Is the future of basketball being influenced by predictive data analysis?

Enrique Moreno[a], David Gil[b], Jose F. Vicent[a,*]

[a]*Department of Computer Science and Artificial Intelligence, University of Alicante, Campus de San Vicente del Raspeig, Ap. Correos 99, E-03080, Alicante, Spain*
[b]*Department of Information Technology and Computing, University of Alicante, Campus de San Vicente del Raspeig, Ap. Correos 99, E-03080, Alicante, Spain*

Abstract

The performance of basketball in the NBA league varies largely from season to season and for different teams. It is evident that a method capable of forecasting and analyzing the future of this sport shall assist the management to great extent. There are several techniques to effectively forecast time series data such as Autoregressive Integrated Moving Average (ARIMA) model, that has demonstrated its performance in precision and accuracy, or machine learning techniques such as Long Short-Term Memory (LSTM) which has proven to be one of the most efficient predictive algorithms for time series. This work aims to provide a deep dive into the evolution of basketball, using techniques based on ARIMA and LSTM, but without going into too many technicalities of the basketball game. The analysis carried out shows clear evidence that the NBA is going through a phase of important changes, both in the style of play and in the characteristics of the players. The pace of the game has picked up over the years, increasing the number of shots per game and reducing the average length of possession. As a consequence, teams are leaning towards more versatile and dynamic rosters.

Keywords: Time Series Forecasting, Autoregressive Integrated Moving Average, Long Short-Term Memory, Sports Analytics, Basketball

1. Introduction

When Dr. James Naismith was requested to come up with an indoor sport that would allow athletes to stay in shape during the cold winters, nobody could have imagined his creation would eventually become the multibillion-dollar business that basketball is today [1]. The rules have deviated significantly from the 13 that were originally proposed [2] and the style of play has continuously developed to suit the changing physical characteristics and skill sets of the players. In fact, some rule modifications that took place over the years were designed to mitigate the physical superiority of certain players, as well as to make the game more entertaining for the spectators. In the case of the National Basketball Association (NBA), notable changes include the incorporation of the 24 second shot-clock in 1954 to increase the tempo of the game, or the addition of the 3-point line in the late 70s, which forced players to spread out on the court and took the game to another degree of difficulty [3]. If we fast-forward to the current era of basketball, the 4-point line is a hot topic of discussion. Notable shooters such as Stephen Curry, Damian Lillard or Trae Young have mastered the art of the 3-point shot and have stretched their shooting range to practically half-court. For this reason, it is not surprising to see that NBA teams have begun to add extra lines beyond the 3-point line [4]. This is one of the many examples of progress in the game of basketball. Observing the trends in historical data can help teams anticipate the changes in the style of play and in the players' characteristics

*Partially supported by the Spanish Government, grant number PID2021-127275OB-I00, FEDER.
*Corresponding author.
Email addresses: emc132@alu.ua.es (Enrique Moreno), david.gil@ua.es (David Gil), jvicent@ua.es (Jose F. Vicent)

in order to gain an edge over the competition. Identifying the strengths and weaknesses of the opponents objectively through the use of data allows teams to prepare strategies to maximise their chances of winning, for example.

Due to its high level of competition and popularity worldwide, the NBA can be considered to be representative of the global state of basketball. This work uses NBA data to analyse and predict the evolution of the game of basketball. The selection of data focuses on the shots and physical characteristics of the players from the last two decades. These provide insight into how the game of basketball has evolved over the years and what the prototype of an NBA player looks like. Combining these two provides a picture of where the game of basketball is heading towards, which is then contrasted with the future predictions of time-series models and machine learning techniques.

The main contributions of this document are:

- Carrying out an empirical study and analysis of with the objective of investigating the performance of traditional forecasting techniques and algorithms based on deep learning in predicting the evolution of basketball.
- Using LSTM and ARIMA with respect to the minimisation achieved in the forecast error rates.
- Investigating shot predictions and player physical characteristics to provide a picture of where basketball is headed.

To achieve the goals, this paper is organised as follows. The *Related Work* section provides context to the presented work integrating key information from a list of related scientific papers. Then, the *Preliminaries* section describes two models used in the work, ARIMA and LSTM, and it serves as a starting point to establish the predictions. The main part of the paper is the *Methodology* section, that proposes the research scheme to formulate the prediction hypotheses. The *Experimental Results* section focuses on measuring the impact of the predictions discussing relevant characteristics of the aforementioned predictions. Finally, we draw some conclusions about our study.

2. Related work

This section introduces two key topics that are in continuous development and are allowing sport organisations to become more data-driven in their decision-making, namely, the fields of sports analytics and activity recognition.

Sports analytics is a field of research that has become very popular over the years. In sports analytics the goal is to gather and analyse the necessary data to obtain useful insights that will allow a player or team to be in an advantageous position against their opponents. Although this field is still in an emerging state, the concept of analysing sports data to facilitate decision making has been around for a while. From 1956 to 1960, Donald Knuth, who is considered to be the "founder of the fields of data structures and algorithm analysis" [5], became the Case Institute of Technology's basketball manager while being on a scholarship for his undergraduate studies. During this time, Knuth designed a computer program that evaluated the players' performance based on manually collected data. Figure 1 shows the sheet on which the data was recorded in real-time. The court drawing also allowed Knuth to annotate the locations of the shots or other in-game events. The data was then put together into IBM punch cards and fed into the IBM 650 computer system, which was able to output the player evaluations in 90 seconds [6]. Fast-forwarding to present day, the tools used to collect and analyse the data have become more efficient, but the overall process does not deviate much from Knuth's work.

Moneyball [8] is the most widely-known case of successful predictive analytics in sports (baseball). Moneyball is the story of how the Oakland Athletics, a small-budget team, assembled a competitive team through an extensive analytical process in 2002. This team's front office took new metrics into account to evaluate players across the Major Baseball League and developed pioneering prospect selection techniques under strict financial constraints. Despite losing their three star players from the 2001 season, the 2002 Oakland

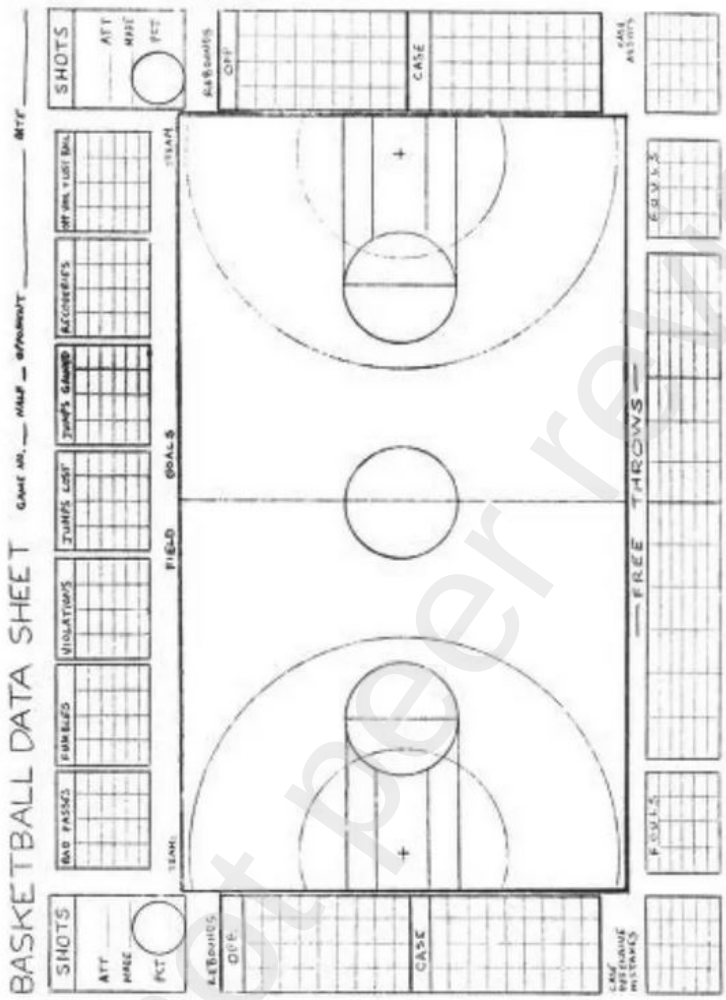

Figure 1: Donald Knuth's basketball data sheet [7]

Athletics were able to improve their previous season's win record. Similarly, the predictions in this project will try to indicate what the prototype of a basketball player will look like in the future. This knowledge would allow teams to narrow their search of new players to add to their roster.

As technology keeps advancing, the importance of the role of the data analyst in the NBA will increase. Recent years have witnessed data analysts climb the ladder up to managerial positions within NBA teams. A famous case for this is computer scientist Daryl Morey, currently the president of basketball operations of the Philadelphia 76ers. Before this, Morey was the General Manager for the Houston Rockets between 2007 and 2020. Morey's data-driven basketball philosophy led the Houston Rockets to become the first team in NBA history to shoot more 3-pointers than 2-pointers in a single season [9]. Morey's studies revealed that long 2-point shots were very inefficient. This influenced the team's strategy to either shoot 3-pointers or layups, while avoiding anything in between as much as possible. This kind of takes on the way basketball should be played have also brought up a lot of controversy, as basketball purists argue that analytics have ruined the NBA by making the style of play more predictable. Kirk Goldsberry is another researcher that made it up the ranks of an NBA team, becoming the vice president for strategic research for the San Antonio

Spurs between 2016 and 2018. After having studied data visualisation, Goldsberry became famously known for coming up with the hexagonal shot chart in basketball analytics (see Figures 10). Goldsberry introduced a research paper on innovative methods for making shot charts [10] in the 2012 MIT Sloan Sports Analytics Conference, which is co-chaired by Daryl Morey. After leaving the Spurs, Goldsberry returned to writing and published the book "SprawlBall: A Visual Tour of the New Era of the NBA" [11] which also addresses Morey's philosophy on the inefficiency of long 2-pointers. In the *Statistical Analysis* section, a variety of data visualisations (including Goldsberry's hexagonal shot chart) are displayed to showcase the trends in the basketball data.

