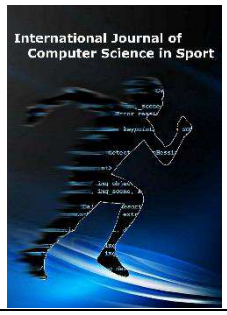
International Journal of Computer Science in Sport
Volume 24, Issue 1, 2025
Journal homepage: http://iacss.org/index.php?id=30
sciendo
DOI: 10.2478/ijcss-2025-0006



NBA Results Forecast: From League Dynamics Analysis to Predictive Model Implementation

Rodrigues F[1,2], Pires F[1]

[1]ISEP, Polytechnic of Porto, R. Dr. Aº Bernardino de Almeida, 431, Porto, 4249-015, Portugal.

[2]INESC TEC, Institute for Systems and Computer Engineering, Technology and Science, R. Dr. Roberto Frias, Porto, 4200-465, Portugal.

Abstract

This study presents a machine learning-based approach to predicting the outcosmes of NBA games, with the aim of enhancing decision-making in sports betting and performance analysis. Using a dataset spanning 20 NBA seasons (2003–2023), we incorporated key features such as team statistics, player performance metrics, and external factors like team fatigue and rankings. The methodology followed the CRISP-DM process, involving data preprocessing, feature selection, and model evaluation.

We experimented with multiple classification algorithms, including Logistic Regression, Random Forest, Gradient Boosting, and ensemble methods, to identify the best-performing models. Feature selection techniques such as LASSO and decision tree-based methods were employed to optimize model performance. Our best model, combining team rankings, statistics, and fatigue factors, achieved an accuracy rate of 64.1% and an F1 score of 72.4%, reflecting the complexity of NBA game outcome prediction.

The study highlights the importance of key features like team rankings and the challenges posed by the dynamic nature of the NBA. Future research will explore additional qualitative factors, such as emotional states and team dynamics, and employ more advanced machine learning techniques like deep learning to further improve prediction accuracy.

KEYWORDS: NBA, MACHINE LEARNING, GAME OUTCOME PREDICTION, FEATURE SELECTION, CLASSIFICATION.

Introduction

Basketball, in particular the National Basketball Association (NBA), has become a multi-million dollar industry. Sports betting in Portugal has experienced exponential growth since the introduction of Placard in 2015. According to the Gambling Regulation and Inspection Service, the authority responsible for granting licenses, there are currently 17 bookmakers legalized in Portugal, with 11 having sports betting modes. Between 2018 and 2022, sports betting mobilized 4 billion euros (Lusa, 2023).

Currently, various online sources, such as websites and mobile apps, offer free predictions for basketball game results. Examples include Vitibet (Vitibet, n.d.) and Forebet (Forebet, n.d.), which take into account the teams' form and the home/away nature of the game. However, these criteria may be insufficient to determine accurate results. Sportytrader (SportyTrader, n.d.) suggests betting based on expert guesses, going beyond the mere prediction of winners, and may include, for example, individual player performance. Regarding mobile apps, there are numerous platforms available, and the criteria are nearly identical to those found on websites. Our analysis reveals that there is a lack of effective decision-support solutions for NBA game betting in the market, which underscores the necessity for enhancements in the prediction of the outcomes of these events.

A vast body of literature exists on the topic of basketball match forecasting using machine learning, which is a highly explored area in sports prediction (Horvat & Job, 2020). A recent study by Horvat et al. (2023) proposed a ML model that combines league statistics and a custom team efficiency index, the Extended Team Efficiency Index (ETEI), created by expanding the index traditionally used in the NBA, with the aim of predicting league games. The case study utilized data from 6567 NBA games across 5 seasons. The authors' model outperformed traditional machine learning models, achieving an average score of 66% across over 2,500 games.

In order to predict NBA game results, Zheng (2022) proposed the introduction of both internal and external factors, including team ELO rankings, average team performance in recent games, the home factor, and player fatigue due to consecutive games. To achieve this, they used historical data from the 2012-13 season to the 2020-21 season. The study analyzed the performance of various ML models, including neural networks (NN), support vector machines (SVM), logistic regression (LR), random forest (RF), and naïve bayes (NB). The best achieved accuracy rate was 67.98% with the RF algorithm. After looking at the data, the author came to the conclusion that the new variables made the trained models more accurate, and RF and NB were the best algorithms for predicting the outcome of NBA games.

Zhao et al. (2023) proposed a method based on graph convolution networks (GCN) to predict NBA results. The study uses a homogeneous graph, where the nodes represent teams and the edges represent past/future games, connecting the teams to each other through game relationships. They explored three distinct approaches, each combining GCN with a specific feature selection technique. The first method used Principal Component Analysis (PCA) to reduce the size of the data. For characteristic selection, the second method used the LASSO technique, while the third method used RF. The study used a dataset with NBA statistics from 2012 to 2018. The conclusion highlights that the most effective combination was CGN with RF, reaching an average accuracy rate of 71.54%. On the other hand, the combination of PCA and CGN did not produce the expected results, suggesting a loss of essential data.

Cheng et al. (2016) proposed the creation of a model for predicting outcomes of NBA playoff games using the Maximum Entropy Principle. The suggested model was able to predict the winner of the match with a 74.4% accuracy rate, demonstrating superior performance compared to the other ML algorithms, such as neural networks (NN), Decision Trees (DT), and Naive Bayes (NB), which face challenges like overfitting due to limited datasets, lack of independence among features, and inability to provide interpretable probability values.

Horvat, Hava, and Srpak (2020) conducted a study whose main objective was to identify the best combination of ML algorithm, validation method, and data preparation to predict the outcomes of NBA games. The authors used supervised machine learning techniques on data from nine seasons, from 2009/2010 to 2017/2018. They used seven different algorithms: RL, NB, AD, multilayer perceptron NN, RF, K-Nearest Neighbors (KNN), and LogitBoost. They applied two different data preparation techniques: disjoint data, which completely separated the training and testing sets without sharing any information, and updated data, which added the known outcome data from the testing phase to the training set after each prediction. The updated data yielded the best results, as the KNN algorithm achieved an average accuracy rate of 60.01% and a maximum of 60.82%, utilizing one season for training and two seasons for testing.

The analysis of the various discussed works reveals the complexity and diversity of existing approaches. Each algorithm presented has its own advantages and limitations, emphasizing the importance of carefully choosing the algorithm most suitable for the specific context. Good estimation of basketball game result requires a global understanding of several internal and external factors, such as detailed game information, collective team statistics, ranking teams, the physical and emotional state of key team players. Additionally, because a team can only play a limited number of matches with another team in a single season, the training data for basketball outcomes is limited. Also, teams may transfer players at the end of each season. As a result, the key players fluctuate from season to season (Cai et al. 2019). In this context, it is difficult to predict basketball outcomes due to the high-dimensional, size-limited, and imbalanced dataset. The studies examined show a wide variation in model accuracy rates. The different data and experimental conditions of each study make objective and reliable comparisons difficult and controversial. These values, however, emphasize the difficulty in obtaining models with high and consistent scoring rates for predicting NBA game outcomes. We believe that optimizing three key factors is necessary to develop a high-performance basketball prediction model: 1) training data selection, 2) feature selection, and 3) algorithm selection. The aim of this research is to evaluate the effects of these factors on basketball prediction performance and obtain an optimized classifier that correctly predicts the team winner of a basketball game. Using the CRISP-DM methodology, we analyze the individual and collaborative effects of different features on basketball prediction performance. We developed several single and ensemble learning models and tested the models to assess the individual and collaborative effects of the features on prediction performance.

Methodology

2.1 Dataset

We used the NBA games dataset (Lauga, 2022), which provides detailed information about NBA games from the 2003/2004 season to the 2022/2023 season. The data includes details related to the games, such as the date of the game, the teams involved, the collective statistics of the teams, and whether the home team was the winner. It also contains the ranking of each team on each day throughout the season. The rankings are updated daily, reflecting the continuous performance of the teams throughout the season. The data includes the team's number of games, wins, and losses in the current season, not only overall but also specifically for home and away games. These data are important because they allow for the inference of a team's strength, as stronger teams have higher wins versus losses. In addition to this information, individual player statistics for each game were also collected. The statistics cover various performance parameters, such as points scored, assists, rebounds, minutes played, and shooting efficiency, among others. These data are critical because they enable the inclusion of the player's individual performance factor in the model.

The dataset covers a total of 20 seasons. It includes information on 26622 games played over this period, as well as statistics for 2686 different players, in addition to the 30 NBA teams. These numbers provide a solid foundation for statistical analysis and predictive modeling, allowing a detailed view of the dynamics and performances in the NBA over two decades.

