an NBA draft pick in terms of the average win-shares added per season depending on where a prospect is selected in the NBA draft.

The purpose of this thesis is to try to predict the amount of win-shares that NBA players contribute to throughout their career depending on where they were picked in the NBA draft. This will be done by testing different machine learning models, then comparing them to one another to see which one has the highest accuracy of classification. The machine learning

[12] Front Office Gurus, M. Feldman, *2020 NBA DRAFT: FLOOR AND CEILING COMPARISON´S FOR EACH TOP PROSPECT*, 19 May 2020, https://frontofficegurus.com/2020/05/19/2020-nba-draft-floor-and-ceiling-comparisons-for-each-top-prospect/, Accessed 2 May 2022

[13] ESPN, ESPN.com, *2022 NBA draft order: Complete picks for the first and second rounds ahead of the lottery*, 20 April 2022, https://www.espn.com/nba/story/_/id/33748170/2022-nba-draft-order-complete-picks-first-second-rounds-ahead-lottery, Accessed 2 May 2022.

[14] See note 12.

models that will be used are Linear Discriminant Analysis, K-Nearest Neighbors, Support Vector Machines as well as Random Forests. The goal is to be able to evaluate an NBA draft pick in a certain position the same way one would evaluate the expected performance of an existing player. By connecting a certain class of win-shares to each draft pick, the hope is that one might be able to use this information to get a clearer understanding of the value of a draft pick. The research question is as follows:

Can we predict how many win-shares NBA basketball players contribute to depending on where they were picked in the NBA draft? Which of the models has the best accuracy?

The remainder of this thesis is structured in the following way. In Section 2 the data is introduced and the variables used are explained. In Section 3 the theory behind the machine learning models are presented as well as the methodology. The results are presented and discussed in Section 4 and the conclusion is given in Section 5.

2. Data

2.1 Data overview

The data for this thesis was collected from www.stathead.com/basketball.[15]
The dataset consists of observations of all players that have been drafted during the NBA draft lottery era, which ranges from the 1985 NBA draft to the latest NBA draft in 2021. The dataset consisted of 2162 observations and 26 variables before the data washing and 1599 observations and 2 variables after the data washing. In the cleansed dataset 563 observations were removed because some players had not played a full single NBA season in their career. The reasoning for this was simply that the data for such a player was deemed to be too small. At the date of this report, the 2022 NBA season is still in process, and thus players drafted in 2021 have been removed from the dataset.

2.2 Variables

In order to answer the first part of the research question;
"Can we predict how many win-shares NBA basketball players contribute to depending on where they were picked in the NBA draft?", two variables were used, Pick and Win-shares per season.

The variable Pick is a variable that measures with which pick each of the players were selected. These variables can assume values from 1 to 60, as there are 60 picks in the NBA draft.

Win-shares is a measurement of individual basketball performance that measures how many wins a player adds to a team. The main idea behind this measure is to divide the credit for a teams´ performance into individual parts. Thus, the sum of all individual win-shares on a team roughly equals a teams´ total amount of wins during a season.[16]

The Win-share formula takes the following form

[15] https://stathead.com/tiny/xA93X, Accessed 20 April 2022.

[16] Basketball Reference, *NBA Win Shares,* https://www.basketball-reference.com/about/ws.html, Accessed 20 April 2022.

Win-shares = Offensive Win-shares - Defensive Win-shares.

(1)

In order to get offensive win-shares, one must perform the following steps,

1. *Calculate points produced by each player*
2. *Calculate offensive possessions for each player*
3. *Calculate the marginal offense for each player*
4. *Calculate marginal points per win*
5. *Credit offensive win-shares to the players*

1. *Calculating points produced by each player.*

This measure can be described as the points produced by an individual player per 100 possessions. Points produced by each player has the following formula

$$Points\ produced\ by\ each\ player\ =\ (Field\ goals\ attempted\ +\ 0.44\ *\ Free\ throws\ attempted\ +\ turnovers)\ *\ (Offensive\ Rating)\ /\ 100, \quad (2)$$

where the 0.44 multiplier is included because not all free throws take up an entire possession.[17]