2.1. Activity recognition in sports

The improvements of data-capturing technologies and the democratisation of data have opened up the range of work available in the field of sports analytics [12]. Teams that are able to handle the data wisely will have an edge over their opponents. From wearable inertial units [13] to motion-capturing sensors [14], these technologies provide valuable information to many sport stakeholders such as players, coaching staffs, sport organisations or fans. For example, basketball players can learn more about their own shooting technique with the data gathered on their shooting arc, point of release or hand placement. The technology required to track these need not be complex nor expensive, allowing amateur players to also benefit from it. Apps like HomeCourt already provide a very useful service to help players improve their shooting skills [15]. The NBA uses an in-game player tracking system by Second Spectrum to obtain the locations of the players and the ball during the whole game [16]. This data is then fed into computer vision algorithms which break down the actions and convert the data into what basketball fans know as game statistics.

There are many elements of basketball game play that do not look quantifiable. How would one model team chemistry or shot difficulty, for instance? This is an active area of study that has a lot of promise. The creation of metrics opens up the possibility to capture new gameplay insights. Research on the tactile communication (physical touch) between players proved that these kind of interactions at the early stages of the competition were correlated with better performance later in the season [17]. A research paper presented at the 2014 MIT Sloan Sports Analytics Conference introduced new metrics for shot analysis that took the difficulty of the shot and the ability to shoot above expectation into account [18]. This paper evaluated several machine learning models, such as decision trees, logistic regression and Gaussian process regression, on their ability to effectively detect shot quality. The motivation behind this is that not all shots from the same location are the same, since the way a player is being defended and the way the attacker gets to that shot can vary. For this reason, a wide-open shot from a given spot should not be evaluated in the same way as a contested shot from the same spot. This information can reveal which teams' or players' shooting efficiency is above average or what their shot selection looks like.

Data science and machine learning techniques are already being used to analyse the impact of injuries on player and team performance [19]. Ideally injuries could be prevented before they occur, and injury forecasting has also been studied with the use of deep learning [20]. Game result prediction is another hot topic where significant money is involved, as it is closely linked with sports betting [21].

3. Preliminaries

Since this work focuses on the evolution of basketball through time, the data handled is in the form of time series. A time series is a set of data points ordered chronologically. The intervals between data points need not be equal. In time series analysis, the goal is to be able to interpret the series and to extract the patterns in the data. The knowledge acquired about these patterns allows to predict unseen data points, either beyond the range of data (extrapolation) or within the series timespan (interpolation). This section will introduce the two models used in this project, ARIMA and LSTM, as well as their applications in sports analytics. The motivation behind this choice of models is to study whether the modern deep learning approaches like LSTMs [22] obtain better results than the traditional ones like ARIMA [23].

3.1. ARIMA

ARIMA models are statistical models that use regressions in order to find patterns in the data that will help predict future data [24]. In other words, an ARIMA model describes a data point as a linear function of previous observations (in time) plus some random error. It can also contain a cyclical or stationary component. The model consists of three key elements:

- (AR) Autoregressive: model that describes that a given point in time can be predicted using observations before that time.
- (I) Integrated: differencing of consecutive observations to make the series stationary.
- (MA) Moving average: model that describes a stationary time-series.

ARIMA is usually described as *ARIMA(p,d,q)* where p is the number of past observations used to make a prediction, d is the number of times differencing takes place, and q is the number of observations used to compute the moving average. Each of these parameters is a non-negative integer. The applications of ARIMA models in sport include the prediction of game outcomes, as they have been proven to obtain around 90% of prediction accuracy [25]. In the context of basketball, ARIMA has been used to study performance variability [26] and for injury analysis [27].

3.2. LSTM

LSTMs are a kind of artificial neural network that works well with sequential data and is therefore suitable for time-series prediction tasks. They are known for being able to mitigate the vanishing gradient problem in recurrent neural networks. An LSTM is capable of "remembering" relevant sequential data and storing it for some time in a similar way to what the human brain does. LSTMs have a hidden layer of memory blocks, each of which contain a recurrent memory cell, an input/output gate and a forget-gate. These gates are trainable, meaning that the block can learn whether to hold on to the information or not [28]. In sport, LSTMs have been used for game outcome prediction [29] and activity recognition [30], among other things. Applications of LSTMs in basketball include the prediction of shot trajectories [31] or offensive play classification [32].

4. Methodology

This section presents the steps performed to carry out the statistical analysis and the subsequent predictions.

4.1. Data extraction

The official NBA data is obtained through the NBA_API [33], an unofficial API Client for the NBA's website (nba.com). The NBA website contains a statistics page where different types of historical NBA information can be retrieved. Because the scope of this project is to understand the evolution of modern basketball, the data obtained includes the entirety of the 21st century so far, ranging from the 2000-01 season up to the most recent 2021-22 season (22 seasons in total). For each season, data from different stages of the competition can be obtained. These are the preseason, regular season, All Star and playoffs. This project will only use the regular season and playoffs data, as these stages are where the outcome of each match will directly affect the league standings. The preseason is the series of friendly matches that occur before the regular season begins. Teams use these stage to get their players in shape and start to build the team chemistry. The All Star is a series of exhibition events which include some games between a minority of selected players from the league. For these reasons, the preseason and All Star cannot be considered representative of the state of basketball competition.

4.1.1. Shooting Data

The data selected to analyse the evolution of shooting across the year comes from the **shotchartdetail** endpoint[1]. A shot chart is a visual representation of the location and efficiency of the shots (see Figure 2). The shots to be analysed are the field goal attempts (FGA). These are all the shots that are worth 2 or 3 points, meaning free throws are not included.

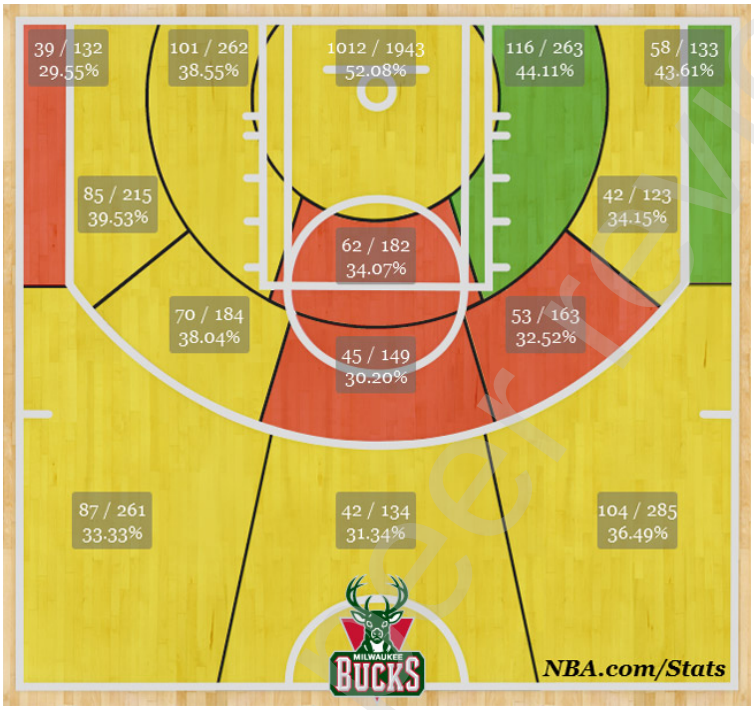

Figure 2: Example of shot chart for the Bucks team

There are a few things worth mentioning about the data preprocessing. The dataset holds 4666475 field goal attempts from 28155 games held in the last 22 seasons. Because every basketball game is meant to last (at least) the standard length of 48 minutes, outlier detection does not play a significant role in this study. However, it is necessary to verify that all games were played for at least 48 minutes. There could have been cases where the game was cancelled before completing the 48 minutes. The verification showed that all games were played for a minimum amount of 48 minutes. However, 6% of the games went into overtime, meaning that more than 48 minutes were played in those games. Because some seasons had different number of games due to special circumstances like COVID-19 or the 2011 NBA Lockout, calculations are made per game (48 minutes) instead of per season to ensure fairness in the experiments. For this reason, overtime periods produce an outlier effect on the data and are therefore discarded from the study.

Since the general trends in each season are the point of emphasis, it is not necessary to track down each shot to the individual or team level. For this reason, columns describing the player IDs, player names, team IDs and team names are considered irrelevant and discarded from the original dataset. The same approach is taken with other columns that describe the game's metadata, such as date, home team and visiting team. The final dataset is described below to grasp a better understanding of the columns' meanings and their purpose in the study. Every row in the dataset represents a different field-goal attempt. Therefore, the dataset contains all shots from every official game (regular season and playoffs) in every season between seasons 2000-01 and 2021-22 (both included).

[1]ShotChartDetail: https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/endpoints/shotchartdetail.md

- **SEASON ID**: Newly created column containing unique season identifiers (e.g., 2000-01, 2001-02, ..., 2021-22).
- **SEASON STAGE**: Newly created column containing unique season stage identifier, either *Regular Season* or *Playoffs*.
- **GAME ID**: Unique identifier for every game.
- **PERIOD**: Each game has 4 periods of 12 minutes each. Since overtime periods have been removed, the values of this column range from 1 to 4 only.
- **MINUTES REMAINING**: Ranges from 0 to 12.
- **SECONDS REMAINING**: Ranges from 0 to 60.
- **SHOT TYPE**: The amount of points the shot is worth (either 2 or 3).
- **LOC X**, **LOC Y**: X- and Y-coordinates for the location of the shot. The location is measured relative to the basket the shot is directed towards [2]. X goes from -250 to 250 (left and right side of the court, respectively). Y goes from -52 (slightly behind the offensive basket) to 884 (the defensive basket).
- **SHOT MADE FLAG**: Boolean denoting whether the shot was made or not.