2.2 Data Processing

2.2.1 Information about the NBA games

We started by eliminating columns with unique values. Next, we checked for missing values in some columns related to game statistics. Since these rows corresponded to preseason games and represented a small sample relative to the complete dataset, they were removed. Boxplot graphs were created for the columns representing game statistics, and through the analysis of these graphs, some games were discovered where some statistics were extremely low or extremely high, particularly the points scored by the home team and the visiting team. Consulting the NBA official website (NBA, n.d.), it was possible to determine that the extremely high statistics were from games with one or more overtimes, while the extremely low statistics were from preseason games that lasted only 30 minutes instead of the usual 48 minutes. Preseason games primarily serve the purpose of testing new tactics and evaluating the potential of recently acquired players at a slower pace and with a reduced workload. Additionally, the best players tend to limit their participation in these games to avoid risking injuries that could compromise the start of the season (AS, 2023). We removed preseason games from the data sample due to their small sample size and distinct dynamics.

Each row in this file represents the statistics of both teams, including points scored, shooting percentage, free throw percentage, three-point shooting percentage, assists, and rebounds, for each game played. Since it would be impossible to access these data when predicting future games, it was necessary to transform these columns to not only represent the current form of the teams but also to provide information available before the game occurs. Therefore, we calculated the average over the last 4 games for each of these statistics, including points conceded. Additionally, we created columns that calculated these averages solely based on the home team's last 4 home games and the visiting team's last 4 away games. The choice regarding the number of games to use for calculating the averages was based on the work of Chen et al. (2021), which concluded that the number of games that yielded the best results was 4.

We found outliers in the data during the boxplot visualization, which could distort the analysis and compromise assumptions, but we decided not to remove them. Removing the games where these outliers occur would affect the calculation of the average statistics over the last four games, compromising the integrity of the data. Thus, keeping these outliers allows for a more faithful analysis of reality.

2.2.2 Information about the Teams

In the NBA, coaches play a crucial role in team performance. A coach's ability to develop effective strategies, adjust tactics during games, develop players, and create team chemistry can differentiate between a mediocre team and a successful one (Bishop, 2023) (Hoop Social, 2023). Therefore, the Head Coach column, which represents the team's current coach, seemed valuable. However, because the column solely reflected the current coach, it was unable to accurately analyze the entire temporal scope of the dataset, spanning 20 seasons. During this period, there were several coaching changes in the teams. We conducted searches on various websites to obtain accurate information about these changes between seasons and even during the ongoing seasons. Although it was possible to find information about coaching changes, the specific dates on which these changes occurred are not available, which would be an

obstacle in introducing accurate information about the coaches into the dataset. We ultimately removed this column as a result. Another column was removed, indicating whether the team has a B team. The removal of the column occurred because it predominantly contained the same value, despite only 3 out of the 30 teams lacking a B team.
The information about wins and losses for each team was combined, so we derived four columns representing the number of home wins, home losses, away wins, and away losses, respectively. The '*conference*' column represents the team's conference location, and due to dummies' requirements, two new columns, *'is_east_conference'* and *'is_west_conference'*, were created.

2.2.3 Team Rankings

The day's ranking already includes the games played that day. As a result, a team's ranking when entering the field should be the previous day's ranking. Since the previous day's ranking relied on preseason information, we set all values related to game count, wins, and losses to 0 on the first day of each season.
The comment column contains an explanation for the absence of a player from the game if they did not actually play. Therefore, we deleted all rows that contained records of players who did not participate in the game. This operation eliminated the column that only contained missing values.
Each team consists of two guards, two forwards, and one center. Therefore, in an NBA game, four guards, four forwards, and two centers should start. We ensured that all games met these requirements by taking this into account. When analyzing the distribution of the '*start_position'* column, the value "C" (center) should have half the occurrences of the values "G" (guard) and "F" (forward), and the occurrences of the last two should correspond to a quarter of the total number of games in the file. This distribution was confirmed. The NaN values in this column represent players who participated in the game but were not starters. To fill in these missing values and give them meaning, they were filled with the value "*didn't start*". Table 1 displays the distribution of this column post-treatment.

Table 1: Distribution of the start_position column after treatment

Value	Frequency	Percentage
DIDN'T START	279367	52,2%
F	102260	19,1%
G	102260	19,1%
C	51130	9,6%

We transformed the column, which represented the player's minutes played in text format MIN:SEC, to only contain the minutes' value. Finally, we verified through boxplots whether there were any incongruent values in the columns related to player statistics, which we did not observe.
We discovered that the teams' rankings remained unchanged after the conclusion of the regular season. From the moment the playoffs began, the number of games, wins, and losses by the teams ceased to change. Therefore, it was necessary to calculate for each team the number of games, number of wins, number of losses, and win percentage, both overall and specifically for home games for the home team and away games for the visiting team. Additionally, we calculated each team's current win/loss streak, specifically focusing on home games for the home team and away games for the visiting team. These operations are critical to ensuring that the dataset accurately reflects the performance of each team throughout the season, not just until the end of the regular season. The inclusion of the win/loss streak is

particularly important as it allows for a more detailed analysis of the teams' current form. Positive or negative streaks can significantly influence a team's morale and confidence, affecting their performance.

The comparison of the original columns with the newly created ones revealed some erroneous data at this point. We discovered the absence of some games from the 2022/2023 season. We removed the games from this season from the dataset, as the dataset only included data from this season up until December 22, 2022, and some games were missing.

Additionally, COVID-19 caused a significant disruption during the 2019/20 season, including a mid-season interruption and a small preseason before the competition resumed. This preseason introduced additional information in the ranking records, which did not reflect the teams' actual performance prior to the interruption. To address this, we corrected the ranking records on the day of the competition's return to align with the information prior to the COVID-19 interruption, ensuring they accurately represented the teams' pre-interruption performance. Furthermore, we eliminated games from the dataset associated with the brief preseason to prevent bias in the training or testing process. This correction was crucial for maintaining the integrity of the data and ensuring the model's predictions were not influenced by the anomalies introduced during this disrupted period.

2.2.4 Team Fatigue

Zheng (2022) demonstrated the importance of player fatigue to team performance. Therefore, we introduced the fatigue factor into the dataset. We incorporated this factor by calculating the following characteristics:

- Number of games in the last 7, 10, and 14 days: This metric is fundamental for assessing the intensity of a team's schedule and quantifying the physical wear accumulated in a short period of time. Teams with a higher volume of games in a short period may show signs of fatigue that affect performance.
- Number of away games in the last 7, 10, and 14 days: Frequent travel for away games can increase team fatigue due to the time spent traveling and lack of adequate rest. This metric helps identify teams that may be at a disadvantage due to a demanding schedule of away games.
- Current sequence of consecutive away games: Playing consecutively away from home can be particularly exhausting, as teams do not have the opportunity to recover in their usual environment and need to travel repeatedly. This characteristic enables the identification of critical periods during which accumulated fatigue and constant travel may impact team performance.
- Average number of players used in the last 4 games: Teams use player rotation as a strategy to mitigate fatigue. Teams that use more players can better manage the workload and maintain more consistent performance. This metric assesses the depth of the roster and the ability to manage fatigue.
- Average minutes per player in the last 4 games: Players who accumulate many consecutive playing minutes tend to become more fatigued, which can harm their performance. This characteristic helps identify teams under greater physical stress.
- Average minutes played by starters in the last 4 games: Starters usually play more minutes and have a greater impact on team performance. Monitoring the minutes played by starters is crucial for assessing the fatigue level of the team's key players.

With these characteristics included in the dataset, it is possible to perform a more complete and accurate analysis of the impact of fatigue on team performance, helping to identify patterns and predict potential drops in performance.

2.2.5 Other Team Statistics

Although the game information file includes team statistics, some important statistics were missing. However, it was possible to calculate these statistics by summing the corresponding values for each player who participated in the game. We performed this sum using the player statistics file, which included not only the game file's statistics but also additional ones. From this file, we calculated the following statistics:

- Offensive and defensive rebounds: We decided to separate the rebounds into two distinct characteristics, even though the game information file already included the teams' total rebounds. Offensive rebounds are crucial because they provide a second chance to score, even after a missed shot, thereby increasing the likelihood of success (Christos et al. 2020). On the other hand, defensive rebounds allow a team to regain possession of the ball, reducing the opponent's scoring attempts and opening the way for a counterattack (Christos et al. 2020). In summary, offensive and defensive rebounds are indicators of the teams' offensive and defensive capabilities, respectively.
- Steals: Steals not only provide more offensive possessions, which can result in more points, but they also disrupt the opponent's possession of the ball, allowing for counterattacks (Goodman, 2014). These are indicators of a team's defensive ability.
- Blocks: Blocks interrupt the opponent's shot attempt. They allow for an assessment of a team's ability to defend its basket.
- Turnovers: When a team loses possession of the ball, it wastes an opportunity to score. Additionally, it gives the opponent the opportunity to score through swift breaks. Teams with a higher number of turnovers are generally teams where the offensive process is not as well defined (Hoop, 2024).
- Fouls committed: The number of fouls can be an indicator of both defensive aggressiveness and discipline. If a team commits a high number of fouls, it will frequently benefit the opponent with free throws, which is a negative point. However, a moderate amount of fouls can indicate aggressive and physical defense.
- Field goal attempts and successful field goals: Generally, teams that make more field goal attempts play at a faster pace, leading to a higher workload for the opponent's defense, causing fatigue and possible mistakes. The number of field goal attempts becomes even more important if the team can successfully convert a significant portion of these attempts, as teams with this ability have a substantial competitive advantage.
- Three-point attempts and successful three-pointers: A high number of three-point attempts may indicate that a team is capable of creating space for its best shooters to take uncontested shots. On the other hand, it may also indicate that the team is struggling to penetrate the opponent's defense and get close to the basket, especially if the success rate of these shots is low. Three-pointers offer an advantage compared to two-point shots because they allow for scoring more with fewer attempts. Even with lower accuracy in three-point shooting, a team can score the same or even more than with two-point shots (CoachAD, n.d.).
- Free throw attempts and successful free throws: A high number of free throw attempts means that a team has been fouled often. These shots give teams an opportunity to score points without the pressure of the opponent's defense and are critical because they can make a big difference in the final result, especially in close and narrow-margin games.