2. *Calculate offensive possessions for each player.*

An offensive possession can be simplified as when a player on a team does something that ultimately ends an offensive possession. For example, shoots and misses so that the other team gains possession of the basketball.[18] Offensive possessions for each player has the following formula,

$$Offensive\ possessions\ for\ each\ player\ =\ 0.96\ *\ [(Field\ Goal\ Attempts)\ +\ (Turnovers)\ +\ 0.44\ *\ (Free\ Throw\ Attempts)\ -\ (Offensive\ Rebounds)], \quad (3)$$

[17] See note 16.

[18] NBA Stuffer, *Possession*, https://www.nbastuffer.com/analytics101/possession/, Accessed 21 April 2022.

where the 0.96 multiplier is included to account for a continuation of a possession without an offensive rebound being credited, for example, a missed shot goes off a defensive player out of bounds.[19]

3. *Calculate the marginal offense for each player*

The marginal offense for each player can be described as the total points scored by a certain player "A" compared to the rest of the NBA players in the league, if these players would have the same amount of opportunity (possessions) as player "A".
The marginal offense for each player has the following formula

$$Marginal\ offense\ for\ each\ player\ =\ (points\ produced)\ -\ 0.92\ *\ (league\ points\ per\ possession)\ *\ (offensive\ possessions), \quad (4)$$

where the 0.92 multiplier is included to standardize the average points per possession.[20]

4. *Calculate the marginal points per win*

The marginal points per win standardizes a player´s offensive contributions by adjusting for that player´s team´ s pace of play compared to the rest of the league. The marginal points per win has the following formula

$$Marginal\ offense\ for\ each\ player\ =\ 0.32\ *\ (league\ points\ per\ game)\ *\ ((team\ pace)\ /\ (league\ pace)), \quad (5)$$

*w*here the 0.32 multiplier is included to standardize the average points per game. [21]

5. *Credit offensive win-shares to the players*

Lastly we credit the offensive win-shares to each player by using the following formula,

Offensive win-shares = marginal offense / marginal points per win.[22] (6)

In order to get defensive win-shares, one must go through the following steps,

[19] See note 16.

[20] See note 16.

[21] See note 16.

[22] See note 16.

1. *Calculate the defensive rating for each player*
2. *Calculate the marginal defense for each player*
3. *Calculate marginal points per win*
4. *Credit defensive win-shares to the players*

1. *Calculate the defensive rating for each player*

The defensive rating for each player can be described as the amount of points allowed by an individual player per 100 defensive possessions while adjusting for a player´s team´s defense. Defensive rating revolves around the concept of an individual's defensive stops, which can be explained as a defensive player ending the opposing team's offensive possession. The calculation of defensive rating for each player has the following formula

$$Defensive\ rating\ for\ each\ player\ =\ Team\ defensive\ rating\ +\ 0.2\ *\ (100\ *\ defensive\ points\ per\ scoring\ possession\ *\ (1\ -\ Stop\%)\ -\ Team\ defensive\ rating) \quad (7)$$

[23],

where the 0.2 multiplier symbolizes that a team's defensive rating increases by 0.2 points for each additional percentage of usage for the individual player per 100 possessions.[24]

2. *Calculate the marginal defense for each player*

The marginal defense for each player can be described as the increase of a teams´ defensive ability while a certain player is on the court. The calculation of marginal defense for each player has the following formula

$$Marginal\ defense\ for\ each\ player\ =\ (player\ minutes\ played\ /\ team\ minutes\ played)\ *\ (team\ defensive\ possessions)\ *\ (1.08\ *\ (league\ points\ per\ possession)\ -\ ((Defensive\ Rating)\ /\ 100)). \quad (8)$$

[25]

3. *Calculate marginal points per win*

[23] Basketball Reference, *Calculating Individual Offense and Defensive Ratings,* https://www.basketball-reference.com/about/ratings.html, Accessed 21 April 2022.

[24] Sonics Central, K. Pelton, *The WARP Rating System Explained,* June 2021, http://sonicscentral.com/warp.html, Accessed 21 April 2022.