4.1.2. Player Bios

In the same fashion as with the shooting data, the player bios from the selected 22 seasons are extracted from the **leaguedashplayerbiostats** endpoint[3]. This dataset originally contains 23 columns, but only 8 of them are kept for the study. The player bios dataset is merged with a second dataset from the **commonteamroster** endpoint[4] to obtain some interesting additional columns. These include the players' positions, their team prior to joining the NBA and their years of experience in the league. The columns used to merge both datasets are **SEASON**, **PLAYER ID** and **TEAM ID**, ensuring the data is appended to the correct player. The final DataFrame for the player data consists of 9482 rows (players) and 10 columns. The **TEAM ID** column is dropped after the merge as it is no longer required for the analysis. The resulting columns are the following:

- **SEASON**: Newly created column containing unique season identifiers (e.g., 2000-01, 2001-02, ..., 2021-22).
- **PLAYER ID**: Unique identifier for every player.
- **AGE**: in years.
- **HEIGHT**: in centimetres.
- **WEIGHT**: in kilograms.
- **COLLEGE**: College the player attended. *None* if the player entered the NBA straight from high school or from a team overseas.
- **COUNTRY**: Player's birthplace (not nationality).
- **POSITION**: Player's main position on the court.
- **EXP**: Years of experience in the NBA.
- **PRE-NBA TEAM**: Player's last high school, college or club prior to making their debut in the NBA.

[2] In-depth explanation of the court coordinate system: https://datavizardry.com/2020/01/28/nba-shot-charts-part-1/

[3] LeagueDashPlayerBioStats: https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/endpoints/leaguedashplayerbiostats.md

[4] CommonTeamRoster: https://github.com/swar/nba_api/blob/master/docs/nba_api/stats/endpoints/commonteamroster.md

4.2. Time-series predictions

The ARIMA model is implemented using the *auto_arima* method from the *pmdarima* package, which is an extension of the popular *Statsmodels* library [34]. The advantage of using *auto_arima* is that it tries to identify the best ARIMA parameters and avoids all the manual search. The data that is fed into this model is in the form of a Pandas Series, which is obtained after doing the necessary operations on the shooting and players DataFrames.

For the LSTM model, a multistep forecasting model is implemented using *Tensorflow*'s popular deep learning API *Keras*. The time-series data is transformed into a supervised learning problem using a rolling-window approach in which each training sample consists of $m+n$ elements, where m is the number of past observations used to predict n future observations. A naive baseline model that propagates the occurrence at the previous time-step is created to measure the effectiveness of the more complex methods ARIMA and LSTM. The LSTM implementation consists of an LSTM layer (one neuron only) and a dense layer, using the mean squared error as the loss function and ADAM [35] as the optimiser.

5. Experimental results

This section shows the results obtained with the statistical analysis and the forecasting methods, providing a picture of the past, present and future state of basketball.

5.1. Statistical analysis

Once the appropriate data has been collected and analysed, we can present some key findings on the evolution of basketball.

5.1.1. Speed of play

To get a feel for the changes in the pace at which the game of basketball is played, the frequency of shot attempts can be analysed. The average time taken between consecutive shots can also be considered as an indicator of the speed of the game. Logically these two metrics should be negatively correlated. As the average time taken between shots decreases, the amount of shots per game should increase (and the other way around). Figure 3 shows how the average amount of shots attempted in each game has increased significantly over the years, as teams are able to get more shot attempts within the same time frame (48 minutes). In the 2000-01 season a game had approximately 160 shots, while in today's games there are around 170 shots being taken. The difference in the number of shots taken in the different season stages is also noticeable. When teams are competing in the playoffs (final stage of the competition), they take fewer shots in general. The most likely reason behind this is that the playoffs are the decisive stage of the competition. Therefore, the defence becomes more intense than in the regular season and the offence struggles more to put shots up [36]. On top of that, the shot selection becomes more disciplined as each shot has to be taken wisely. This trend is true for all seasons except 2000-01 and 2014-15, although the difference is minimal. The steep increment in the playoffs curve in 2014-15 coincides with the Golden State Warriors' first championship title. The Warriors will be referenced on multiple occasions throughout the project, as this team's pioneer style of play has lead them to success and it can provide a glimpse of the direction basketball is heading towards.

Analysing the same question from the time standpoint facilitates getting a better picture of what this increment in the pace of the game means. Figure 4 shows that the mean time between consecutive shots has decreased over the years. Although it might not seem like a significant numerical difference, even a 1 second decrease in the length of the average possession can have a significant impact in the overall amount of shot attempts per game. For example, if we take the oldest and most recent regular season average times (18.23s and 16.53s respectively), we see that modern basketball is played at a speed that is 10% faster than it was 20 years ago. This is equivalent to saying that for every 10 shots taken in 2000-01, there is an extra shot attempt nowadays. The aforementioned variation in the results across the different competition stages is also visible from this graph. During the playoffs, teams take on average more time to attempt a shot.

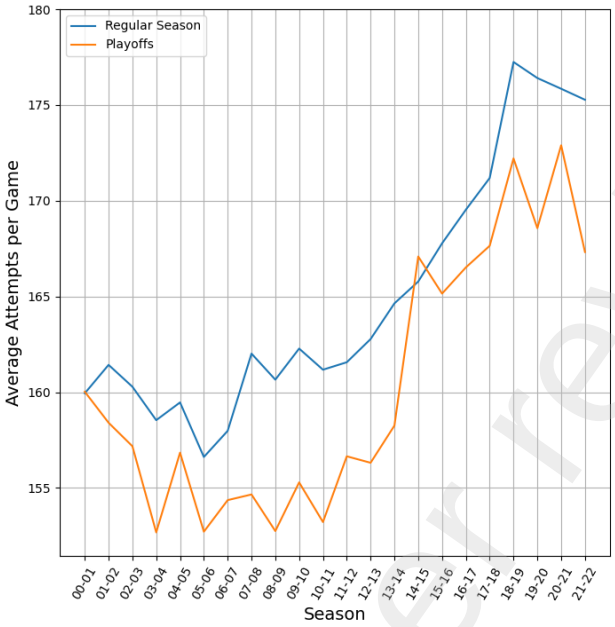

Figure 3: Evolution of shot attempts per game

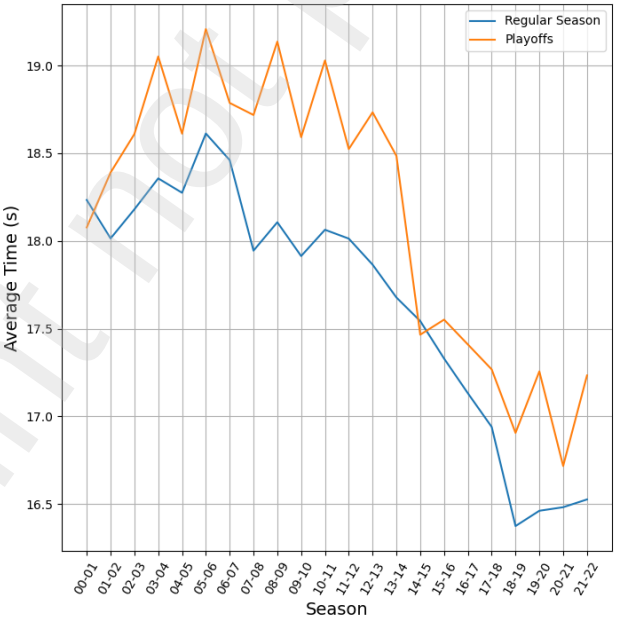

Figure 4: Evolution of time between consecutive shots

The trend observed in the speed of the game can be linked to the increase of popularity of run-and-gun playing systems. This style of play relies on fast offensive possessions rather than slowing down the game to call set plays [37]. In the 2005-06 season, the Phoenix Suns coached by Mike D'Antoni implemented the "7 seconds or less" attacking strategy which, as it name suggests, consisted of shooting early in the offensive possession before the defense was ready. This was ahead of its time, as most teams were still heavily relying on their dominant (and slower) centers for scoring. This style of play took the Phoenix Suns to the Western Conference Finals for two consecutive seasons and established the fast-paced offensive style as a game changer. In the last decade, the Golden State Warriors have been known to pick up this strategy with great success, reaching six of the last eight NBA Finals and winning four championships.

5.1.2. 3-point shooting revolution

Figure 5 reveals what has been discussed to be the most indicative sign of the progress of basketball; that is, the rise of the 3-point shot. The graph shows how the average amount of 3-point shots per game has more than doubled in the last 22 seasons, going from approximately 30 in the 2000-01 season to 70 in the 2021-22 season. This increment has lead to a decrease in the amount of 2-point shots taken per game, dropping from 130 shots on average to around 100 in the same timespan. The game speed changes can also be spotted in this graph, as the increase in 3-point shots is greater than the decrease in 2-pointers. The graph suggests that basketball is heading towards a playing style where 3-pointers will be the primary offensive tool. If the trend continues, it would not be surprising to see the 2-point- and 3-point lines overlap in the next decade. Another thing to highlight is how similar the curves for each shot type are depending on the season stage. This suggests that the teams' strategies during the playoffs do not differ much from those in the regular season.

The 3-point revolution did not happen overnight, but it has been a gradual process [38] that has sparked the emergence of specialised 3-point shooters such as the Golden State Warriors players Stephen Curry or Klay Thompson, also known as the Splash Brothers[5] due to their unmatched ability in "splashing" the net. Both of them have set numerous records related to 3-point shooting. In 2016, Curry improved the previous record for most 3-pointers made in a game from 12 to 13, but was later dethroned by his teammate Klay Thompson who scored 14 against the Chicago Bulls in October 2018. Curry also broke the record for most 3-pointers made in a season (previously 269 by Ray Allen in 2005-06) in several occasions and Curry's record now stands at 405 (2015-16 season), which is 50% greater than Allen's record. This achievement becomes more impressive considering that both seasons had the same amount of games.

5.1.3. Shooting efficiency

The 3-point revolution can be justified with mathematical evidence. The 1-point difference between 2-pointers and 3-pointers leads to an interesting question of which shot attempt is more valuable based on the shot efficiency. The longer distance in the 3-point shot creates a higher risk but also higher reward compared to the 2-point shot. The most recent season, 2021-22, is the only season in the study which had a higher value of expected points per shot attempt for 2-pointers (see Figure 6) than for 3-pointers. This graph shows that the expected points for every 3-pointer oscillates between 1.05 and 1.1, while the 2-pointer value has experienced a steep growth since the 2014-15 season, going from 0.96 to 1.06 nowadays. This season seems to be a pivotal moment in the evolution of modern basketball. It is also highly likely that this 2-point value increase is correlated with the increase in 3-point shot attempts, as teams will defend the 3-pointer harder and allow to concede 2-pointers more frequently. Prior to this season, the 80s was the last decade where the expected points for 2-pointers can be observed to be higher than those for 3-pointers (see Figure 7).