Before calculating the aforementioned statistics, we performed a verification to identify any

potential incorrect data. We identified discrepant values during this analysis, particularly in the 2003/2004 and 2004/2005 seasons, which led to the removal of games corresponding to those seasons. We made the necessary corrections to the remaining games by consulting the NBA official website to validate and update the necessary data. Table 2 displays the distribution of games won by the home team and the visiting team following this correction.

Table 2: Home_team_wins after 2003/2004 and 2004/2005 seasons removal

Value	Frequency	Percentage
Yes	12814	58,8%
No	8972	41,2%

We calculated their average over the last four games, both for the home team and the visiting team, since it would be impossible to obtain the values of these statistics before the game. We also calculated their average over the last four home games for the home team and the last four away games for the visiting team.
Following the several data processing steps, we conclude with a dataset consisting of 119 features and 21786 lines.

2.3 Data Exploration

We will explore the data through analyses that investigate various quantitative and qualitative aspects of the NBA, aiming to understand the league's dynamics and transformations over time. We will conduct four analyses, each focussing on a distinct aspect of the game and its influence on players and teams, namely:

1. The evolution of the NBA over the seasons: This analysis highlights the main changes and trends that have shaped the style of play.
2. The impact of team changes on player performance: This examines how transfers and trades influence individual and collective performance.
3. The evolution of positions in the NBA over the years: involves observing how these roles have evolved and adapted to the new demands of the game.
4. The evolution of player performance with age: Comprises evaluating how factors related to age influence the performance and longevity of athletes' careers.

We emphasized that we supplemented the dataset for these analyses with information from two additional sources. We obtained data from the 2003/2004 and 2004/2005 seasons, previously removed from the dataset due to a high number of games with incorrect statistics, from Basketball-Reference (Basketball-Reference, n.d.). We also recovered data from the 2022/2023 season, which we had previously excluded due to its incompleteness. Additionally, we collected detailed player statistics by season, which included age, games played, average minutes per game, average points per game, and more, to support analyses on the NBA's evolution over time, the impact of team trades, and the evolution of player performance with age. In order to analyze the evolution of positions in the NBA, we acquired the averages of statistics per position on the court for each season of the study period from StatsMuse (n.d.).

2.3.1 Evolution of the NBA over the seasons

Here we analyze the evolution of the NBA's game dynamics over the past 20 seasons, focusing on changes in offensive and defensive strategies. The analysis reveals a significant increase (22,8%) in the average points per game, driven primarily by a rising frequency of shot attempts (10,65%), particularly three-pointers (129,53%), and an improvement in team play, as reflected by the increase in assists (18,78%).

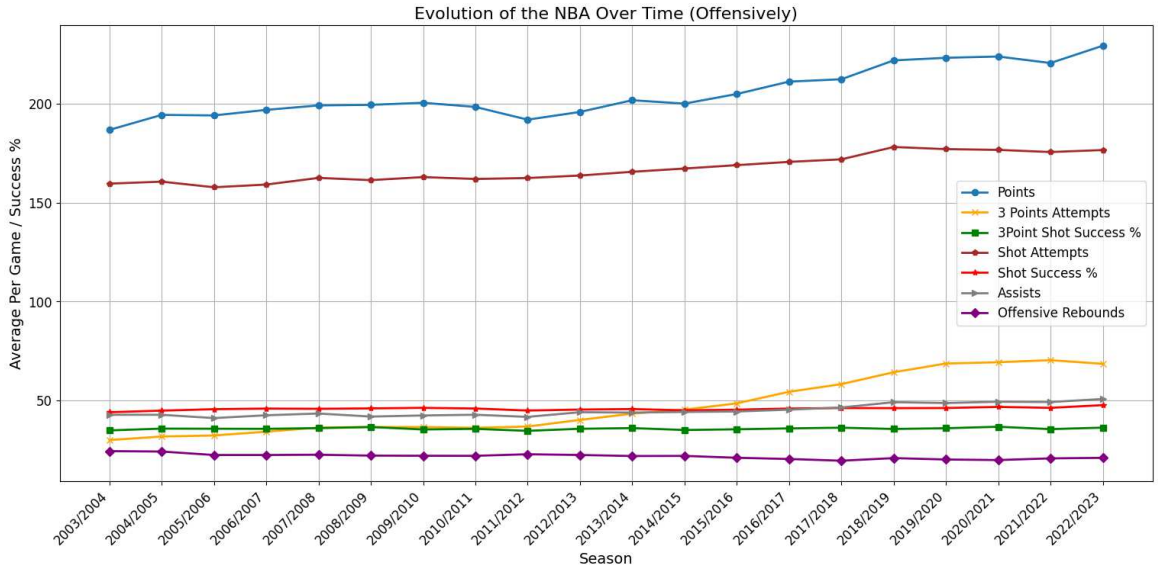


Figure 1-NBA evolution offensively

Offensively (Figure 1), the NBA has transitioned to a faster style of play with more shot attempts, emphasizing long-distance shots, without significant changes in shooting success rates.

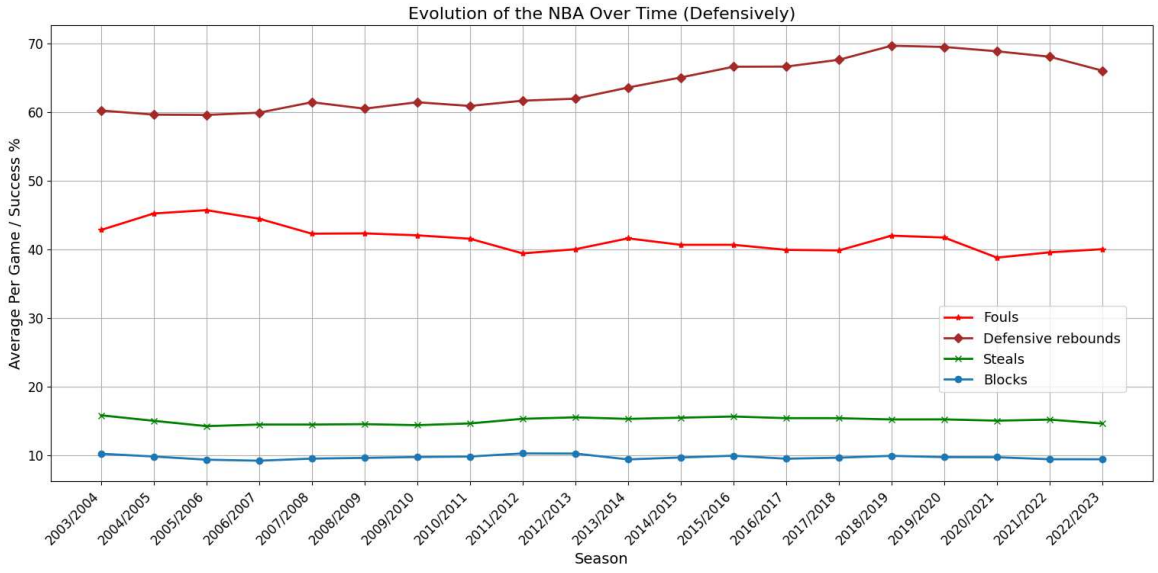


Figure 2-NBA evolution defensively

Defensively (Figure 2), there has been an increase in defensive rebounds (9,63%) and a reduction in fouls committed (6,54%), suggesting improved control of the game without compromising defensive aggressiveness. Overall, the analysis highlights the transformation of contemporary basketball towards a more collective, efficient game focused on three-point shooting, with refined defensive strategies that balance aggressiveness and consistency.