[25] See note 16.

This calculation is the same as for the offensive win-shares. Where marginal points per win standardizes a player´s offensive contributions by adjusting for that player´s team´ s pace of play compared to the rest of the league. The marginal points per win has the following formula

$$Marginal\ offense\ for\ each\ player\ =\ 0.32\ *\ (league\ points\ per\ game)\ *\ ((team\ pace)\ /\ (league\ pace)), \quad (9)$$

*w*here the 0.32 multiplier is included to standardize the average points per game.[26]

4. *Credit defensive win-shares to the players*

Lastly we credit the defensive win-shares for each player by using the following formula,

Defensive win-shares = (marginal defense) / (marginal points per win).[27] (10)

The variable Win-shares per season was not in the original dataset and thus had to be coded. The original dataset included a variable named WS (win-shares). This variable included each player's total accumulated win-shares over their careers. In order to code this variable into Win-shares per season, the WS variable was divided by the career length of each player, thus getting each player's average Win-shares per season. The lowest observed value of this variable was -1.5 and the highest 13.67. This variable was split up into the prior probabilities by using the percentiles of the observations for this variable and is presented in Table 1 below. These percentiles being “0-25th percentile","25-50th percentile","50-75th percentile","75-100th percentile”. With 75-100th percentile meaning that the average expected winshare of a player is in the 75-100th percentile compared to the rest of the data.

Table 1. *Prior probabilities of Win-shares per season*

Classes	**0-25th percentile**	**25-50th percentile**	**50-75th percentile**	**75-100th percentile**
Number of observation	421	380	403	395
Percentage of observations	0.263	0,238	0,252	0.247

[26] See note 16.

[27] See note 16.

3. Theory

In this section the machine learning methods used in the thesis will be explained.

When building machine learning models one splits up the data in testing and training data in order to first build the model and then test the model on the remaining unseen data. However if one were to simply split the data down the middle, chances are that the model wouldn't do especially well since there might be underlying patterns that will be missed by the model. One method to fix this problem is through Cross-validation. Cross-validation is a commonly used method when building machine learning models. Cross-validation offers the opportunity to estimate the model performance on the unseen data not used while training the model. In this thesis, we will specifically look at k-fold cross validation.[28]

K-fold cross validation randomly splits the dataset of observations into k-groups, also called folds, of roughly the same size. The first fold is treated as the testing set, also called the validation set, and the model is fit on the remaining k - 1 folds. Then the mean squared error, MSE_1, is calculated on the observations from the validation fold. This process is repeated k-times and on every repetition, a different fold is used as the validation fold. This procedure results in k number of estimations of the test error, MSE_1, MSE_2,..., MSE_k. The k-fold cross validation estimate is then calculated by taking the average of these values,

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i. \quad [29] \qquad (11)$$

The data in this thesis was split into training and testing sets using a 10-fold cross validation approach. That is, the 1599 observations remaining from the cleansed dataset was split into k = 10 folds with roughly 159 observations in each fold.

[28] Towards Data Science, M. Alhamid, *What is Cross-Validation?*, 24 December 2022, https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75, Accessed 4 May 2022.

[29] Springer Texts in Statistics, , G. James Et al., *An Introduction to Statistical Learning with Applications in R*, 2013, https://link-springer-com.ezproxy.its.uu.se/content/pdf/10.1007/978-1-4614-7138-7, Accessed 4 May 2022.

3.1 Bayes´ classifier

Bayes´classifier is a classification method that assigns each observation to the most likely class, given its predictor values. That is, a test observation with x_0 is assigned to the class *j* for which

$$\Pr(Y = j|X = x_0) \tag{12}$$

is the largest. This classifier is called the Bayes´ classifier. In theory one would always prefer to predict qualitative responses using Bayes´classifier. However, for real data the conditional distribution of Y given X is unknown which makes the calculation of Bayes classifier impossible. That being said, other methods such as K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) attempts to estimate the conditional distribution of Y given X, and later classify an observation the the class with the highest estimated probability.[30]

3.2 Linear Discriminant Analysis

Linear Discriminant Analysis is a linear model that can be used for regression and classification problems. First formulated in 1936 by Fisher, it was solely a model for two classes. The model was later generalized for multiple classes in 1948, by C.R Rao.[31]

In Linear Discriminant Analysis, the objective is to predict group memberships into already known groups. That is, the true class membership of the observations in the data is already known when training a model. Once a model has been trained, it can be used in order to classify new observations where their true class membership is not known.