The shot efficiency per season and competition stage is plotted in Figure 8. The 2-point shot percentage shows an improvement throughout the years, reaching the widely pursued 50% mark[6] from 2016-17 up to

[5] Splash Brothers: https://en.wikipedia.org/wiki/Splash_Brothers
[6] 50–40–90 club : https://en.wikipedia.org/wiki/50-40-90_club

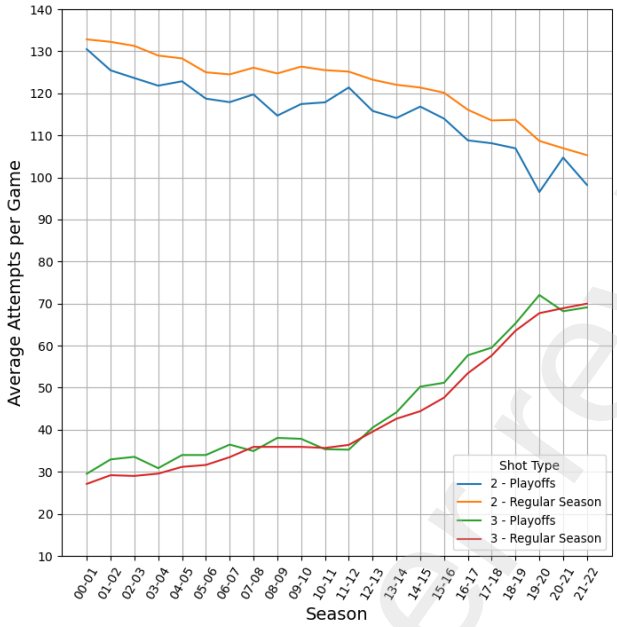

Figure 5: Evolution of 2- and 3-point shots attempts per game

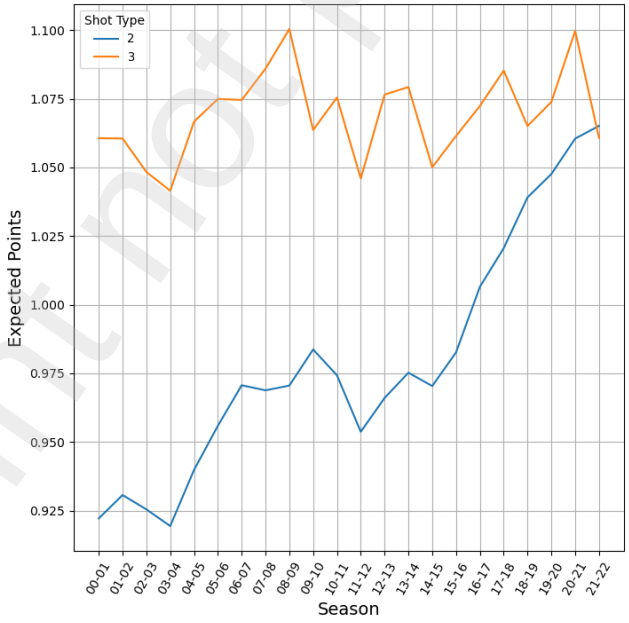

Figure 6: Expected points per shot attempt

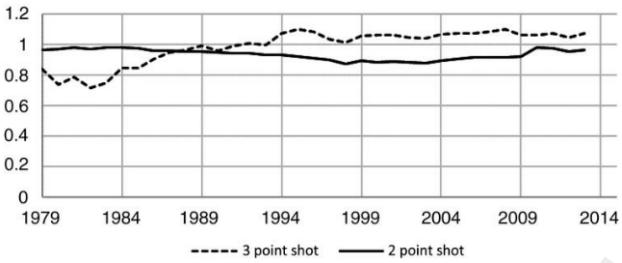

Figure 7: History of expected points per shot attempt [39]

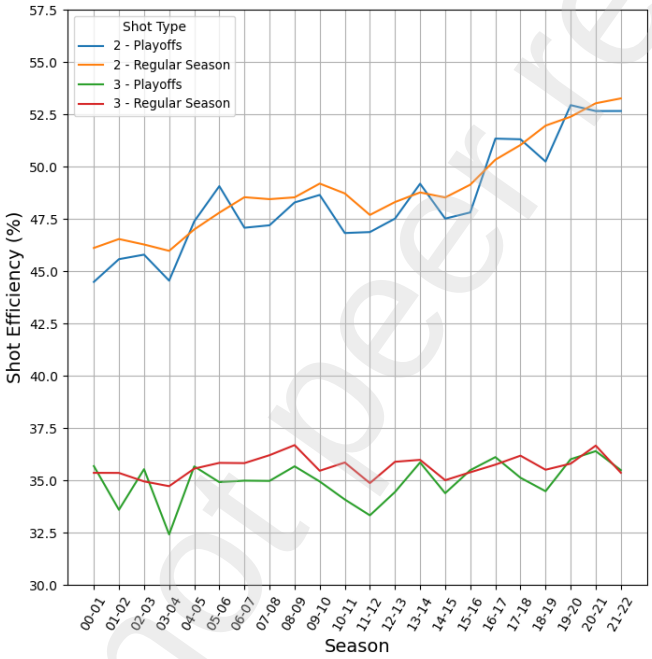

Figure 8: Evolution of shooting percentages

present day. The decrease in the number of 2-point shots has had a positive effect on its efficiency, which is likely to be linked to better shot selection. Meanwhile the 3-point percentage has stayed constant. This can be considered a good sign because the number of 3's taken per game has increased significantly throughout the years without causing a decrease in its efficiency.

5.1.4. Shot locations

The variations in the style of play during the years has had an effect of the usage of the court zones. From Figure 5 it is clear that the areas under the 3-point line are less popular in today's game than they have ever been in the past, while the zones behind the 3-point will have experienced an increase in popularity. In Figure 9, the average amount of shots in each game per location is plotted on a basketball court using the X- and Y-coordinates. Note that the mean frequency per shot location is taken as the lower end of the colour spectrum in order to highlight the more extraordinary shot locations. Since the 75th percentile is also 1, the upper end of the spectrum is set slightly above that value to 1.02, and all values above that value are treated as the maximum too. Within the 2-point range, the only area that seems to stay constant

both throughout the seasons and competition stages is the one directly below or next to the hoop. Since the distance to the hoop is minimal, it is considered as the easiest location to score from. Another area that is also very popular is the baseline corner (both sides), as it is a common spot to find open shooters when the defense collapses after a penetration or an offensive rebound. The playoff plots can be thought of as more reliable to understand the usage of the different court zones, as this is the most decisive stage of the competition and teams tend to improve their shot selection.

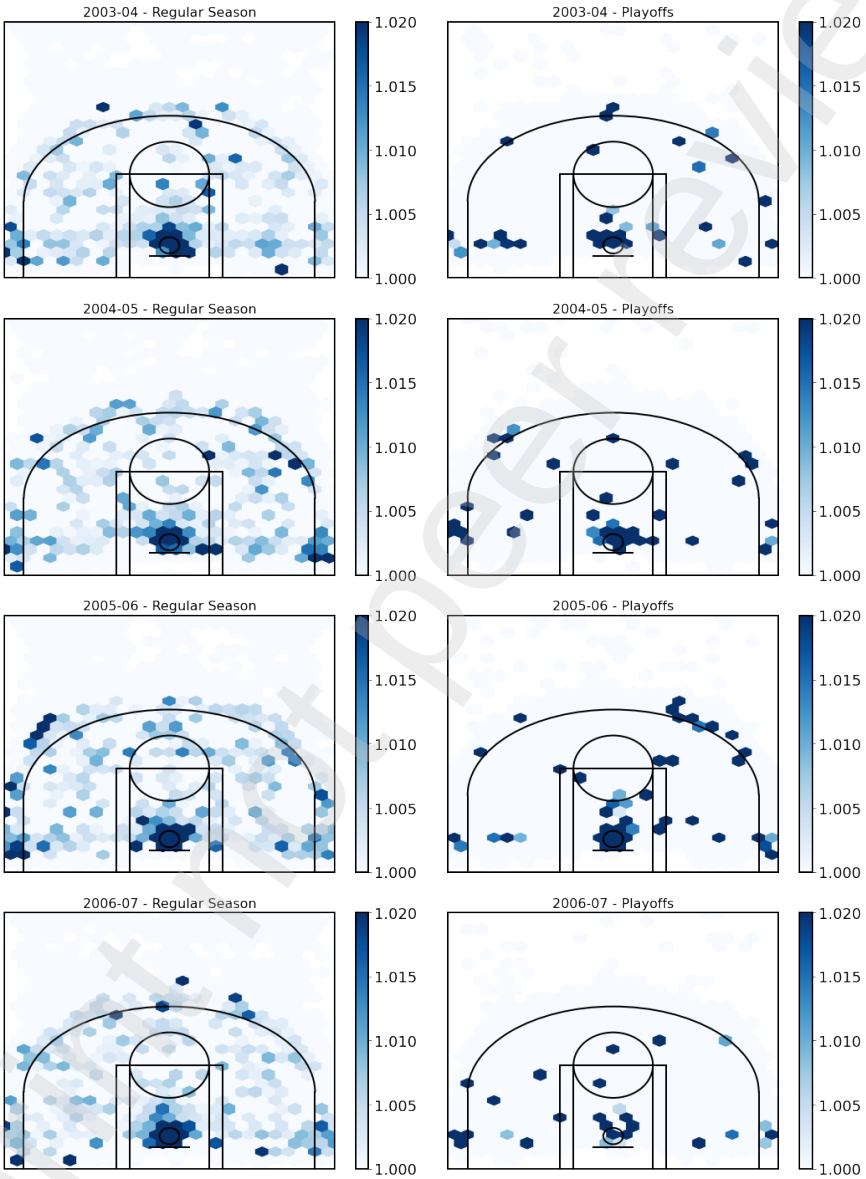

Figure 9: Sample of average shot frequency per location in the 2000s

The contrast between the early 2000s and the most recent seasons is observable in Figure 10. The 2-point shot has lost all its popularity. Focusing on the playoff data, it is striking to see how rare it is to find 2-pointers nowadays between the 3-point arc and the paint (rectangular area surrounding the hoop). Lately, teams have prioritised 3-pointers and shots closer to the basket over other shot types. This insights helps to explain the rise in popularity of the 3-pointer "step back move", where players pretend they want

