International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00523-y

REGULAR PAPER

An innovative method for accurate NBA player performance forecasting and line-up optimization in daily fantasy sports

George Papageorgiou[1] · Vangelis Sarlis[1] · Christos Tjortjis[1]

Received: 11 July 2023 / Accepted: 16 February 2024

Abstract

This study presents a novel approach for predicting NBA players' performance in Fantasy Points (FP) by developing individualized models for 203 players, using advanced basketball metrics from season 2011–2012 up to season 2020–2021 from reliable sources. A two-step evaluation and validation process secured validity, while applying linear optimization methodology, considering constraints such as salary and player position to recommend an eight-player line-up for Daily Fantasy Sports (DFS). Four scenarios with 14 machine learning models and meta-models with a blending approach with an ensembling methodology were evaluated. Using individual per-player modeling, standard and advanced features, and different timespans resulted in accurate, well-established, and well-generalized predictions. Standard features improved MAPE results by 1.7–1.9% in the evaluation and 0.2–2.1% in the validation set. Additionally, two model selection cases were developed, with average scoring MAPEs of 28.90% and 29.50% and MAEs of 7.33 and 7.74 for validation sets. The most effective models included Voting Meta-Model, Random Forest, Bayesian Ridge, AdaBoost, and Elastic Net. The research demonstrated practical application using predictions in a real-life DFS case evaluated in a DFS tournament on a specific match day. Among 11,764 real users, our Daily Line-up Optimizer ranked in the top 18.4%, and profitable line-ups reached the top 23.5%. This unique approach proves the proposed methodology's effectiveness and emphasizes its profitability, as the optimizer process delivers positive results.

Keywords Data mining (DM) · Sports analytics (SA) · Machine learning (ML) · Daily fantasy sports (DFS)

1 Introduction

Sports analytics (SA) is a vast emerging field with significant domain value for teams, players, and organizations [1, 2]. In recent years, it has been verified that analytics and performance prediction in the sports domain facilitates the evolution of any sport, team, and players [3]. The biggest clubs have departments focused on their, as well as opponent, team analytics, trying to optimize their playstyle and detect valuable information that staff, players, and coaches cannot see [4]. For this reason, data have immense value for teams, and via different methods (cameras and sensors), collect as much data as possible for evaluation and forecasting purposes [5]. In basketball, many statistics and formulas refer to a player's overall performance in a match. Some important basketball metrics are efficiency (EFF), player impact estimate (PIE), player efficiency rating (PER), and usage rate [6].

Moreover, the daily performance of a player can be translated into Fantasy Points (FP), another metric highly correlated with these mentioned above. FP is a metric that betting companies use for player performance ranking. Basketball is a sport that entertains and engages people in many ways, from watching it and supporting their favorite team up to betting on it, with plenty of choices available (Win, points) [7]. In recent years, a new trend for DFS sports came up. People can become coaches and build their teams. Betting companies offer tournaments with many modes, awarding the users based on the total FP, their built team has achieved.

✉ Christos Tjortjis
c.tjortjis@ihu.edu.gr

George Papageorgiou
gpapageorgiou2@ihu.edu.gr

Vangelis Sarlis
e.sarlis@ihu.edu.gr

[1] School of Science and Technology, International Hellenic University, 14th km Thessaloniki, 570 01 Thermi, Moudania, Greece

Published online: 19 March 2024

Predicting daily player performance is a complex task, as numerous factors can influence a player's performance on any given day because of the complexity of its sport and exogenous parameters that affect each individual performance. This research focuses on basketball, specifically the NBA, with the objective of generating accurate daily player performance predictions, while excluding rookies and certain underperforming players from the 2020 to 2021 season.

In this study, 14 machine learning (ML) models were employed to forecast NBA player daily performance in FP, utilizing historical advanced and standard player and team data. By carefully collecting, cleaning, pre-processing, and conducting feature engineering on the data, highly accurate ML models were generated and compared, ultimately selecting the best model for predicting the FP of each eligible player in a game or event.

Eligible players are those who meet specific criteria, such as having over 100 appearances from the 2017–2018 to 2019–2020 seasons, more than 30 appearances in the 2020–2021 season, and an average participation time of over 18 min in the 2020–2021 season. This selection process results in a pool of players that individual modeling is effective and excludes rookies and those with long-term injuries [8].