The way LDA approximates the Bayes classifier differs depending on how many predictor variables that are used. Since only one predictor variable was used for classification in this thesis, the LDA method approximates the Bayes classifier by plugging estimates for π_k, μ_k , and σ^2 into,

[30] See note 29.

[31] Analytics Vidhya, SK. Dash, *A Brief Introduction to Linear Discriminant Analysis*, 18 August 2021, https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/, Accessed 4 May 2022.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{13}$$

in particular, the following estimates are used,

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{14}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \tag{15}$$

where n is the total number of training observations, and n_k is the number of training observations in the kth class. The estimate for μ_k is simply the average of all the training observations from the kth class, while σ^2 can be seen as a weighted average of the sample variances for each of the K classes. Sometimes we have knowledge of the class membership probabilities $\pi_1,...,\pi_K$, which can be used directly. In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class. In other words,

$$\pi^k = n_k/n. \tag{16}$$

The LDA classifier plugs the estimates given in (14) and (15) into (13), and assigns an observation X = x to the class for which,

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \tag{17}$$

is largest.[32]

LDA has two main assumptions;

[32] See note 29.

- The data is multivariate normally distributed
- The covariance matrix of x_i is equal within all k-classes.

In order for LDA to be a good approximation of Bayes´classifier, these assumptions must be fulfilled. Unfortunately, these assumptions were not fulfilled for the data in this thesis, however even if this means that LDA is not deemed a good approximation of Bayes´classifier, the method can still be used for classification.[33]

3.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a machine learning model that can be used for both regression and classification problems. The main idea of the KNN algorithm is that it assumes that similar things exist in close proximity to one another.[34]

Given a positive integer K and a test observation x_0, the K-Nearest Neighbors firstly identifies the neighbors K points in the training data that are closest to x_0, which in turn is represented by N_0. Secondly, the KNN classifier estimates the conditional probability for class *j* as the fraction of points in N_0 whose response values equal *j*:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j) \tag{18}$$

Lastly , the KNN classifier applies Bayes´ classifier in order to classify the test observation x_0 to the class with the largest probability. [35]

3.4 Support Vector Machines

Developed by the computer science community in the 1990s, the Support Vector Machine (SVM) is an approach for classification problems. The SVM is often considered to be one of

[33] Towards Data Science, R. Gotesman, *How Did Linear Discriminant Analysis Get Its Name?*, 2 July 2019, https://towardsdatascience.com/mathematical-insights-into-classification-using-linear-discriminant-analysis-9c822ad2fce2, Accessed 4 May 2022.

[34] Towards Data Science, O. Harrison, *Machine Learning Basics with the K-Nearest Neighbors Algorithm*, 10 September 2018, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 , Accessed 4 May 2022.

[35] See note 29.

the best “out of the box” classifiers since it has been proven to perform well in a lot of different situations. The SVM is an extension of the linear Support Vector Classifier. The Support Vector Classifier is converted into the SVM classifier by converting the linear classifier into a classifier that produces non-linear decision boundaries. This is done by enlarging the feature space using kernels. A kernel is a function that quantifies the similarity of two observations. There are a variety of different kernels that can be used, however, in this thesis the radial kernel was used. The radial kernel takes the form,

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{19}$$

where gamma is a positive constant. If a given test observation

$$x^* = (x^*_1 \ldots x^*_p)^T \tag{20}$$

is far from a training observation x_i in terms of Euclidean distance, then

$$\Sigma^p_{j=1} (x^*_j - x_{ij})^2 \tag{21}$$

will be large, and $K(x_i , x_{i'})$ (19) will be very small. Thus, training observation x^* will play essentially no role in the predicted class label for x^*. That is, the radial kernel has a very local behavior because only the nearby training observations have an effect on the class label of a test observation. [36]