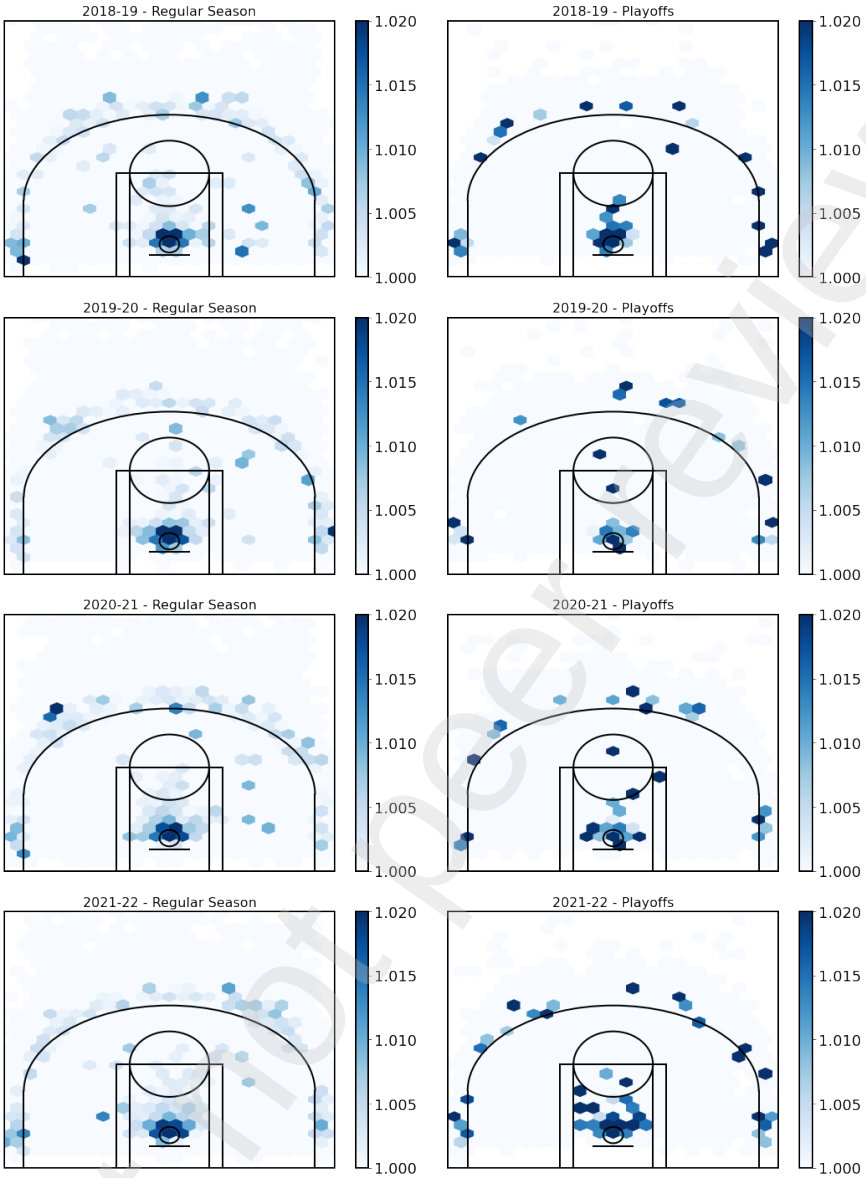

Figure 10: Sample of average shot frequency per location in the last 4 seasons

to drive in the lane for a 2-pointer but end up stepping back behind the arc for a 3-point shot. It is not a coincidence that James Harden, a master of the step back 3-pointer, was the star player of the first team in history to shoot more 3's than 2's in a single season (Houston Rockets, 2017-18).

5.1.5. Player Ages

The NBA is transitioning from a league that is led by veteran players to a league full of young talent. The physical demand required to compete in this elite competition is never decreasing [40], with young players bringing a wave of unprecedented athleticism. The new generations are developing at a vertiginous speed due to the broader access to resources that previous generations did not have, such as the Internet or better training facilities. Nowadays anyone who dreams of becoming a professional player has access to unlimited basketball-related content online, including tutorials, game replays and in-depth game analyses.

As a consequence, the NBA will most likely experience a bottleneck effect on the access of young players to the league in the upcoming years. Potential solutions that are being discussed are the expansion beyond 30 teams and increasing the current roster capacity of 15 players [41]. Figure 11 shows the NBA's rejuvenation, dropping from an average player age of approximately 28 in the 2000-01 season to 26 in the current days. The curve had a steep decrease in the early 2000s, then stabilised around the 27-year age mark between 2005 and 2016 before dropping almost linearly to the current 26-year mark at the rate of (-)0.25 years per season.

Figure 12 shows how the age distribution has been skewing towards the younger side of the graph.

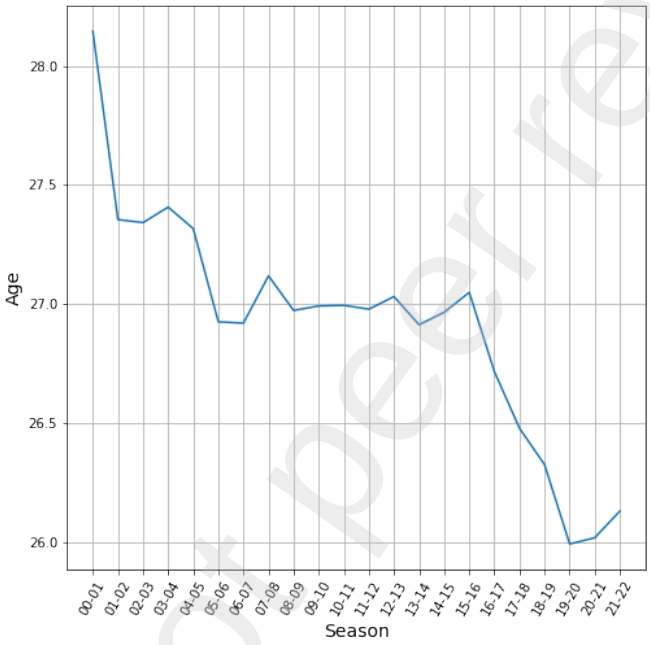

Figure 11: Average player age per season

5.1.6. Height and weight

Height has always been considered to be a necessary attribute in basketball. However, recent studies show that team height does not play a decisive role in the outcome of a game [42]. "Small ball" is amongst the latest trends in the NBA. It is a style of play which prioritises speed and agility over height and physical strength by having a lineup that meets these criteria. As a result, the pace of play increases and the spacing on the court opens up to allow 3-point shooters to excel [43]. Figure 13 shows how that adaptation to modern basketball is taking place. There are two interesting and almost-linear trends in the data. Between the 2011-12 season and the 2015-16 season, there was a continuous increase in average height and decrease in average weight. After then, every season players' height and weight decrease together. The latter coincides with the Golden State Warriors' rise to stardom and shows how influential this team has been for the sport of basketball. This is also a proof that it takes a single revolutionary playing style to suddenly change the course of the sport and hence predictions should be interpreted cautiously.

Figure 14 shows how the average player height is tending to decrease. "Small ball" also poses a threat to the old-school big and slow centers, as they will need to adapt to the new roles that the evolution of basketball is bringing to the table [44].

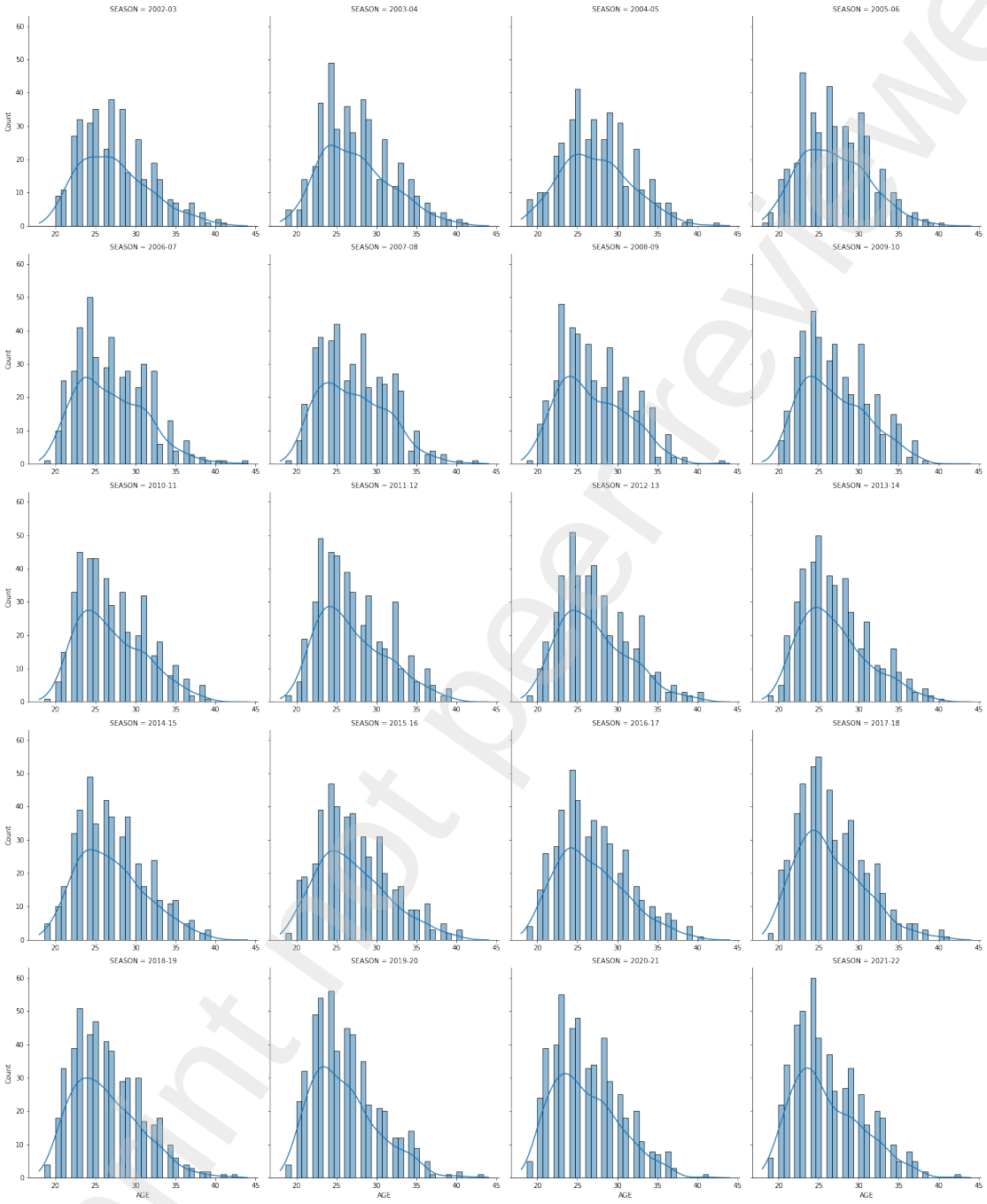

Figure 12: Age distribution per season

5.1.7. Nationalities

Since the NBA is a competition with 29 teams from the United States and 1 from Canada, it is natural that the majority of players will be Americans. In the 2000-01 season, 90% of the players had been born in the United States. The league reached an all-time high of foreign players in the 2019-20 season (24.4%).