To thoroughly evaluate the performance of the 14 ML models, this research conducts four distinct scenarios, each designed to assess the effectiveness of the techniques in predicting player performance. These scenarios involve various data configurations, such as comparing the results obtained from the last three seasons (LTS) and ten seasons (TS) periods, as well as analyzing the differences in prediction outcomes when using standard features versus a combination of standard and advanced features. This comprehensive approach ensures a robust evaluation of the models and helps to identify the most suitable techniques for daily Basketball Player Performance Forecasting (BPPF).

Following the development of accurate ML models for predicting player performance, this study also incorporates a linear optimization process to recommend an optimal eight-player team for a match day, considering constraints such as salary and player position. The primary goal of this process is to maximize the total FP of the selected team while adhering to the restrictions imposed by the DFS platform.

To demonstrate the practical application of the proposed methodology, the research examines a real-life DFS case using the generated predictions. This real-world example not only showcases the effectiveness and potential profitability of the chosen ML models and linear optimization approach, but also emphasizes the value of this research for professionals seeking to enhance their DFS strategies and decision-making processes in the realm of basketball and SA.

This paper presents several contributions to ML and sports analytics. It introduces an innovative approach by developing individualized ML models for 804 high-performance NBA players in four different forecasting scenarios. It uses data from the last three seasons and last ten seasons, as well as standard and advanced basketball metrics. This personalized modeling begins with generic models and offers a tailored approach on player performance. Additionally, the research employs linear optimization techniques to optimize the selection of player line-ups for DFS, based on forecasted performance. Hence, it incorporates real-world constraints, such as player fantasy salaries and positions, emphasizing the practicality of the predicted ML models.

Furthermore, the paper provides a detailed comparative analysis, which tests 14 different ML models and meta-models across the four forecasting scenarios. This analysis provides valuable insights into which models and techniques are most effective for each scenario. Moreover, the study goes beyond theoretical testing, by using its models in a real-world DFS competition, demonstrating remarkable performance, ranking in the top 18.4% among 11,764 participants and providing profitable results. Lastly, the paper addresses key research questions regarding model accuracy, generalizability, and real-world applicability, offering a comprehensive view of its methodology and outcomes. Overall, this research advances the field theoretically and demonstrates real-world practical application, particularly in the DFS context.

2 Literature review

SA is an emerging field, and all big sports organizations and professional teams use it to develop the team, improve results, and identify problems that are hard to be spotted by human abilities [9, 10]. Technology improvements have created new playstyles, strategies, and tactics over the years. In addition, results evaluation with the aid of analytics is important for every sport. Nowadays, coach experience is not enough to be competitive at the highest professional level [11]. However, decades ago, data were handwritten and hard to observe, while data collection was manual and time consuming. For this reason, there are limited statistical records for several sports. SA appeared in the nineteenth century, along with the idea of evaluating each player's skills by analyzing their play [3].

2.1 Basketball players' performance prediction overview

Predictions for basketball player performance using DM and ML algorithms are a new research subject. Decades ago, evaluation and predictions for player performance were only based on coach experience. However, in recent years, SA and players performance prediction are becoming significant research field [12, 13].

A study in [14] used DM methods and tried to predict NBA player performance by working with data from seasons 2005–2006 up to 2013–2014. Firstly, the authors clustered players using K-means on their historical performance and proven skills, aiming at detecting changes in performance-based clusters to predict their next game's performance. The problem was transformed into a classification task, using Naive Bayes with clusters as labels based on historical performances. They tested three players' performance predictions using both methods and compared results. Two out of three players were classified into the same labeled cluster that was assigned in the clustering experiment. Lastly, they used a multiple regression model and exponential smoothing based on athletes' historical statistics to predict their performance. Results have shown that the exponential smoothing algorithm performs better.

Another study in [15] designed a unique network based on NBA data from all line-ups and matchups of teams from season 2007 up to 2019. Using ML and graph theory, the authors created a metric called Inverse Square Metric and an edge-centric multi-view network aiming to predict the performance of an NBA line-up anytime. Specifically, the edge-centric approach provides a thorough examination of any situation of the teams from 16 perspectives, working with data like defensive or offensive rebounds and many other features. According to their findings, they constructed a highly accurate system with an edge-centric multi-view method with 80% average accuracy. Their results were improved by 10% compared to the baseline methods, illuminating how efficient graph theory is for line-up performance prediction.

The study in [16] tried to predict points scored and winning scores, using mixed models with random effects. Also, the authors tried to find out which feature-metric was essential to make these predictions. In their study, they considered all the possible variables that may affect player performance. As a result, they created a dataset of 2187 examples, focusing on 27 NBA players in the 2007 regular NBA season. Results show that variables such as the player, position, difference in team quality, if the player started the match, the minutes he played, and his usage rate was crucial to predict the points scored successfully. In addition, crucial variables to predict the winning score were the player, his age, position in the field, difference in team quality, relationship between his age and his position, the minutes that he played, and the usage percentage. Lastly, they made their predictions using a single model with all the data instead of creating daily models.