3.5 Random Forests

A classification tree is a type of decision tree that is used to predict a qualitative response. In the classification tree approach, an observation is predicted to belong to the most commonly occurring class of training observation in the region to which the observation belongs to. When interpreting a classification tree, the focus lies not only on the predicted class corresponding to a specific region, but also the proportions of the classes among the training observations that fall into that specific region. The process of growing a classification tree is done by binary recursive splitting. Binary recursive splitting is considered a top-down approach, as it begins at the top of the tree where all observations belong to a single region

[36] See note 29.

and then sequentially splits the tree in two new branches based on what branches that best minimizes the classification error rate. The classification error rate can be described as the fraction of the training observations in a specific region that do not belong to the most common class. [37]

The previously mentioned classification trees often suffer from high variance. That is, the results from testing unobserved data on the trained classification tree provide inconsistent results. In order to lower the variance, bootstrap aggregation (Bagging) can be used.
The main idea of Bagging is to average a set of observations in order to reduce the variance and thus increase the prediction accuracy. This is done by drawing samples of the training observations with replacement in order to create a lot of training sets from the population (known as bootstrapping), and then build a separate prediction model using each training set and averaging the resulting predictors. [38]

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{22}$$

The problem with Bagging is that it might lead to highly correlated trees. Random Forests is an approach that is based on bagging, but by using a small tweak, decorrelating the trees. Just like in bagging, a number of decision trees are built based on the bootstrapped training samples. However when trees are built using Random Forests, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is only allowed to use one of these *m* predictors. [39]

[37] See note 29.

[38] See note 29.

[39] See note 29.

4. Results

In this section the results are presented and discussed.

In this section the result for each model is presented with a confusion matrix followed by a summarized table of the accuracy for all the model and random guessing measures. A confusion matrix gives a clear overview of how many observations that were correctly classified and also the observations that were not.The observations correctly classified are represented in the diagonal of the confusion matrices and have also been color coded to lightblue. While the observations that were incorrectly classified have been color coded to orange. The accuracy of the models represents the amount of observation the models managed to correctly classify into their respective classes.[40]

Table 2. Confusion matrix for LDA

	LDA		**Actual**		
	Classes	0-25th percentile	25-50th percentile	50-75th percentile	75-100th percentile
	0-25th percentile	272	0	63	86
Prediction	25-50th percentile	180	0	61	139
	50-75th percentile	150	0	62	191
	75-100th percentile	89	0	52	254

According to table 2, the accuracy is equal to,

$$(272+0+62+254) / 1599 \approx 0.368. \quad (23)$$

That is, about 36.8% of the observations were correctly classified. When observing table 2 the majority of the correctly classified observations was classified as 0-25th percentile or 75-100th percentile, while the class 25-50th percentile had zero correctly classified observations. A potential reason for this might be that the assumptions of the data being

[40] Machine Learning Mastery, J. Brownlee, *What is a Confusion Matrix in Machine Learning?*, 18 November 2016, https://machinelearningmastery.com/confusion-matrix-machine-learning/ , Accessed 10 May 2022.

multivariate normally distributed and the covariance matrix of x_i is equal within all k-classes were not fulfilled.

Table 3. Confusion matrix for KNN

	KNN	Actual			
	Classes	0-25th percentile	25-50th percentile	50-75th percentile	75-100th percentile
Prediction	0-25th percentile	244	64	74	39
	25-50th percentile	130	102	78	70
	50-75th percentile	111	64	136	92
	75-100th percentile	75	49	83	188

According to table 3, the accuracy is equal to,

$$(244+102+136+188) / 1599 \approx 0.419. \quad (24)$$

That is, about 41.9% of the observations were correctly classified. When observing table 3 the majority of the correctly classified observations was classified as 0-25th percentile or 75-100th percentile similar to table 2.