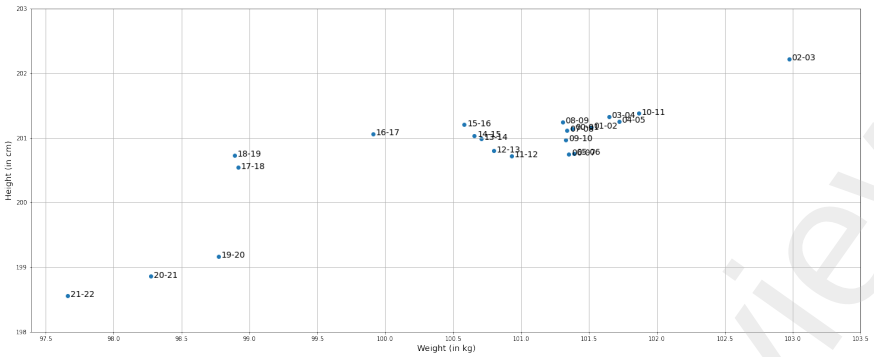

Figure 13: Height-weight relationship

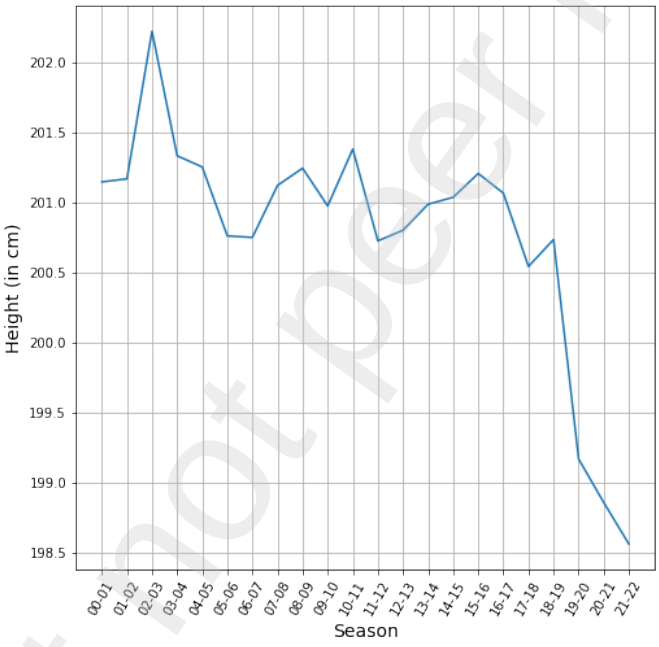

Figure 14: Average height per season

The bubble maps in Figures 15 and 16 show the globalisation at the player level. Note that the USA is not represented as a bubble to highlight the international aspect. Europe is the continent that "exports" the most players to the NBA every year. The Balkan countries have a strong basketball culture and are Europe's main ambassadors in the NBA [45] along with France. On the other hand Asia is very mildly represented. Argentina and Brazil are the two South American basketball powerhouses. The more recent bubble map shows the emergence of the African continent as a promising source of talent, nowadays almost tripling the number of players in the 2000-01 season. Two countries that also come into the scene during the 2010s are Canada and Australia. In the last 4 seasons, the podium for most (non-USA) players in the league has always been shared by these two countries and France.

Figure 15: Bubblemap of non-USA players in NBA season 2000-01

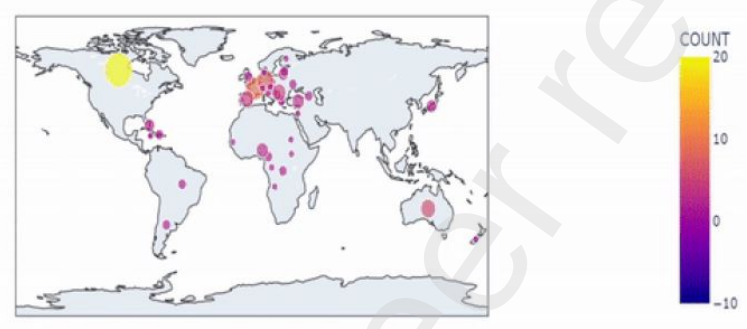

Figure 16: Bubblemap of non-USA players in NBA season 2021-22

5.2. Forecasts

Three different factors are chosen as representative to describe the evolution of basketball. These are:

- Average time between consecutive shots (as an indicator of the speed of play)
- Average amount of 2-point and 3-point shots per game (as an indicator of the shooting locations/distributions)
- Average shooting percentages (as an indicator of the shooting efficiency)

The goal of the forecasts is to predict the value for each of these forecasts for the 2022-23 NBA regular season. Since every team plays 82 regular season games, the forecasts will compute the average of the aforementioned factors by the end of the regular seasons. There are 30 teams in the NBA, thus there are 15 games taking place in each fixture. The original data is collected per game and is prepared for the forecasting task by performing a rolling mean of 15 samples, thus obtaining the averages per fixture. Once the data has been reduced to represent each fixture, it can be used to forecast the following 82 fixtures (length of the regular season). The training and test data is fed to the forecasting models using rolling windows of 164 values (82 past observations + 82 future observations), which simulates predicting one season based on the previous one. The metric that is used to evaluate the performance of the models is the mean absolute error (MAE).

5.2.1. Speed of play

In Figure 17 it is shown that the longer ahead in time the forecast is made, the less precise the prediction is. The LSTM obtains better results than the naive persistence model for all timesteps. Overall, ARIMA seems to perform better than the persistence model and the LSTM, although it is not true for all time steps. The increasing trend over time is also slightly noticeable from the MAE on the ARIMA forecasts. Finally, the models' predictions are shown in Figure 18. The LSTM predicts a mean time between consecutive shots

of 16.59s that follows the increasing linear trend of the last four seasons, while the ARIMA forecast drops to an all-time low of 16s.

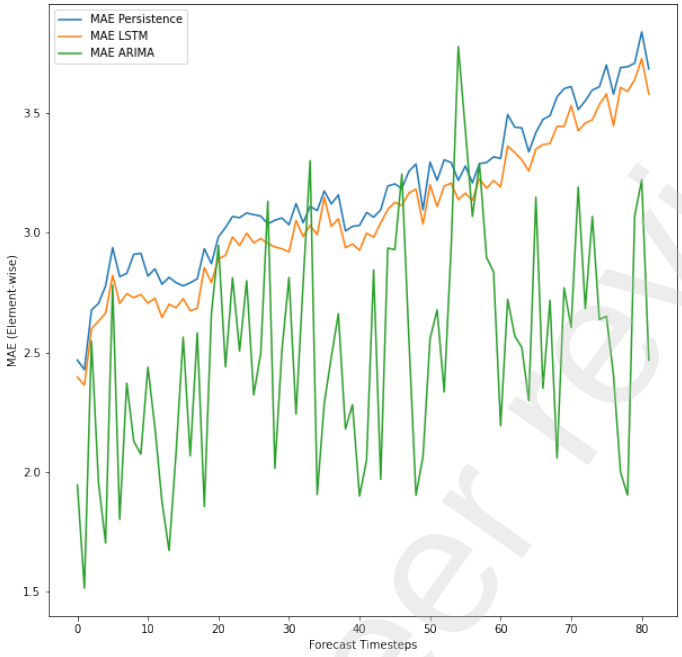

Figure 17: MAEs per look-ahead forecast (speed of play)

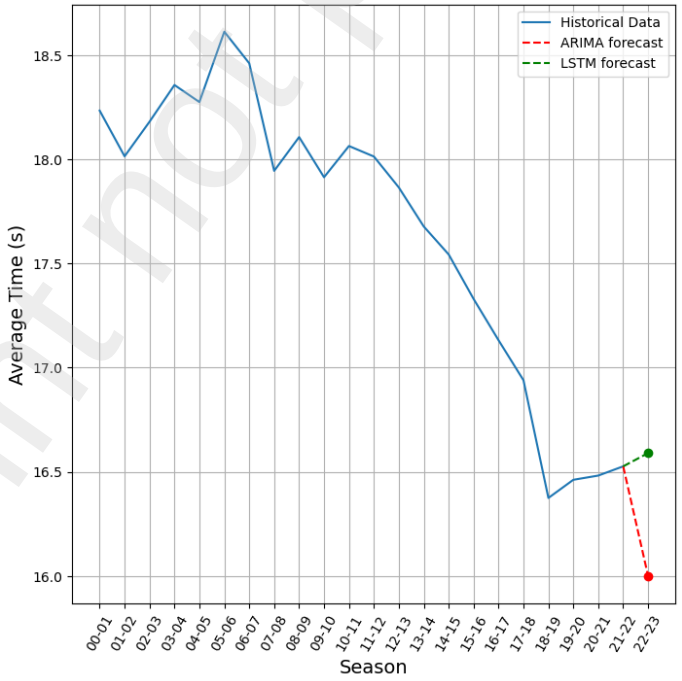

Figure 18: LSTM and ARIMA forecasts for speed of play

5.2.2. Shot distributions

The MAEs for the 2-point shot and 3-point shot forecast (see Figures 19 and 20) show very similar trends to those referring to the speed of play in Figure 17, and also with respect to each other. This time the MAEs for the LSTM are still lower than those of the persistence model, but the difference between them is larger.