The researchers in [17] successfully predicted the NBA MVP for 2017–2018, 2018–2019, and 2019–2020 seasons. In addition, they predicted the Best Defender for the following NBA season 2017–2018, 2018–2019, and 2019–2020. These forecasting scenarios were performed based on certified data from seasons 2017 up to 2020. Every season of the dataset comprised 82 games split into four groups (Q1–Q4). They selected 20 NBA players who participated in at least 30 games per season and at least 15 min average participation time per game. They created two advanced basketball analytics formulas, aggregated performance indicator (API) and defensive performance indicator (DPI). API is a composition of box score statistics and important rating basketball analytics, a synthesis of variables that illustrate the athlete's general performance and was adopted to predict the season MVP. DPI is a combination of advanced analytics variables focused on player contribution to Defense and was used for forecasting the Best Defender of the year. They accurately predicted the NBA MVP for seasons 2017 up to 2020 and the Best Defenders of the year.

2.2 Fantasy points and daily fantasy line-ups

Over the past 15 years, a new method for fans participating in basketball has become very popular worldwide. Companies offer the chance to users to take the role of Team Manager or Coach and create their Fantasy Basketball line-up. Fantasy is a vast sector in the betting industry, with millions of users trying to predict the best in terms of performance daily basketball line-up [18]. Basketball Fantasy is highly competitive, while users compete against each other, and the best line-up predictions are rewarded. Basketball is a sport filled with analytics, yet professionals and amateurs make predictions using raw statistics or advanced analytics and ML, building models, and making up strategies [19].

The study in [20] examines the use of self-exclusion features in Responsible Gambling (RG) tools on DraftKings, a major Daily Fantasy Sports (DFS) platform, by analyzing over 3 years of player data. The researchers found that less than 0.5% of the users opted for self-exclusion, with nearly a third doing so multiple times. Also, they found that self-excluders, compared to those who did not self-exclude, generally engaged in a broader range of contests and sports and opted for higher entry fees. Interestingly, repeat self-excluders initially chose shorter self-exclusion periods and participated in a greater variety of games. The study concludes that these findings offer insights into how RG tools are used in DFS and can help identify risk markers for problem gambling. Compared to our study, this research focuses on using RG and user interactions, which introduces a methodology of predicting an optimal line-up that ranked in the top 18.4% on a real-world DFS event.

The study in [21] proposes new algorithms and models by utilizing mixed-integer programming (MIP) and various time-series prediction techniques to optimize team formation in DFS in the NFL. The researchers compare machine-based approaches against human heuristics and existing benchmarks, testing on data across four NFL seasons, concluding that they outperformed existing methods by returning a profit of 81.3% of DFS game weeks. According to this study, the quality of the optimization model has a more significant impact on cash wins than the capacity to model uncertainty and forecast outcomes. Their study poses challenges and opportunities for the DFS industry by establishing a new performance baseline. Unlike this NFL-focused research, which relies on mixed-integer programming, our study advances DFS models for the NBA using a decade-long dataset and 14 types of ML models.

In the study presented in [22], two mathematical programming models were developed to act as "virtual coaches" for Argentinian fantasy soccer game participants. The first model created "a posteriori," optimizing team line-ups based on known outcomes. In contrast, the second "a priori" model generated real-time decisions based on existing data to improve team performance. The "a priori" model was tested in real competitions and consistently ranked among the top 4% of participants, reaching the top 0.1% in one tournament. The study highlights the potential of mathematical programming in sports decision-making and suggests that such models could also be adapted for real-world sports coaching. This study uses mathematical programming in Argentinian soccer, compared with our study, which employs a robust set of ML models in the NBA context, validated by real-world tournament rankings.

In [23], researchers analyzed data from 5000 participants in an online fantasy soccer game in the English Premier League (EPL) during the 2007/08 season. They inferred that the predictions based on fantasy sports data outperformed random forecasts and those based on team attendance, but those outperformed by the bookmakers. The study concludes that the collective decision-making of fantasy game players can produce a valuable dataset for predicting real-world soccer match outcomes. Compared to our study, which uses multi-season NBA metrics for individualized player models, this research focuses on collective decision-making in fantasy soccer based on a single-season dataset.