Table 4. Confusion matrix for SVM

	SVM	Actual			
	Classes	0-25th percentile	25-50th percentile	50-75th percentile	75-100th percentile
Prediction	0-25th percentile	306	15	37	63
	25-50th percentile	203	18	50	109
	50-75th percentile	172	21	49	161
	75-100th percentile	112	13	37	233

According to table 4, the accuracy is equal to,

$$(306+18+49+233) / 1599 \approx 0.379. \quad (25)$$

That is, about 37.9% of the observations were correctly classified. When observing table 4 the majority of the correctly classified observations was classified as 0-25th percentile or 75-100th percentile, similar to table 2 and 3.

Table 5. Confusion matrix for Random Forest

	RF	Actual			
	Classes	0-25th percentile	25-50th percentile	50-75th percentile	75-100th percentile
Prediction	0-25th percentile	250	62	70	39
	25-50th percentile	132	100	78	70
	50-75th percentile	112	65	134	92
	75-100th percentile	75	49	83	188

According to table 5, the accuracy is equal to,

$$(250+100+134+188) / 1599 \approx 0.420. \quad (26)$$

That is, about 42% of the observations were correctly classified. When observing table 5 the majority of the correctly classified observations was classified as 0-25th percentile or 75-100th percentile, similar to table 2,3 and 4.

Presented below are the accuracy for the different machine learning models used as well as the random guessing measures. A common rule of thumb when it comes to classification is to beat what is called random guessing by 25%. That is, if a classification model cannot beat one of the random guessing measures it is hardly of use. These criterions are based on the prior probabilities found in table 1. The maximum chance criterion is one of the random guessing measures which uses the most common class of prior probabilities and multiplies it by 1.25. The proportional chance criterion is the second of the random guessing measures which uses the sum of the squares of the prior probabilities of each respective class then multiplies this sum with 1.25.[41]

[41] Revolutions, S. Gauher, *Is your Classification Model making lucky guesses?*, 22 March 2016, https://blog.revolutionanalytics.com/2016/03/classification-models.html, Accessed 4 May 2022.

Table 6. Accuracy for the models and random guessing values

LDA	KNN	SVM	RF	Maximum chance criterion	Proportional chance criterion
0.368	0.419	0.379	0.420	0.329	0.250

According to table 6, Random Forest does the best job at predicting win-share class based on draft pick with an accuracy of 42%, with K-Nearest Neighbors being a close runner-up with 41.9%. When observing table 6, all of the machine learning models used beat each of the two random guessing measures, thus all of the machine learning models are deemed as relevant. These results were a bit surprising since the SVM method is known as a very good out of the box classifier, but performed about 4 percentage points worse than Random Forest.

5. Conclusion

The objective of this thesis was to evaluate the expected performance of a basketball player coming into the NBA based on where the player was selected in the draft. This was done by using different machine learning models on the data and comparing their results in terms of classification accuracy. The results have shown that all of the models managed to beat the random guessing measures and the best performing model, Random Forests, had a 42% accuracy in predicting the class of percentiles of these win-shares, followed closely by K-Nearest Neighbors at 41.9%. Since all of the models beat the random guessing measures, the models may be considered of use for those interested in evaluating expected performance of NBA players.

References

Adams, J. & Maroon-News Staff, The Colgate Maroon-News, *Tanking in the NBA: Good or Bad For the League?,* 11 February 2016, https://thecolgatemaroonnews.com/4167/sports/nattysports/tanking-in-the-nba-good-or-bad-for-the-league/, Accessed 2 May 2022.

Alhamid, M., Towards Data Science , *What is Cross-Validation?*, 24 December 2022, https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75, Accessed 4 May 2022.

Beck, H., Sports Illustrated, *The Tanking Era As We Know It Is Over*, 26 April 2021, https://www.si.com/nba/2021/04/26/tanking-era-is-dead-play-in-tournament, Accessed 2 May 2022.

bhussey, NBA, *Charlotte Bobcats Become NBA´s 30th Team, 6 May* 2004, https://www.nba.com/hornets/news/official_team.html, Accessed 2 May 2022

Brownlee, J., Machine Learning Mastery, *What is a Confusion Matrix in Machine Learning?*, 18 November 2016, https://machinelearningmastery.com/confusion-matrix-machine-learning/, Accessed 10 May 2022.