2.3.2 Impact of Team Trades on Player Performance

Here, we examine the impact of team trades on NBA player performance over the past 20 years, focusing on trades made during a season and between seasons. We analyzed 1126 trades made during the season and 800 trades made between seasons. Five key metrics were

analyzed: offensive contribution (OFC), defensive contribution (DFC), minutes played, games played, and matches started.

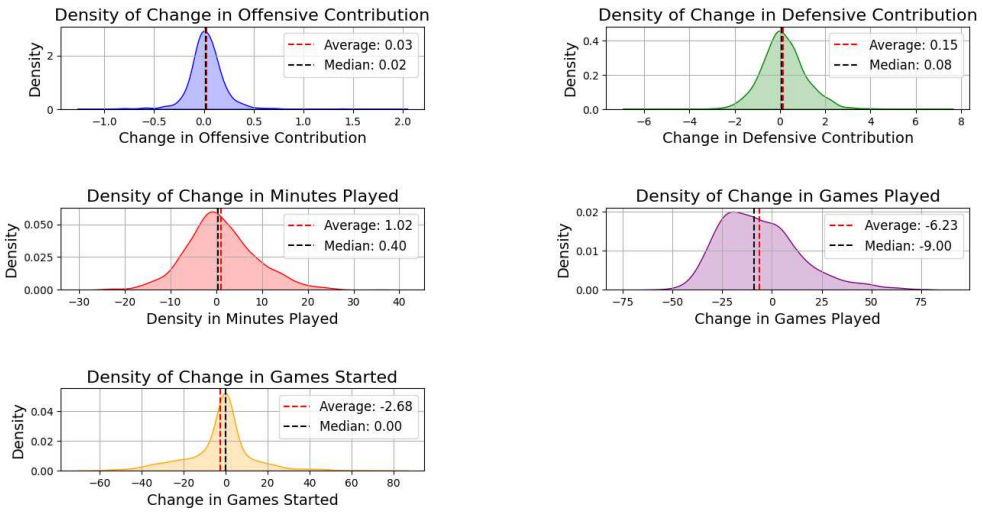


Figure 3-Impact of In-Season Team Trades

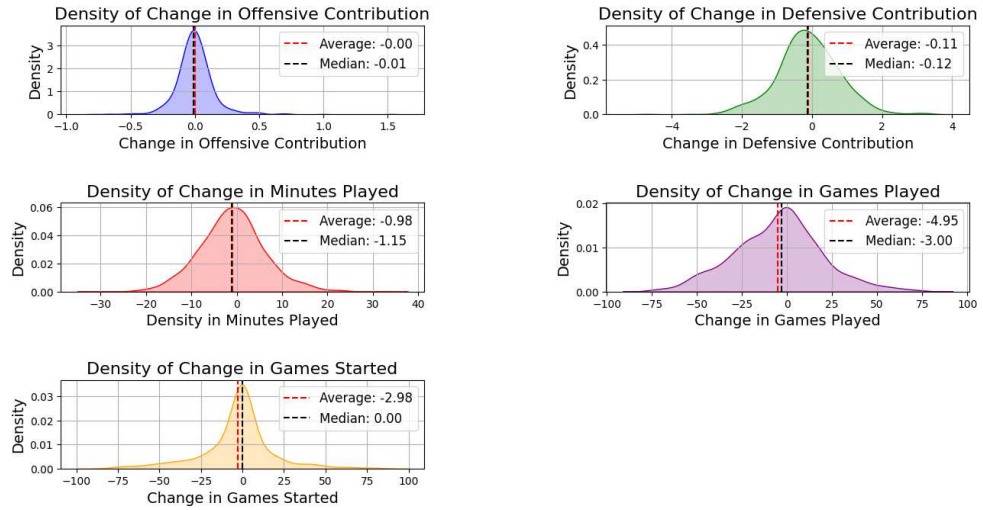


Figure 4-Impact of Team Trades between seasons

Trades made during a season (Figure 3) showed minimal changes in offensive (on average, players experienced an increase of 0.03 in offensive impact, with a median improvement of 0.02) and defensive contributions (average increase of 0.15 and median improvement of 0.08), indicating players generally maintain consistent performance. However, there was a significant decrease in games played (on average, players participated in approximately 6 fewer games, with a median decrease of 9 games), suggesting reduced participation post-trade. Similar stability was seen in offensive contributions (on average, there was no change in offensive impact, while the median showed a small decrease of 0.01), but there was a small decrease (on average, there was a decrease of 0.11, with a median decrease of 0.12) in defensive contributions and games played (on average, players played in about 5 fewer games, with a median decrease of 3 games). Overall, while trades can influence player participation, individual performance in terms of contribution to the game tends to remain stable, reflecting

players' adaptability to new teams.

2.3.3 Evolution of Positions in the NBA Over the Seasons

The analysis of the evolution of the five NBA positions from 2003/2004 to 2022/2023 focuses on key metrics such as points per game (PPG), assists (APG), field goal attempts (FGA), field goal percentage (FG%), three-point attempts and percentage (3PA and 3P%), offensive rebounds (OREB), defensive rebounds (DREB), blocks (BPG), and steals (SPG). The goal is to understand how each position has adapted to changes in strategies and modern gameplay, considering only starting players to ensure accuracy.