The study in [24] introduced a way to predict player FP and develop a system predicting the best combination of players in Daily Fantasy Line-ups, having as target the best overall score with a sure salary cap. They trained their models with data from season 2013–2014 and used their system in season 2015–2016, evaluating their predictions against actual results. They followed two methods; firstly, they used a Bayesian random effects model to predict Daily NBA player performance and generate a team baseline based on the game's rules having a specific salary cap and a constraint on the number of players who play in the same position. Secondly, they developed a k-nearest neighbors (KNN) model using the results from the previous Bayesian model to identify "successful" line-ups. Both methods successfully generated profit in a hypothetical experiment for the season 2015–2016, with KNN generating more profit than the Bayesian one on its own. Unlike this single-season, dual-method study in NBA DFS, our study uses a decade of data and 14 ML models, achieving top 18.4% real-world DFS performance.

The researchers in [25] investigate inconsistencies and potential biases in the recorded box score statistics in NBA. They used optical player tracking data from season 2015 to 2016 and developed a model for predicting recorder assists, with higher performance than previous methods. Findings reveal evidence of inconsistencies in the awarding of assists by team-hired scorekeepers. Also, they suggested potential biases related to "passers" or their positions. Afterward, they investigated their approach in the real-world implications, such as monetary consequences for Daily Fantasy Sports participants. Finally, they suggested that individual players could benefit from adopting a more proactive stance for monitoring the attribution of subjective box score statistics. Following the literature review, we will present our innovative methodology for predicting NBA player performance to enhance the effectiveness of DFS line-up selection. Employing state-of-the-art ML techniques and advanced basketball analytics, we are developing a robust workflow that addresses various aspects of the prediction process, from data collection to DFS team evaluation. This study explores biases in NBA statistics compared to ours which focuses on NBA player performance prediction through comprehensive metrics, validated by high real-world DFS rankings.

3 Methodology

The methodology of this study is grounded in a quasi-experimental design, where we focus on the empirical analysis and application of ML models for predicting NBA players' daily Fantasy Points performance. While a comprehensive review of existing literature supports our approach to contextualizing our research within the current state of sports analytics and Fantasy Sports, the core emphasis is on data-driven experimentation and the application of ML models. Beginning with presenting the research questions, aim, and objectives that guide the research, we continuously delve into the details of our methodology, which includes data collection, examination, data wrangling, feature engineering, and post-feature engineering pre-processing. Furthermore, we analyze the modeling phase, encompassing feature selection and data modeling, and present the two-step evaluation and validation processes that are selected, to assess the performance of our models. Finally, we introduce the DLO procedure, outlining its restrictions and implications. This section aims to provide a clear understanding of the methodological framework that helps achieve our research questions and objectives.

3.1 Research questions (RQs)

The research questions subsection addresses three key questions that fundamentally form this study. These questions aim to address the development of individual ML modeling for each NBA player to predict their daily FP performance accurately, the impact of using standard versus advanced features and data from different timespans, and the practical application of linear optimization on generated predictions to recommend an optimal eight-player team for a match day, considering constraints.

The study strives to answer these research questions, which are significant for organizations, teams, coaches, players, DFS fans, and DFS platform holders seeking to enhance their decision-making. Moreover, the potential applications of DM and ML techniques in DFS are underscored since this study aspires to provide a complete workflow for performance prediction, line-up creation, and optimization.

Hence, the research questions try to answer are the below:

Primary research question (RQ1) How can individual ML models for each NBA player be developed to accurately predict their daily FP performance?

This primary question centers on the development of unique ML models for NBA players, setting the stage for a detailed exploration of the data features in the subsequent question.

Secondary research question (RQ2) What is the impact of using standard versus advanced features, and data from different time periods on the accuracy of ML models in predicting player performance?

This question investigates how the selection of standard or advanced features and the incorporation of data from various timespans influence the accuracy of the ML models used in player performance predictions.

Tertiary research question (RQ3) How can linear optimization on generated predictions be effectively applied to recommend an optimal eight-player team for a match day, considering constraints?

The third question extends to the practical application of our models in Daily Fantasy Sports. It explores how linear optimization can be utilized to construct the most effective team compositions based on the model predictions within the specific constraints of DFS platforms.

The selected research questions hold significance for organizations, teams, coaches, and players, as they aim to improve decision-making and performance at various levels through accurate performance predictions. Furthermore, these questions highlight the potential applications of DM and ML techniques in Daily Fantasy Sports (DFS), specifically in the realms of performance prediction and line-up creation and optimization. This emphasizes the growing relevance of advanced analytics in both player evaluation and strategic planning within the basketball industry.