Dash, SK., Analytics Vidhya, *A Brief Introduction to Linear Discriminant Analysis*, 18 August 2021, https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/, Accessed 4 May 2022.

Feldman, M., Front Office Gurus, *2020 NBA DRAFT: FLOOR AND CEILING COMPARISON´S FOR EACH TOP PROSPECT*, 19 May 2020, https://frontofficegurus.com/2020/05/19/2020-nba-draft-floor-and-ceiling-comparisons-for-each-top-prospect/, Accessed 2 May 2022

Gauher, S.,Revolutions, *Is your Classification Model making lucky guesses?*, 22 March 2016, https://blog.revolutionanalytics.com/2016/03/classification-models.html, Accessed 4 May 2022.

Gotesman, R.,Towards Data Science, *How Did Linear Discriminant Analysis Get Its Name?*, 2 July 2019, https://towardsdatascience.com/mathematical-insights-into-classification-using-linear-discriminant-analysis-9c822ad2fce2, Accessed 4 May 2022.

Harrison, O., Towards Data Science, *Machine Learning Basics with the K-Nearest Neighbors Algorithm*, 10 September 2018, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761, Accessed 4 May 2022.

James ,BasketballNoise, *How does the NBA draft work?*, 23 October 2021, https://basketballnoise.com/how-does-the-nba-draft-work/, Accessed 2 May 2022.

James, G., Witten, D., Hastie, T., Tibshirani, R., Springer Texts in Statistics,, *An Introduction to Statistical Learning with Applications in R*, 2013, https://link-springer-com.ezproxy.its.uu.se/content/pdf/10.1007/978-1-4614-7138-7, Accessed 4 May 2022.

Pelton, K., Sonics Central, *The WARP Rating System Explained,* June 2021, http://sonicscentral.com/warp.html, Accessed 21 April 2022.

Rudd, D. & Solomon, N. , wbhsbullseye, *The Rise of the NBA,* 14 March 2019, https://wbhsbullseye.com/808/sports/the-rise-of-the-nba/, Accessed 2 May 2022.

Zaccaneli, F. , Insider, *Former Mavericks president and GM explains the risks and rewards of evaluating NBA talent*, 20 June 2017, https://www.businessinsider.com/evaluating-nba-players-draft-2017-6?r=US&IR=T, Accessed 2 May 2022.

Basketball Reference, *Calculating Individual Offense and Defensive Ratings,* https://www.basketball-reference.com/about/ratings.html, Accessed 21 April 2022.

Basketball Reference, *NBA Win Shares,* https://www.basketball-reference.com/about/ws.html, Accessed 20 April 2022.

Data: https://stathead.com/tiny/xA93X, Accessed 20 April 2022.

ESPN.com, ESPN, *2022 NBA draft order: Complete picks for the first and second rounds ahead of the lottery*, 20 April 2022, https://www.espn.com/nba/story/_/id/33748170/2022-nba-draft-order-complete-picks-first-second-rounds-ahead-lottery, Accessed 2 May 2022.

NBA, *A Chronology of the Teams in the NBA,* https://www.nba.com/celtics/history/nba-teams-chronology#:~:text=On%20June%206%2C%201946%2C%20the,Louis%20the%20West%20Division, Accessed 2 May 2022.

NBA.com Staff, NBA, *NBA Draft Lottery: Odds, history and how it works,* 15 April 2022, https://www.nba.com/news/nba-draft-lottery-explainer, Accessed 2 May.

NBA Stuffer, *Possession*, https://www.nbastuffer.com/analytics101/possession/, Accessed 21 April 2022.

The Editors of Encyclopaedia, Britannica, *National Basketball Association*. *Encyclopedia Britannica*, 22 July 2021, https://www.britannica.com/topic/National-Basketball-Association, Accessed 2 May 2022.