As can be seen in Figure 21, both ARIMA and LSTM predict a continuation of the decrease of the amount of 2-point shot attempts per game with values close to 104. As for the 3-point shot, ARIMA predicts an increase by two shot attempts per game while LSTM predicts almost no change from the previous season.

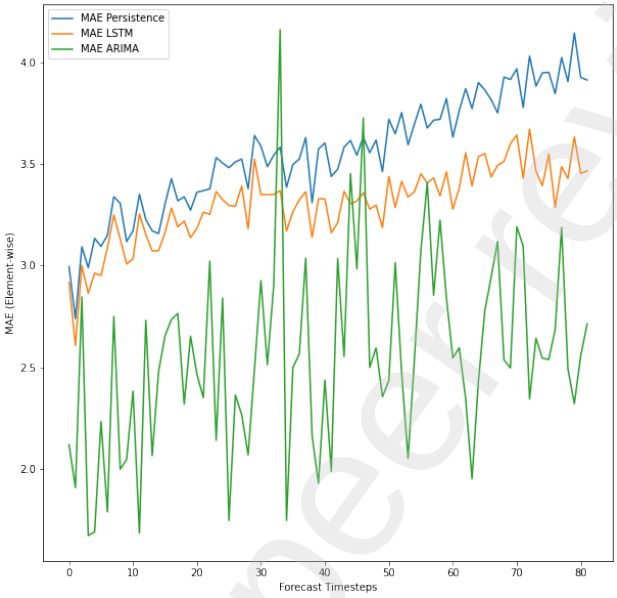

Figure 19: MAEs per look-ahead forecast (2-point shot)

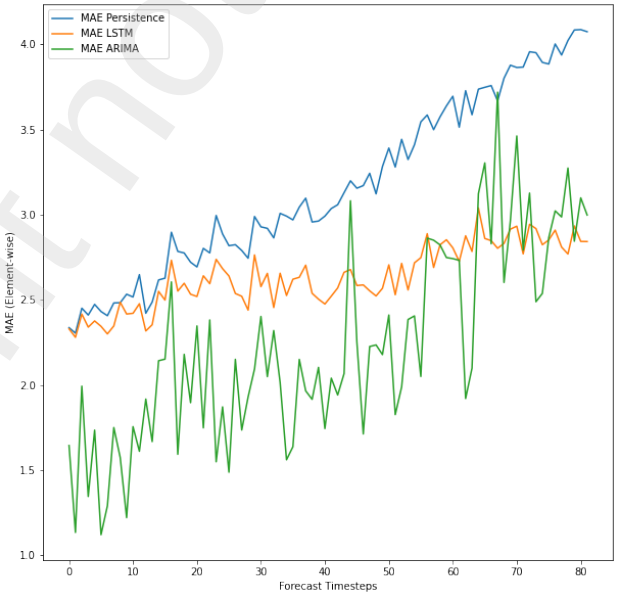

Figure 20: MAEs per look-ahead forecast (3-point shot)

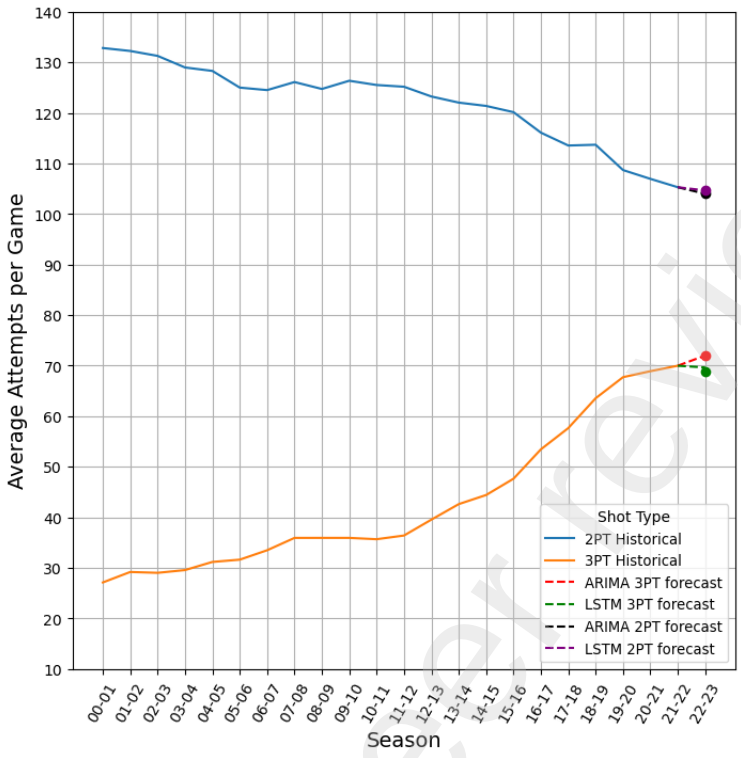

Figure 21: LSTM and ARIMA forecasts for shot distributions

5.2.3. Shooting efficiency

The MAEs for the 2-point shot efficiency predictions in Figure 22 show how ARIMA consistently performs better than the persistence model and LSTM over time. This is not the case for the 3-point shot efficiency predictions (see Figure 23), where for some timesteps ARIMA is significantly worse than the other two models (e.g., around timestep 75). Figure 24 shows the predictions emitted by the model. ARIMA and LSTM predict that the 2-point percentage will continue to increase to around 54%. The 3-point predictions also show an improvement from the 2021-22 season, with ARIMA predicting 36% and LSTM predicting an all-time high of 37.31%. These results can be used to obtain an estimate of the expected points per shot attempt. In this case the expected points per 2-point attempt would be 1.08. Taking 36.65% as an average of the 3-point prediction, the expected points per 3-point attempt would be 1.1. This would make the 3-point shot attempt more worthy than the 2-point attempt again, just one season after the 2-pointer had taken the limelight from the 3-pointer for the first time since the 80s (see Figure 6).

6. Conclusions

The statistical analysis shows clear evidence that the NBA is going through a phase of important changes, both in the style of play and the characteristics of the players. The pace of the game has sped up over the years, increasing the number of shots per game and reducing the average possession length. Teams tend to play with less risk in the playoffs than in the regular season. The overall dynamics of change in the league do not differ across competition stages, meaning that a trend that is visible in the regular season data also shows up in the playoffs data. The data-driven basketball strategies have led teams to put more emphasis on 3-point shooting over the years, to the point where the 3-point shot is getting closer to becoming the go-to action on the offensive side. These strategies have also allowed team to prioritise shot locations that will maximise their scoring probability, making some court zones look almost obsolete. Although it seems

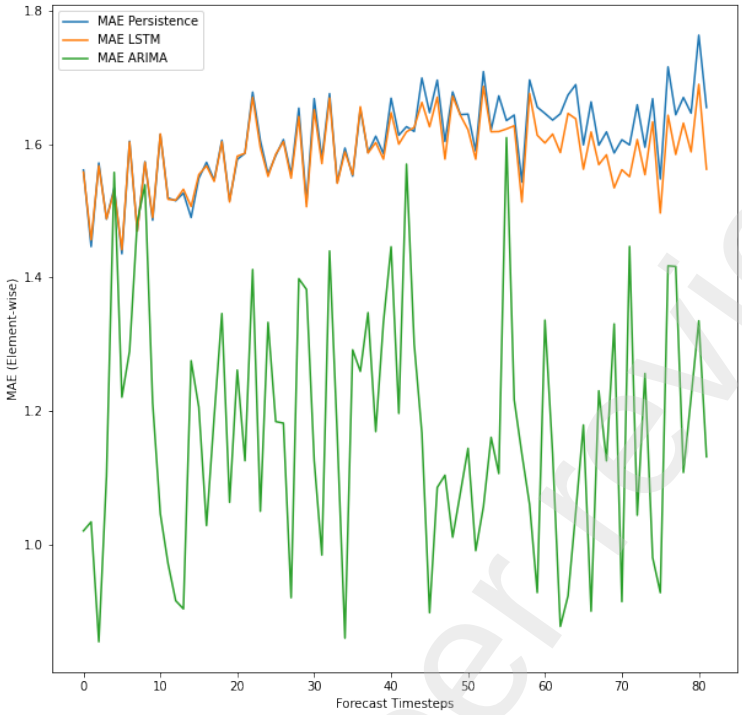

Figure 22: MAEs per look-ahead forecast (2-point efficiency)

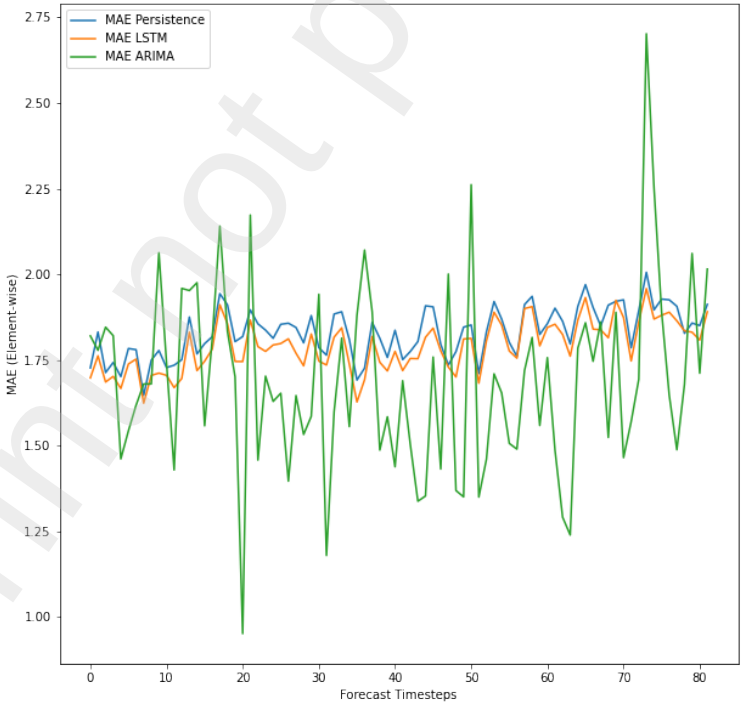

Figure 23: MAEs per look-ahead forecast (3-point efficiency)