3.2 Aims and objectives

This study aims to develop a novel comprehensive approach for predicting daily NBA player performance in FP using individualized models for each player's case, incorporating both standard and advanced features with different timespans. This approach is built to improve the accuracy of player performance predictions and provide more effective models for generating optimal line-ups for DFS, ultimately showcasing the practical application and profitability of the proposed methodology. FP is calculated by formula (1).

$$FP = P + 1.2 \times REB + 1.5 \times AST + 3 \times STL + 3 \times BLK - TOV. \tag{1}$$

where P = Each point scored, REB = Rebound, AST = Assist, STL = Steal, BLK = Block, and TOV = Turnover

By leveraging cutting-edge ML techniques and advanced analytics, this research aims to create a robust workflow that addresses various aspects of the prediction process, from data pre-processing to model evaluation. The following objectives have been presented above to provide a structural understanding of the research's objectives and guidance throughout the research.

The research focuses on developing ML models for each eligible NBA player to accurately predict their daily FP performance, considering historical advanced and standard player and team data. It also investigates the impact of using standard versus advanced features and data from different periods (such as 3-season and 10-season periods) on the accuracy of the ML models in predicting player performance.

Additionally, the study applies linear optimization to the generated predictions to effectively recommend an optimal eight-player team for a match day, considering salary and player position constraints. The performance of the proposed methodology is evaluated through a series of scenarios, comparing the effectiveness of various ML techniques and data configurations.

This research seeking to demonstrate its real-life application and potential profitability, uses the generated predictions to examine a real-life DFS case, contributing to growing basketball DFS and SA knowledge by showcasing the practical applications and benefits of employing advanced analytics in performance prediction and line-up optimization.

3.3 Data engineering

The NBA domain offers various types of responsive data accessible online, but the credibility of some sources is uncertain. This study primarily acquired data through the NBA API, accessing the official NBA website [26] for player and team data from the 2011–2012 season up to the 2020–2021 season. The data types accessed for individual players and teams included traditional, advanced, miscellaneous, scoring, usage, four factors, and opponent statistics [27]. These data types encompass raw performance statistics for each player and team and their final ranking scores. The latter were removed during pre-processing to avoid incorporating future information from the final season.

After scraping relevant data, the methodology involved data cleansing and joining player performance with their corresponding team. The TS dataset with all accessed statistics was split into four subsets for the scenarios: TS data with all available statistics, TS data with only standard statistics, LTS data with all available statistics, and LTS data with only standard statistics. A short overview of each corresponding feature type in the datasets is shown in Tables 5 and 6 in the appendix.

In continuation, ML applications, 1-game lag features, and momentum features were generated for all statistics. Additionally, 3-, 5-, 7-, and 10-game lag features were specifically created for FP. Anomaly detection was also applied to create extra features, and 'rest days' features were incorporated. Before modeling, an individual dataset was created for each player in each scenario. Multicollinearity was removed to avoid unstable model estimates, reduced interpretability, and overfitting. Low variance processes were also applied to eliminate statistics with limited predictive power, reduce dimensionality, enhance computational efficiency, and prevent overfitting.

In the modeling phase, a permutation feature importance technique was applied to provide robustness against multicollinearity, improve model performance, reduce overfitting, enhance interpretability, decrease dimensionality, and increase computational efficiency. The corresponding dataset was split, sorted from the oldest to newest, into 70% for training, 20% for testing, and 10% newest for validation as unseen data. The Pycaret library was used to compare model performance. For each scenario and player's case study, the top three models based on MAE scores in the training process were fine-tuned using randomized grid search. A meta-model, a voting estimator, was created using a blending approach within the ensemble method, based on the three best performers. The best performing model from the 10-fold cross-validation (CV) process, including the voting estimator, was finalized and further trained with the test dataset. In conclusion, each player's finalized model was evaluated and validated with unseen data using mean absolute percentage error (MAPE) [28] and mean absolute error (MAE) [29]. The overview of the process presented in Figure 1.

To demonstrate the real-life application of the research, linear optimization [24] was applied to the generated predictions from all eligible players' unseen data to create an optimal eight-player team for a match day, considering four types of constraints: unique player selection, game diversity, position requirements, and salary cap. Salary and position data were acquired from one of the most famous Fantasy organizations, Draft Kings. The proposed methodology was tested for a match day (May 15, 2021), yielding profitable results in an actual fantasy tournament compared to real individual users who participated in the tournament.