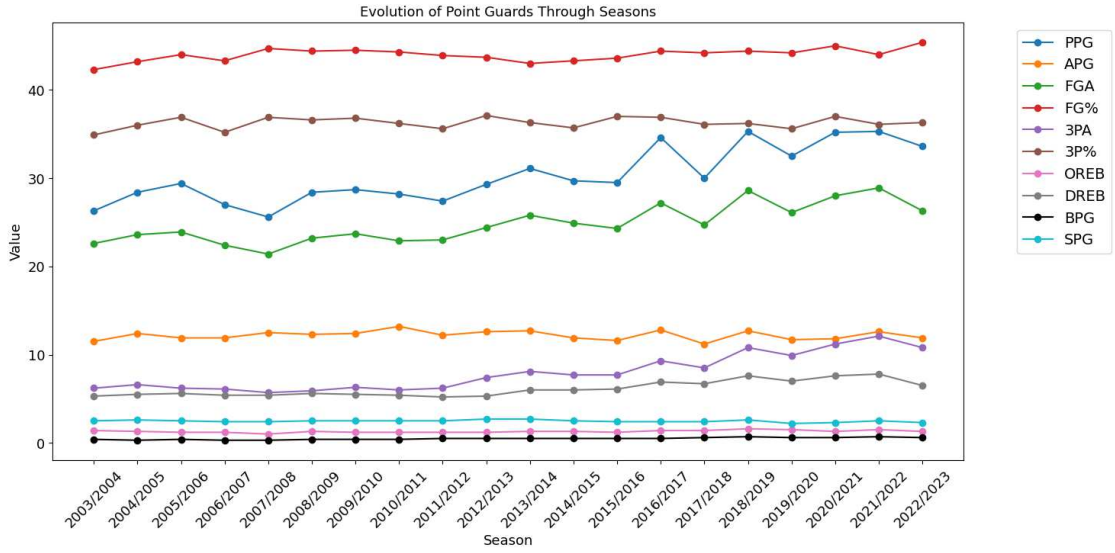


Figure 5-Evolution of Point Guards

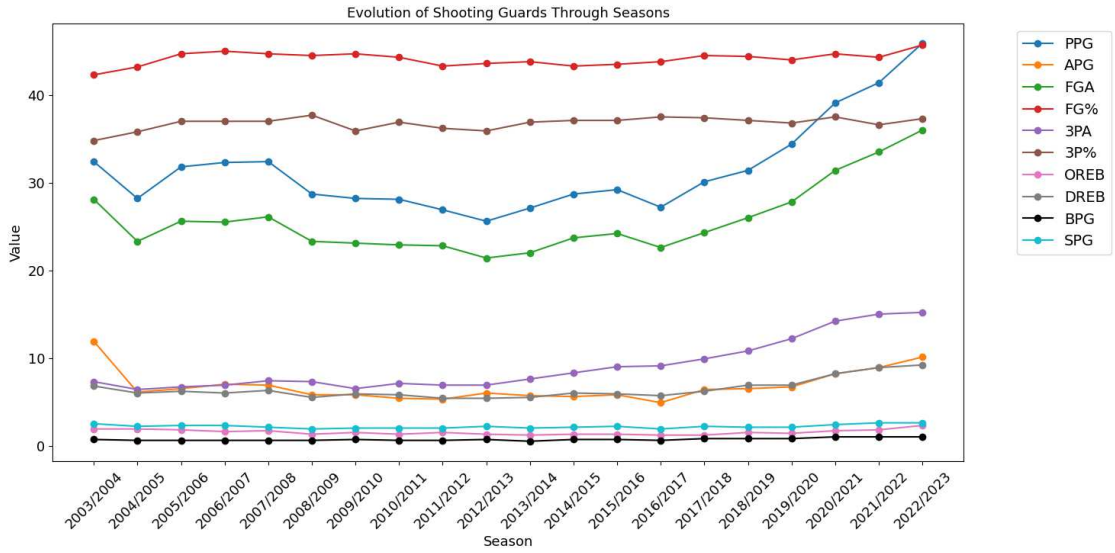


Figure 6-Evolution of Shooting Guards

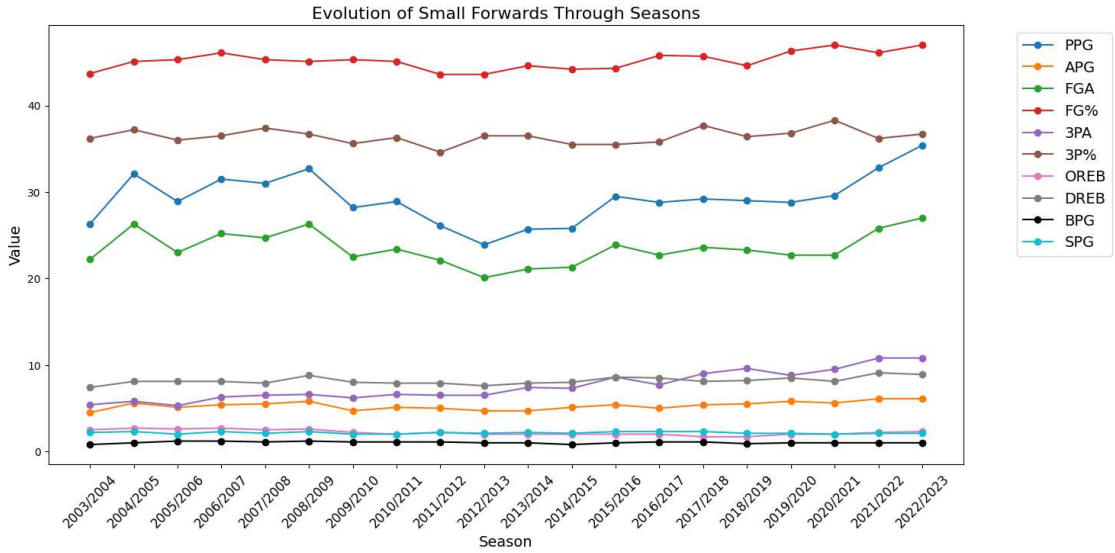


Figure 7-Evolution of Small Forwards

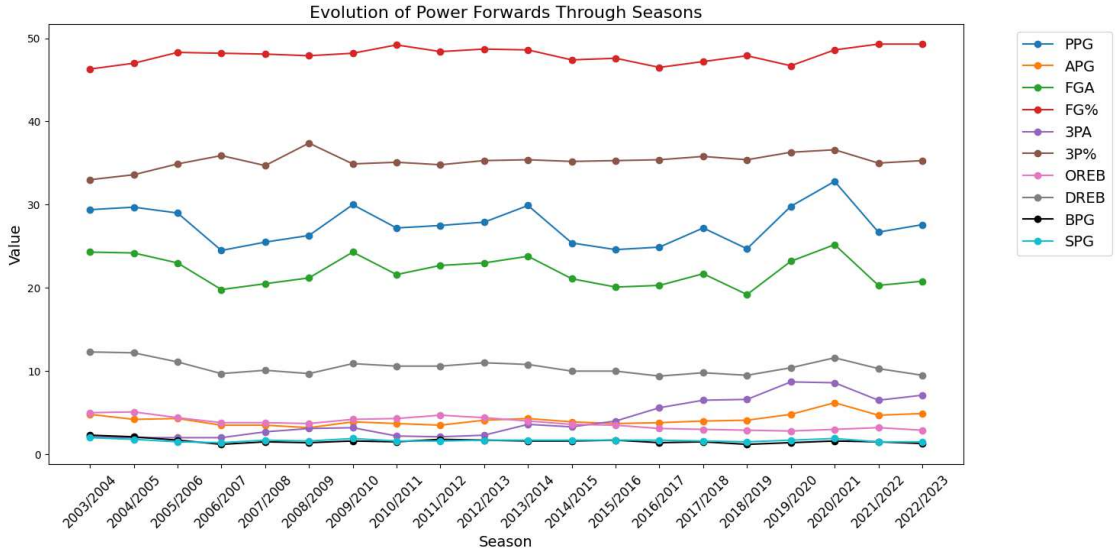


Figure 8-Evolution of Power Forwards

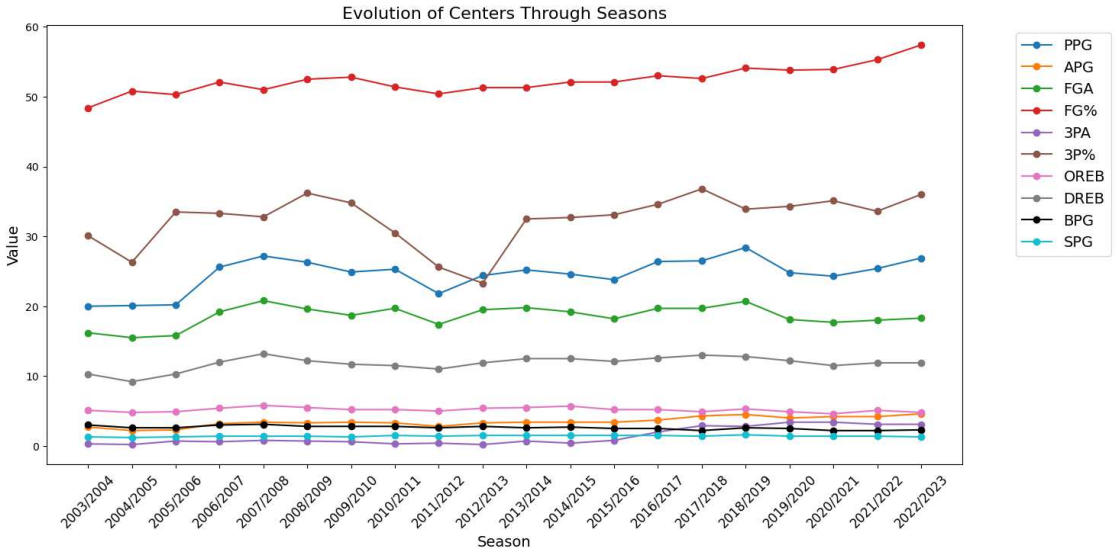


Figure 9-Evolution of Centers

The evolution of basketball positions has led to a notable shift toward greater offensive capabilities and versatility. Point Guards (PG) (Figure 5) and Shooting Guards (SG) (Figure 6) have increasingly focused on offensive play (PG experienced a 27.8% increase in points per game, while SG saw an increase of 41.7%), characterized by significant rises (74.2% for PG and 108.2% for SG) in three-point attempts and modest improvements (2.2% for PG and 3.4% for SG) in shooting efficiency. Small Forwards (SF) (Figure 7) have also transitioned to more offensive roles (34.7% increase in points per game), exhibiting a 100% increase in three-point attempts and enhanced overall shooting efficiency (3.3%). Power Forwards (PF) (Figure 8) have adapted by incorporating more long-distance shooting (with an increase of 23.8% in 3 Points Attempts) and demonstrating improved shooting efficiency (3%), though there has been a slight decline in defensive rebounds (22.7%). The Centers (C) (Figure 9) have shown substantial improvements in shooting efficiency and versatility, with notable increases in shooting accuracy (9%), three-point attempts (93.3%), and defensive rebounds (15.5%). Collectively, these positional shifts reflect a broader trend toward a more dynamic and flexible style of play, emphasizing three-point shooting and multifaceted skill sets. In conclusion, overall, NBA positions have become more versatile over the analyzed period, with increased offensive capability and adaptation to modern gameplay. Players have evolved to focus more on three-point shooting and multifaceted skills, reflecting a trend toward a more dynamic and flexible style of play.

2.3.4 Player Performance with Age

This analysis examines NBA player performance across their NBA carrer using data from 20 NBA seasons. Key metrics analyzed include not only metrics related to players' participation like games played, games started and minutes played, but also stats that measure their efficiency and contribution in multiple game contexts, including field points, assists, rebounds, etc. Only players with an average of at least 5 minutes per game were included to avoid skewing the results with marginal players.
The age groups in years considered are '18-23', '24-27', '28-30', '31-33', '34-36', '37-39', and '40+'.