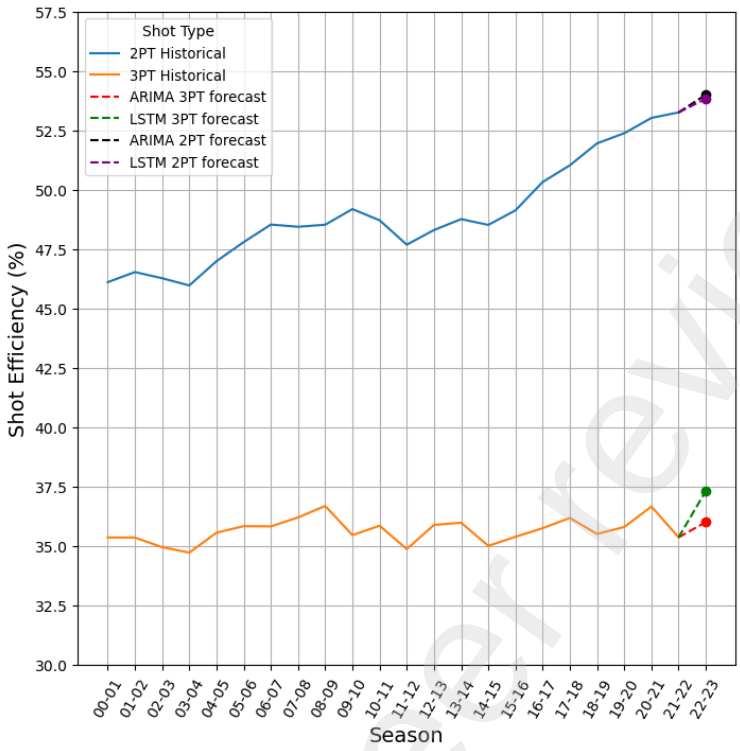

Figure 24: LSTM and ARIMA forecasts for shot efficiency

like the 3-point era is here to stay, it is noteworthy to say that the data indicates that in the most recent season, 2021-22, 2-point shot attempts were more profitable than 3-point attempts, and that strategies may need to be reevaluated. In terms of the players, the data shows that the success of "Small Ball" lineups is influencing the characteristics of the NBA player prototype, as players' heights and weights are decreasing together. Forecasting techniques can be applied to understand what the future of the sport could look like. At the same time, one should be cautious about the predictions as the sport is malleable and can change course at any moment due to pioneer playing styles or rule changes, for example.

This project aimed to provide a deep dive into the evolution of basketball without getting into too many technicalities of the game of basketball. Predictive analytics is a field which shows a lot of promise in sports. Future lines of research could include analysing the root causes of injuries and applying machine learning techniques for injury prediction, which would hopefully lead to injury prevention. Following the example of "Moneyball" for roster building, a similar strategy could be applied to fantasy sports [46], trying to maximise the performance of a fictitious team given a series of constraints. Another study that teams would benefit from is finding the relationship between the NBA Draft Combine drill results and performance at a young age [47], as teams would be able to narrow down their selection criteria during the Draft process. Although basketball is not an exact science, data analysis and machine learning will continue to provide teams with useful insights that they will be able to use to design beneficial strategies, both from an in-game and business standpoint.

References

[1] Yanbo Jin. Analysis of NBA Business Strategy. In *Proceedings of the 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021)*, pages 706–709. Atlantis Press, 2021.

[2] Dr. James Naismith. The Triangle: Basket Ball, 1892. Accessed: 4 Aug 2022.
[3] Jeremy Mertz, L. Donald Hoover, Jean Marie Burke, David Bellar, M. Lani Jones, Briana Leitzelar, and W. Lawrence Judge. Ranking the Greatest NBA Players: A Sport Metrics Analysis. *International Journal of Performance Analysis in Sport*, 16(3):737–759, 2016.
[4] Malika Andrews. How the '4-point line' and other court markings are changing the NBA, 2018.
[5] Richard M. Karp. Combinatorics, Complexity, and Randomness. *Commun. ACM*, 29(2):98–109, feb 1986.
[6] Keith Lyons. Donald Knuth, basketball and computers in sport, Sep 2018.
[7] Donald Ervin Knuth. Selected Papers on Fun and Games. In *CSLI lecture notes series*, 2011.
[8] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.
[9] Thomas H Davenport. Analytics in sports: The new science of winning. *International Institute for Analytics*, 2:1–28, 2014.
[10] Kirk Goldsberry. CourtVision : New Visual and Spatial Analytics for the NBA. 2012.
[11] Kirk Goldsberry. *Sprawlball: A visual tour of the new era of the NBA*. Mariner Books, 2019.
[12] Nitin Singh. Sport analytics: a review. *learning*, 9:11, 2020.
[13] Le Nguyen, Daniel Rodríguez-Martín, Andreu Català, Carlos Pérez, Albert Samà Monsonís, and Andrea Cavallaro. Basketball Activity Recognition using Wearable Inertial Measurement Units. 2015.
[14] David W. Chen. The Newest Player on the High School Basketball Court: High Tech, 2021.
[15] Kornel Tokolyi and Maher Elshakankiri. Internet of things in the game of basketball. In *International Conference on Internet of Things as a Service*, pages 421–435. Springer, 2020.
[16] Adrià Arbués Sangüesa, Thomas B. Moeslund, Chris H. Bahnsen, and Raul Benítez Iglesias. Identifying Basketball Plays from Sensor Data; Towards a Low-Cost Automatic Extraction of Advanced Statistics. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 894–901, 2017.
[17] Michael Kraus, Cassey Huang, and Dacher Keltner. Tactile Communication, Cooperation, and Performance: An Ethological Study of the NBA. *Emotion (Washington, D.C.)*, 10:745–9, 10 2010.
[18] Yu-Han Chang, Rajiv T. Maheswaran, Sheldon J.J. Kwok, Tal Levy, Adam D. Wexler, and Kevin Squire. Quantifying Shot Quality in the NBA. 2014.
[19] Vangelis Sarlis, Vasilis Chatziilias, Christos Tjortjis, and Dimitris Mandalidis. A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance. *Information Systems*, 99:101750, 2021.
[20] Alexander Cohan, Jake Schuster, and Jose Fernandez. A deep learning approach to injury forecasting in NBA basketball. *Journal of Sports Analytics*, (Preprint):1–12, 2021.
[21] Fadi Thabtah, Li Zhang, and Neda Abdelhamid. NBA game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116, 2019.
[22] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.
[23] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401, 2018.
[24] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
[25] Andrew Yiannakis, Michaël JP Selby, John Douvis, and Joon Young Han. Forecasting in sport: the power of social context—a time series analysis with English premier league soccer. *International Review for the sociology of sport*, 41(1):89–115, 2006.
[26] Paola Zuccolotto, Marco Sandri, Marica Manisera, and Rodolfo Metulini. Modelling basketball players' performance and interactions between teammates with a regime switching approach. *arXiv preprint arXiv:1912.10417*, 2019.
[27] Rongkai Duan and Pu Sun. Basketball sports neural network model based on nonlinear classification. *Journal of Intelligent & Fuzzy Systems*, 40(4):5917–5926, 2021.
[28] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.
[29] Qiyun Zhang, Xuyun Zhang, Hongsheng Hu, Caizhong Li, Yinping Lin, and Rui Ma. Sports match prediction model for training and exercise using attention-based lstm network. *Digital Communications and Networks*, 8(4):508–515, 2022.
[30] Jun Chen, R Dinesh Jackson Samuel, and Parthasarathy Poovendran. Lstm with bio inspired algorithm for action recognition in sports videos. *Image and Vision Computing*, 112:104214, 2021.
[31] Yu Zhao, Rennong Yang, Guillaume Chevalier, Rajiv C Shah, and Rob Romijnders. Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction. *Optik*, 158:266–272, 2018.
[32] Kuan-Chieh Wang and Richard Zemel. Classifying NBA offensive plays using neural networks. In *Proceedings of MIT Sloan Sports Analytics Conference*, volume 4, 2016.
[33] Swar Patel. NBA API. https://github.com/swar/nba_api.
[34] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 10–25080. Austin, TX, 2010.
[35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
[36] Masaru Teramoto and Chad L. Cross. Relative Importance of Performance Factors in Winning NBA Games in Regular Season versus Playoffs. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.
[37] Thomas L. Moore and Benjamin Kramer Johannsen. Keys to Success in a Run-and-Gun Basketball System. 2013.
[38] Lucas Freitas. Shot distribution in the NBA: did we see when 3-point shots became popular? *German Journal of Exercise and Sport Research*, 51:237 – 240, 2020.

[39] Mark Fichman and John O'Brien. Three point shooting and efficient mixed strategies: A portfolio management approach. *Journal of Sports Analytics*, 4:1–14, 2017.
[40] Blake Mclean, Donald Strack, Jennifer Russell, and Aaron Coutts. Quantifying Physical Demands in the National Basketball Association (NBA): Challenges in Developing Best-Practice Models for Athlete Care and Performance. *International Journal of Sports Physiology and Performance*, 14:1–22, 07 2018.
[41] NBA Communications. NBA Board of Governors approves play-in tournament for 2021-22 season, Jul 2021.
[42] Masaru Teramoto and Chad L Cross. Importance of team height to winning games in the National Basketball Association. *International Journal of Sports Science & Coaching*, 13(4):559–568, 2018.
[43] Andrew Fenichel. The modernization of NBA offenses and why small ball is here to stay, Jul 2022.
[44] Federico Bianchi, Tullio Facchinetti, and Paola Zuccolotto. Role revolution: towards a new meaning of positions in basketball. *Electronic Journal of Applied Statistical Analysis*, 10(3):712–734, 2017.
[45] Vjekoslav Perica. United they stood, divided they fell: nationalism and the Yugoslav school of basketball, 1968–2000. *Nationalities Papers*, 29(2):267–291, 2001.
[46] Jack W Porter. Predictive analytics for fantasy football: Predicting player performance across the nfl. 2018.
[47] Tobias Berger and Frank Daumann. Jumping to conclusions – an analysis of the NBA Draft Combine athleticism data and its influence on managerial decision-making. *Sport, Business and Management: An International Journal*, 07 2021.