3.3.1 Data collection

Utilizing the NBA API from the official NBA website [30, 31], which offers a wealth of player and team statistics, data from the 2011–2012 season up to the 2020–2021 season were acquired. The relevant data included player and team performances per game, spanning both regular and playoff seasons. For individual players, traditional, advanced, miscellaneous, scoring, and usage data were collected for each player's game performance. Traditional, advanced, miscellaneous, opponent, scoring, and four factors data were obtained for teams [32].

The diverse data types gave us a comprehensive view of players' and teams' performance, enabling us to make predictions. A short overview of those:

Traditional statistics include basketball statistics for players and teams, such as field goals attempted and made, blocks, assists, rebounds, steals, turnovers, and others.

Advanced statistics encompass metrics beyond traditional statistical measures, such as shooting percentage, assists percentage, and assist-to-turnover ratio.

Miscellaneous Statistics contain various statistics for players and teams, for instance, points off turnovers, second-chance points, and points in the paint.

Scoring statistics provide scoring-related metrics for both players and teams, including the percentage of points from different court areas and assists and other related measures.

Usage statistics focus on player and team usage, capturing percentages between players and teams in various categories, such as field goals, rebounds, and assists.

Four factors statistics include metrics first proposed by Dean Oliver, aimed at determining the most critical factors in a game's outcome, such as effective field goal percentage, turnover percentage, offensive and defensive rebound percentages, and free throw attempt rate.

Opponent statistics include metrics that measure the performance of opposing teams.

3.3.2 Data wrangling

The collected data included performance metrics for every player who participated in at least one game from the 2011–2012 season up to the 2020–2021 season. However, the research focused on players active in the 2020–2021 season. Inactive players were excluded based on an active ID key generated from the latest season. For excluding rookies

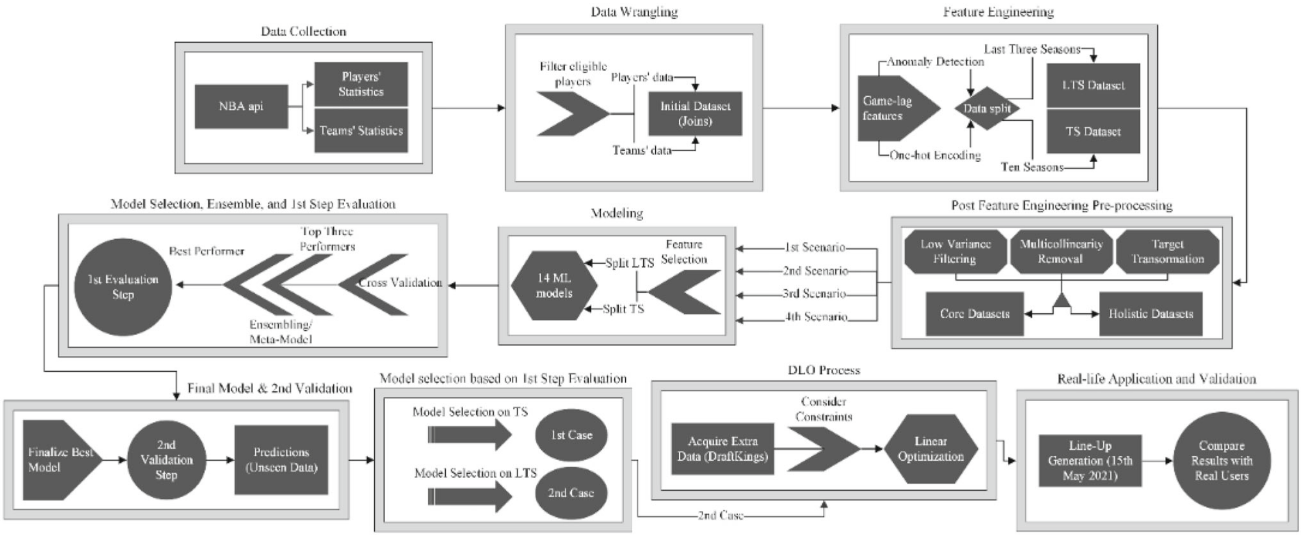


Fig. 1 Research workflow

and long-term injured players, specific criteria were applied. Eligible players in this study had at least 100 appearances between the 2017–2018 and 2019–2020 seasons, over 30 appearances in the 2020–2021 season, and an average of at least 18 min of playing time per game in the 2020–2021 season. Additionally, players scoring season 2020–2021 FP mean of less than half of the total FP mean for all players were dropped. These constraints were applied to ensure the effectiveness of individual modeling and to predict an optimal eight-player line-up for a match day. Users should select players more likely to participate in games and accumulate FP while avoiding benchwarmers and injured players. After applying these filters, the dataset included 203 out of 540 players.