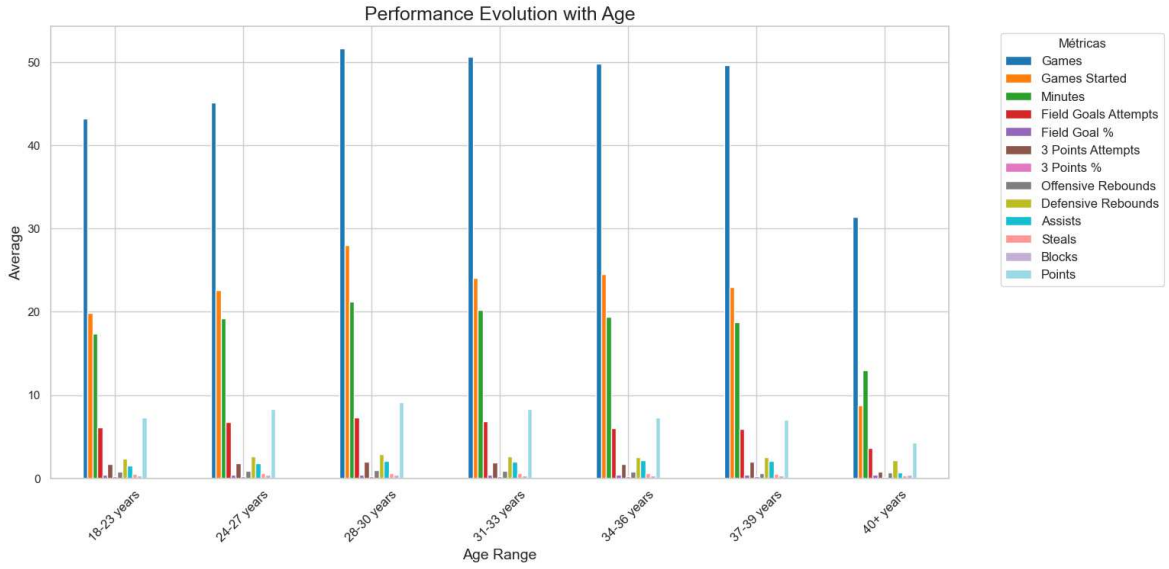


Figure 10-Performance Evolution with Age

Table 3 – Key metrics by age range

Age	Freq	%	Games			Field Goals		3 Points		Defensive Rebounds	Assists	Points
			Num	Start	Min	Attemps	%	Attemps	%			
18-23	1338	35.4	43,2	19,8	17,4	6,1	43,2	1,7	28	2,3	1,6	7,3
24-27	1264	33.5	45,1	22,6	19,2	6,7	44	1,9	28,1	2,6	1,8	8,3
28-30	566	15.0	51,7	28	21,2	7,3	44,7	2	28,9	2,9	2	9,1
31-33	356	9.4	50,6	24	20,1	6,8	44,1	1,8	27,7	2,6	2	8,3
34-36	174	4.6	49,8	24,4	19,4	6	45,2	1,7	29,4	2.5	2,1	7,3
37-39	66	1.7	49,6	23	18,8	6	41,6	2	28,9	2,5	2	7
40+	13	0.3	31,4	8,8	13	3,7	46	0,8	20,5	2,2	0,7	4,3

During ages 18-23, players show limited field presence, with fewer games, starts, and minutes played, reflecting their adjustment to the NBA. This group includes 1338 players, indicating many young entrants. From ages 24-27, players experience increased performance metrics, showing maturity and stability. Games, starts, and minutes rise, with slight increases in field goal attempts and points. There are 1264 players in this range, suggesting successful adaptation and retention in the league. Ages 28-30 mark peak performance with 566 players. Here, players leverage experience and peak physical condition, showing the highest results in games, starts, minutes, and scoring. The number of players declines, possibly due to teams favoring younger talents. After age 30, performance gradually declines. Metrics like minutes, games, and starts decrease, and scoring begins to drop, reflecting physical decline. The number of players drops significantly, with 356 between 31-33, 174 between 34-36, and 66 between 37-39, indicating a challenging league environment for older players. At age 40 and above, performance metrics drop sharply, with fewer games, starts, and minutes. Scoring is at its lowest, reflecting reduced physical capabilities. This age group has only 13 players, highlighting the rarity of maintaining high-level performance at such an advanced age. Overall, NBA players generally reach peak performance between ages 28-30. After this, there is a gradual decline, becoming more pronounced after age 40, reflecting natural aging and the need for role adjustments within teams.

2.4 Modeling

The prediction models were trained using a systematic and comprehensive approach, with the goal of predicting whether the home team will win an NBA game, by selecting the best

algorithms and optimizing them to achieve the maximum performance.
Initially, we divided the data into 70% for training and 30% for testing. The training set is comprised of approximately 12 NBA seasons, which include 15250 games played between the 2005/2006 and 2015/2016 seasons, as well as a significant portion of the 2016/2017 season. The test set spans approximately 5 seasons and includes 6536 games from the 2017/2018 to 2021/2022 seasons, and the remaining portion of the 2016/2017 season, not included in the training set. It is important to mention that the games are chronologically ordered from the oldest to the most recent; because sporting events are not completely independent, this detail is essential.
The training period (2005/06 to some time in 2016/17) was selected to ensure the model was exposed to a substantial volume of historical data, allowing it to learn diverse patterns in game results and team performance. We chose this period to ensure stable training conditions by providing consistent data with minimal interruptions or anomalies. The testing period (remaining from 2016/17 to 2021/22) was deliberately selected to evaluate the model's performance across two distinct temporal contexts: the first seasons served as an immediate testing set to assess the model's generalisation to recent data from the same era as the training period. The seasons after the COVID-19 interruption were included to evaluate the model's robustness against long-term changes in gameplay, team composition, and any unforeseen influences, such as post-COVID-19 impacts or rule changes.
Finally, we normalized the data in both sets, a crucial step for the KNN and SVM algorithms that are sensitive to the variable scales.
We performed cross-validation only with the training set, utilizing a variety of base algorithms such as LR, DT, NB, KNN, and SVM, along with ensemble learning techniques like Stacking, Voting Classifier, Bagging (Bagging Classifier and RF), and Boosting (AdaBoost, GB, XGBoost, and LGBM). We determined the base algorithms for the Stacking and Voting Classifier algorithms by analyzing the correlation between them and selecting those with the lowest correlation.
An exhaustive evaluation of the models' performance was carried out, utilizing the criteria of accuracy, precision, recall, and F1 score. The accuracy rate is the proportion of games successfully predicted by the model given the total number of games forecasted. The precision metric quantifies the proportion of games in which the model accurately projected the home team to win, and indeed, the home side did win. The recall metric quantifies the proportion of home team victories accurately detected by the model per the total number of home team victories in the dataset. The F1 measure is calculated as the harmonic mean of precision and recall. This metric serves as a comprehensive evaluation of the model by considering both its capacity to predict accurate victories and its ability to identify actual wins. Subsequently, we chose the top five algorithms with the highest F1 score measures. The RandomizedSearchCV technique was then employed to further optimize the algorithms' parameters, with the objective of improving their performance. The subsequent parameters were optimized for each algorithm:

- LR: C, penalty, solver, max_iter
- DT: max_depth, min_samples_split, min_samples_leaf, max_features, criteria
- NB: not applicable
- KNN: n_neighbors
- SVM: kernel, C
- Bagging: estimator, n_estimators, max_samples, max_features
- RF: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features
- AdaBoost: not applicable, as it never ranked among the top 5 algorithms.
- GB: learning_rate, n_estimators, max_depth

- XGBoost: Eta, max_depth, min_child_weight, subsample, colsample_bytree, n_estimators
- LGBM: num_leaves, learning_rate, n_estimators, max_depth, min_child_samples

It is important to note that for the Stacking and Voting classifier algorithms, the parameters of their base algorithms were optimized. Following optimization, we conducted a new cross-validation using only the training set and the best algorithms with optimized parameters. Finally, we selected the three best algorithms based on the previous stage's results, and performed validation using the holdout technique with the optimized parameters.

Two iterations of model training were performed following the described methodology. In the first iteration, models were trained with different datasets to evaluate the impact of attributes on the prediction of the winning team. In the second iteration, model training was repeated with the same datasets, but this time applying feature selection algorithms. By following this rigorous methodology, we were able to identify the features relevant to discern the winning team and the best prediction models for the problem at hand, ensuring their robustness and ability to generalize to future data.

2.4.1 First Iteration

In this first iteration, models were trained with four different datasets:

- The first dataset included the averages of team statistics, both overall and in home games for the home team and away games for the visiting team.
- The second dataset included all the features from the first dataset, along with information on team rankings.
- The third dataset included all of the second's data, as well as the team fatigue factor.
- The fourth dataset includes all the features from the third dataset, along with additional team statistics.

The performance of the models obtained with all four datasets is described in Table 3.