Moreover, per-game outliers were removed from the dataset to ensure more reliable and consistent predictions and optimal line-up creation. Records were excluded where players underperformed with less than 10 FP due to injury, unexpected benching, or playing less than one period (12 min). By eliminating these outliers, the modeling process can better predict a player's potential performance, focusing on their distinctive capabilities without being influenced by rare occurrences. Additionally, statistics related to the final ranks of each player in each statistic were dropped to avoid data leakage. At this pre-processing stage, the initial players' dataset comprised 145,415 instances and 106 statistics, while the initial teams' dataset contained 25,148 instances and 100 statistics.

3.3.3 Feature engineering

This step applies dataset's feature engineering and post-feature engineering. First, the dataset is split for each player, and the "rest days" feature is created by calculating the differences between each player's respective games. Statistical anomaly detection [33] is applied to the FP using standard deviation, with a boundary set at − 2, generating two features: One called "anomaly" for marking where an anomaly in FP is detected, and a second one called "smoothed FP," which smooths the performed FP based on standard deviation.

One-hot encoding is applied to the categorical features, including home/away, matchups, team name, season type, double–double, and triple–double. Momentum features are created for all features, capturing the performance differences between each game and the previous one.

Subsequently, each player's dataset is split individually, and 1-game-lag features are generated for player and team data. Additionally, 3-, 5-, 7-, and 10-game-lag averages for player FP and 3-game-lag sums for past double–double and triple–double achievements are calculated. Lastly, momentum features were created with calculated player's performance last match's difference from the match before last.

Next, each player's opponent dataset is created based on their future matchup using the team's dataset. Finally, each player's dataset is joined with their corresponding team's performance and the next opponent's performance data.

Finally, the primary TS datasets for each player with all available historical statistics were created. Following the above methodology, data were transformed into historical for applying regression ML modeling, and there is insurance that any data leakage option is prevented [34]. For constructing our final databases, a subtraction of the primary datasets for each player is generated, keeping only the standard statistics and standard opponent data. The "complete" dataset, with advanced features dataset, finally contained 79,036 instances and 269 features, while the dataset will only standard data named "core" also contained 79,036 instances but only 67 features.

3.3.4 Post-feature engineering pre-processing

To finally refine the datasets, further pre-processing steps are employed after the competition of the feature engineering. Those are considered highly important to address issues related to the distribution of the target variable and the presence of irrelevant or redundant features. Because of the form of the reformatted datasets, features generated from the feature engineering process needed to be further adjusted based on each player's performance score characteristics.

Target variable transformation Each player's performance is also adjusted based on exogenous parameters for each game. For this reason, our target variable, FP, represents players' performance and follows a non-symmetrical distribution. This issue is faced with the Yeo–Johnson method. This method creates a more symmetric distribution, which is desirable for modeling [35].

Feature reduction Considering a large number of features and the specified data available per player (with up to 82 regular-season games plus playoffs), we apply the following methodology for feature reduction, while some features did not contribute significantly to the model's explained variance. The feature reduction process consisted of the two following steps:

1. *Multicollinearity removal* The features which exhibit high linear correlation with another feature and lower correlation with the target variable were dropped. Datasets with highly correlated features may increase coefficient estimates' variance, making them unstable and noisy [36]. A multicollinearity threshold of 0.50 was set, and features with inter-correlations higher than this threshold were removed.
2. *Low variance filtering* This method was applied for the categorical features [37], some of which were "PLAYOFFS," "OPPONENT," "SEASON_YEAR," and "TEAM_ABBREVIATION." Features with statistically insignificant variances were removed from the dataset. The above rules are justified for filtering since both conditions should be met.

2.1. Proportion of Unique Values: Count of unique values in a feature/sample size < 10%

2.2. Dominance of Most Common Value: Count of most common value/count of second most common value > 20 times.

3.4 DM and ML algorithmic models

Throughout this research, 14 traditional from different categories of ML models and meta-models with one model or meta-model per use-case, utilizing a blending approach within the ensembling method, with Pycaret [38], were challenged on 203 individual real-life players' use cases for investigating the optimal and most balanced forecasting methodology for producing accurate predictions.

A brief description of each model is described below:

Huber uses a different loss function from the traditional least-squares; it is less sensitive to outliers in data [39].