Table 3: Best models resulting from first iteration

Dataset	Model	Accuracy	Precision	Recall	F1
First	LR	0.596	0.607	0.823	0.698
	LGBM	0.586	0.594	0.858	0.702
	GB	0.580	0.592	0.844	0.696
Second	VCH (NB, LR)	0.623	0.687	0.619	0.651
	VCS (NB, LR, SVM)	0.634	0.664	0.720	0.691
	VCS (NB, LR)	0.627	0.669	0.677	0.673
Third	VCH (NB, LR)	0.621	0.693	0.597	0.641
	VCS (NB, LR, KNN)	0.637	0.669	0.714	0.691
	VCS (NB, LR)	0.630	0.677	0.666	0.671
Fourth	VCS (NB, LR)	0.630	0.691	0.626	0.657
	VCS (NB, LR, KNN)	0.633	0.678	0.669	0.674
	VCH (NB, LR)	0.619	0.705	0.562	0.625

With the first dataset, which includes information about the games and information about the teams, we achieved a precision of around 59%–60%. At this stage, the metrics of the best models are quite low, except for recall, which lies between 82% and 86%, suggesting that models may be classifying almost all instances as a home team victory.

The second dataset, which incorporates team rankings into the first dataset, yielded a 6–8%

improvement in precision. All models in this phase have superior accuracy and precision compared to the models from the initial dataset, indicating a clear improvement. Therefore, team ranking can be regarded crucial for the efficacy of predictive models.
The third dataset adds information about team fatigue to the previous one. This iteration shows a very slight, practically negligible increase in precision, ranging from 67% to 69%, but it does not coincide with an increase in recall, resulting in identical F1 measures in both iterations.
The fourth dataset, besides all the information from the previous dataset, includes additional team statistics. Also, the improvement obtained with the addition of these new features is debatable since the accuracy remains the same, the precision shows a very slight increase, varying between 69% and 70%, and the recall and consequently the F1 measure are very variable.
The analysis of the performance of the best models in these four iterations clearly demonstrates the effectiveness of the information about the games and the teams added with team rankings in predicting the winning team. In terms of the remaining features added in the third and fourth iterations, team fatigue and more team statistics did not significantly contribute to an improvement in prediction performance. This leads us to consider new attribute combinations using feature selection algorithms.

2.4.2 Second Iteration

In the second iteration, we repeated the first iteration's training with the same four datasets, but previously we made feature selection in each dataset using three techniques:

- LASSO: where relevant features are those whose coefficients are not null.
- DT-based method: selects features according to their significance.
- Feature shuffling: assesses the effect of eliminating each feature on the model's performance.

We applied each of these techniques to the four datasets from the previous iteration, then performed cross-validation on each of the four resulting subdatasets using the ML algorithms: Bagging, GB, XGBoost, and LGBM, to select the datasets with the best predictive capabilities. Of the three feature selection techniques used, the DT-based technique selected the features that produced the best results across all four datasets. Next, we trained the models using these best datasets, adhering to the same methodology as the previous iteration. Table 4 describes the best model's performance with each dataset.

Table 4: Best models obtained in the second iteration

Dataset	Model	Accuracy	Precision	Recall	F1	Number of Features
First	LR	0.614	0.619	0.828	0.708	14
Second	LGBM	0.631	0.632	0.835	0.719	29
Third	LR	0.641	0.641	0.831	0.724	41
Fourth	VCS (NB, LR, KNN)	0.634	0.683	0.659	0.671	61

The second dataset's model, with F1 of 71.9% and accuracy of 63.1%, trailed the third dataset's model in accuracy (64.1%) and F1 (72.4%). Despite achieving the highest precision of 68.3%, the model associated with the fourth dataset exhibits the lowest F1 and recall scores, with the latter being notably lower than those achieved by the other models.
The third dataset yielded the highest accuracy and F1 among all evaluated models, making it the best choice in terms of performance and efficiency. Although it had lower precision than the model obtained with the fourth dataset, it significantly outperformed the latter in recall and F1. Although it uses more features than the models obtained with the first and second

datasets, its superior performance justifies its choice.
Additionally, this model represents an improvement compared to the best model from the previous iteration because, despite the slightly lower precision, it shows better accuracy, recall, and F1 using only 41 features, representing a reduction of about 45% in the original number of features. This reduction makes the model simpler and easier to use. Thus, the third dataset was chosen because it yielded the best results.

2.5 Deployment

The model was made available on the website (https://nbaprevision.streamlit.app/) using the Streamlit framework (Streamlit, n.d.). This framework allows for the rapid development of interactive web applications with minimal code. The Streamlit platform integrates with GitHub, automatically installs all dependencies, and makes the application available online for free.
The architecture of the developed web application is simple and integrated. The Python script requests data from the NBA-API, processes it, and transforms it to feed the prediction model. Any device with web browsing software can access the application. The interaction with the user takes place through an intuitive web interface created with Streamlit. Users are required to input the specific date for the desired forecasts. Failure to find any games planned for that day will result in the following message: "*No games found for the chosen date.*" Alternatively, Figure 11 depicts the forecasts for the games scheduled on a specific day, such as 24/10/2023.

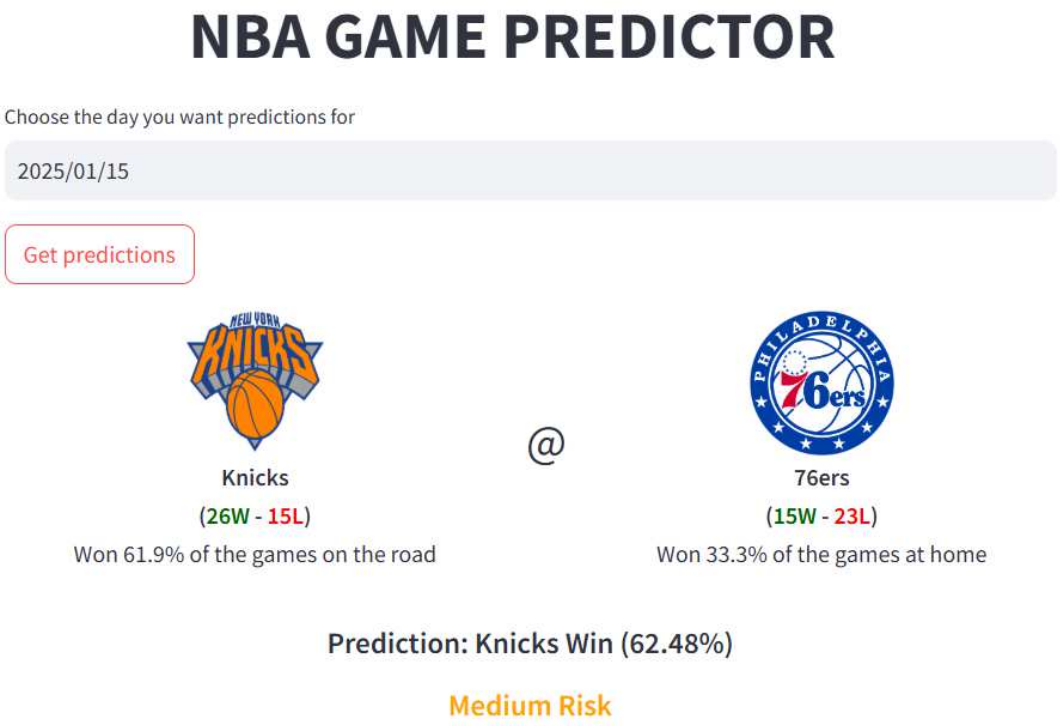


Figure 11: Forecasts presented on the website.

Given the academic nature of the developed system, the application does not contain usage restrictions, nor is any authentication or user data registration required.
The findings from this study have several potential applications. One key outcome is the use of the predictive model to assist stakeholders in gaining insights into game outcomes. For instance, coaches and clubs can leverage these predictions to optimize game strategies, identify potential areas for improvement, and better prepare for upcoming matches. Additionally, the model can serve as a decision-support tool for sports analysts and

commentators, enhancing the depth of pre-game and post-game analyses.
In the context of betting, which was briefly mentioned in the introduction, the model's predictions could provide valuable insights for betting organizations and participants. To further support informed decision-making, the web page presents the probabilities calculated by the logistic regression model, alongside the model's predictions. This additional information allows users to better assess the confidence level of the predictions and make more informed decisions based on the provided data.
While the web page does not explicitly highlight the risks associated with betting or provide detailed disclaimers, it is important to acknowledge the inherent limitations of the outcome prediction model. The predictions are influenced by the complexity of NBA games and numerous unpredictable factors, and users should recognize that predictions cannot guarantee profits. Future iterations of the web page could incorporate more comprehensive information to encourage responsible use and ensure ethical considerations are addressed. It is important to emphasize that the primary focus of this study is not betting but rather exploring predictive modeling as a tool for understanding and analyzing sports performance.
Overall, this study lays the groundwork for tools that can enhance decision-making across various stakeholders, from teams to fans, while also providing a foundation for further advancements in sports analytics.

Results Discussion

The chosen model showed moderate performance, with an accuracy rate of 64.1% and an F1 of 72.4%. These metrics suggest that although the model correctly predicts the majority of the outcomes, there is still room for improvement, especially in its ability to predict both possible outcomes, i.e., both the home team's victory and defeat. The high recall of 83.1% indicates that the model correctly identifies 83% of home team victories. However, the precision of 64.1% reveals that there is a significant number of false positives (36%), which compromises the reliability of the predictions. The F1 measure of 72.4% reflects the harmonic mean of these two metrics, which, while reasonable, underscores the need for a more robust balance between these metrics.
The comparison of the results reveals that, although the developed model does not achieve the highest accuracy rates among the analyzed models, it presents a more robust and comprehensive approach.

Table 5: Characteristics of the analyzed studies and the present study

Study	Seasons	Results
Horvat, Job et al. 2023	2013 – 2018 (2500 games)	Average Accuracy Rate: 66% Maximum Accuracy Rate: 78%
Ozkan 2020	2015/16 (240 games)	Accuracy Rate: 79.2% Sensitivity: 72.7% Specificity: 79.1%
Zhao, Du, and G. Tan 2023	2012 – 2018 (2460 games)	Average Accuracy Rate: 71.54% Maximum Accuracy Rate: 73.78%
Cheng et al. 2016	2007 – 2015 (10271 games)	Accuracy Rate: 74.4%
Horvat, Hava, and Srpak 2020	2009 – 2018 (11578 games)	Average Accuracy Rate: 60.01% Maximum Accuracy Rate: 60.82%
Zheng 2022	2012 – 2021 (10197 games)	Accuracy Rate: 67.98%
Best model of this study	2005 – 2022 (26622 games)	Accuracy rate: 64.4% Precision: 64.5% Recall: 82.5%

Study	Seasons	Results
		F1: 72.4%

Most of the models analyzed focus only on a single performance metric, such as the accuracy rate, which can provide a limited view of actual performance. In contrast, the developed model evaluates multiple performance metrics, allowing for a more comprehensive analysis of its capability. Although the first three studies in Table 5 report higher accuracy rates, their datasets are much smaller than the dataset used in this study, suggesting that the models' high performance may be due to overfitting. Among the remaining three studies, which used larger datasets (though only half the size of ours), the Zheng (2022) study stands out for its significantly higher accuracy. The main difference in Zheng's dataset is the inclusion of more domain-specific features, such as team ELO rankings, which likely made the model more specialized but potentially less generalizable for other basketball leagues. Moreover, while Zheng's study reports high accuracy, it does not comprehensively discuss metrics such as recall, precision, or F1-score. These metrics are critical for evaluating the trade-off between false positives and false negatives. Thus, the higher accuracy reported by Zheng may come at the expense of these other metrics, potentially limiting the broader applicability of the model. In summary, while the model may not achieve the highest levels of accuracy in comparison to the examined models, its strong methodology and thorough assessment of several performance indicators guarantee a more complete and reliable analysis, which ensures a higher capacity for generalization and applicability to various game scenarios.

Conclusion

This study was undertaken to evaluate the effects of several features on the performance of a model to predict the victory or defeat of the home team in an NBA game.

The study tested several single and ensemble machine learning models using a methodology to establish the effects of several statistical and historical features on the prediction results. The model achieved an accuracy rate of 64.1% and F1 of 72.4%, which indicates a fragile performance in percentage terms but robust when considering the complexity and competitiveness of the games in the NBA league.

The study also addressed four complementary questions of great relevance for the analysis of the NBA: how the dynamics of the league have changed over time, the impact of team changes on player performance, the evolution of positions in the game, and the relationship between age and athlete performance. Regarding the dynamics of the NBA, the analyses indicated a significant evolution in both the offensive and defensive capabilities of the teams throughout the period under study. Regarding the impact of player trades on the team, it was found that changes during the season affect the number of games played more than other performance metrics. As for the evolution of positions, they reflect a trend towards a more dynamic and flexible style of play, with athletes capable of performing multiple roles on the field. Regarding the evolution of player performance with age, players' performance typically peaks between 28-30 when technical skills and physical capacity combine. After this, a gradual decline occurs, notably more pronounced from age 40, necessitating adaptation of roles within teams due to natural aging changes. These analyses contribute to a deeper understanding of the factors that affect performance in the NBA, both at the team and individual levels.

This study highlights the robustness of the predictive model across different time periods; however, it is important to recognize certain limitations. Minor changes in rules and competition formats during the analyzed period—such as adjustments to timeout durations and restrictions on defensive plays—may have influenced gameplay patterns and outcomes. The current model didn't directly take these changes into account, but this could be fixed in

the future by either engineering features to capture their effects or retraining models on new datasets that show how competition is changing.

Additionally, future investigations could explore incorporating qualitative factors, such as emotional aspects and team ELO rankings, which may significantly influence the performance of teams and players. Leveraging advanced techniques, such as neural networks, could further enhance predictive performance and provide deeper insights into the complex interactions underlying game outcomes. Such advancements would pave the way for more comprehensive and adaptive models capable of addressing the ever-changing nature of sports competitions.

References

Basketball-Reference.com. (n.d.). *Basketball Stats and History*. Retrieved May 16, 2024, from https://www.basketball-reference.com/

Bishop, E. (2023). The role of coaching in the NBA: Analyzing the impact of head coaches on team performance. *Sportskeeda*. Retrieved Jun 3, 2024, from https://www.sportskeeda.com/basketball/the-role-coaching-nba-analyzing-impact-head-coaches-team-performance

Cai, W., Yu, D., Wu, Z., Du, X., & Zhou, T. (2019). *A hybrid ensemble learning framework for basketball outcomes prediction*. Statistical Mechanics and its Applications, 528, 121461. https://doi.org/10.1016/j.physa.2019.121461

Chen, W. J., Jhou, M. J., Lee, T. S., & Lu, C. J. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *Entropy*, *23*(4), 477. https://doi.org/10.3390/e23040477

Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. *Entropy*, *18*(12), 450. https://doi.org/10.3390/e18120450

Christos, K., Dimitrios, L., Christos, G., Georgios, K., & Nikolaos, S. (2020). Effect of offensive rebound on the game outcome during the 2019 Basketball World Cup. *Journal of Physical Education & Sport*, *20*(6). https://doi.org/10.7752/jpes.2020.06492

Lusa. (2023). Apostas desportivas online movimentam 4.000 milhões de euros em cinco anos. *[Online sports betting moves 4 billion euros in five years]. Diário de Notícias*. Retrieved August 27, 2023, from from https://www.dn.pt/arquivo/diario-de-noticias/apostas-desportivas-online-movimentam-4000-milhoes-de-euros-em-cinco-anos-15836237.html

Forebet.com. (n.d.). Retrieved August 29, 2024, from https://www.forebet.com/

CoachAD. (n.d.). *Unleashing the power of a proven 3-point shot attack*. Retrieved Jun 4, 2024, from https://coachad.com/play/unleashing-the-power-of-a-proven-3-point-shot-attack/

Hoop Social. (2023, August 8). The role of coaching and its impact on team success in the NBA. *Hoop Social*. Retrieved June 3, 2024, from https://hoop-social.com/the-role-of-coaching-and-its-impact-on-team-success-in-the-nba/

Horvat, T., Havaš, L., & Srpak, D. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *Symmetry*, *12*(3), 431. https://doi.org/10.3390/sym12030431

Horvat, T., & Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1380. https://doi.org/10.1002/widm.1380

Horvat, T., Job, J., Logozar, R., & Livada, Č. (2023). A data-driven machine learning algorithm for predicting the outcomes of NBA games. *Symmetry*, *15*(4), 798. https://doi.org/10.3390/sym15040798

Lauga, N. (2022). *NBA games* [Data set]. Kaggle. Retrieved May 15, 2024, from https://www.kaggle.com/datasets/nathanlauga/nba-games

Goodman, N. (2014). The hidden value of the NBA steal. *FiveThirtyEight*. Retrieved Jun 3, 2024, from https://fivethirtyeight.com/features/the-hidden-value-of-the-nba-steal/

NBA.com. (n.d.). *National Basketball Association*. Retrieved December. 21, 2023, from https://www.nba.com/

AS. (2023). Do NBA stars play in pre-season games? *AS*. Retrieved May 16, 2023, from https://en.as.com/nba/do-nba-stars-play-in-pre-season-games-n/

SportyTrader. (n.d.). *Sports Betting Tips*. Retrieved August 29, 2024, from https://www.sportytrader.com/en/

Streamlit. (n.d.). *Streamlit: The fastest way to build and share data apps*. Retrieved July 25, 2024, from https://streamlit.io/

Vitibet.com. (n.d.). *Betting tips Vitibet.* Retrieved August 29, 2024, from https://www.vitibet.com/

Zhao, K., Du, C., & Tan, G. (2023). Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm. *Entropy*, *25*(5), 765. https://doi.org/10.3390/e25050765

Zheng, X. (2022). NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors. *Proceedings of the 4th International Conference on Big Data Engineering*, 71–80. https://doi.org/10.1145/3538950.353896