Ridge is a specialized technique for data that suffer from multicollinearity. The parameters are shrunk, preventing multicollinearity, and finally, the complexity of the model is reduced by coefficient shrinkage [40].

Linear regression is commonly used for predictive analysis. It refers to a linear approach for modeling between a scalar and explanatory variable [41].

Least angle is preferred for high-dimensional data. Finding the higher correlated features to the target value pushes the regression line in this directive until it contacts another variable with the identical or more increased correlation [42].

Bayesian is usually selected for insufficient or inadequately distributed data by formulating linear regression. It works by employing probability distributors instead of point estimates. The response output (y) is assumed to be computed by a probability distribution instead of estimated as a single value [43].

Orthogonal matching pursuit is used to recover a high-dimensional sparse signal from a small set of noisy linear measurements. It is an iterative greedy method that selects the most correlated feature at each stage [44].

Passive aggressive belongs to the category of online learning in ML. This technique works by feeding its instances sequentially, individually, or in groups called mini batches. It is mostly used in procedures where data stream in a continuous flow [45].

AdaBoost is a meta-estimator that works by matching a regressor on the original data, and in the next phase, copies of this regressor on the same dataset using modified weights of instances based on errors in the first prediction [46].

Random forest works by using ensemble methods (bagging). This technique starts by constructing many decision trees and delivers the mean/mode of prediction of these [47].

Gradient boosting implements a regression tree by fitting it on the negative gradient of the given loss function. This method enables the optimization of any differentiable loss function [48].

Extra trees uses a meta estimator that fits several different decision trees on different sum samples of the dataset and improves accuracy by averaging [49].

Lasso is a regularization technique and a linear regression method that uses shrinkage. Usually, this method is preferred when a high level of multicollinearity is present. When there are many features, it automatically performs feature selection [50].

Light gradient boosting machine is an extension of the gradient boosting method. It follows an automated feature selection procedure and boosts examples with more considerable gradients [51].

Decision tree breaks down the dataset into smaller subsets samples. In this way, a decision tree is incrementally produced. The tree is constructed of internal and leaf nodes in its final form [52].

3.4.1 Data modeling scenarios

In forecasting procedures, aiming to achieve the highest possible accuracy and stability in predicting each player's performance, this study utilizes complete and core datasets. These initial datasets were divided into per-player datasets, allowing for four distinct scenarios—two using the most recent records (data from the last three seasons) and two using all available data from the past TS. Since we do not use multivariate time-series forecasting, we want to ensure that our models are well-generalized, stable and provide valid predictions. Therefore, we forecasted two separate batches of the latest continuous performance with unseen data (30%).

For each scenario, the data were split into 70% for training, 20% for testing (sorted by date), and 10% as unseen data for model evaluation. The standard deviation of the LTS datasets was 9.47 for the train/test set and 8.93 for the unseen set, and for TS datasets was 9.73 for the train/test set and 9.07 for the unseen set. Subsequently, our results need to be compared based on our evaluation metrics to ensure the validity of our models.

In order to ensure the validity of the models, a 10-fold CV technique was employed during the training process for each model. It is important to note that our training and forecasting methodology adopts a non-temporal focus; expressly, game dates were not incorporated into the forecasting scenarios, as our objective was not to make time-sensitive predictions. Despite that, we used temporal trend features, such as game-lag features, to structure well-generalized models based on past performance, even without explicitly counting on time order subsequential data. Considering the imbalanced nature of the datasets, MAPE was selected as the most appropriate evaluation metric. Based on the performance during CV, the top three models were identified for further optimization. Furthermore, our models undergo a two-step evaluation process, validated on unseen data, except for CV, to ensure generalizability and confirm no overfitting.

The feature selection methodology was based on each feature's importance [53]. It was targeted to constrain the feature space using a mix of permutation importance approaches, Random Forest, AdaBoost, and linear correlation with the target variable to improve modeling efficiency. The adjusted rule on the feature selection threshold was set to 0.90, meaning that only features accounting for at least 90% of the dataset's variance were retained, thereby ensuring the preservation of the most influential features for modeling.

These top performing models were fine-tuned using a randomized grid search [54] to identify the optimal hyperparameters for each model. Subsequently, a blending ensembling method was employed to design a meta-model known as the "voting estimator," based on the three fine-tuned models.

Afterwards, the examination of four distinct scenarios with different combinations of data and timespans follows:

Scenario 1 Models were trained on complete datasets using TS data. The performance of the models varied for each player.

Scenario 2 Models were also trained in complete datasets with a subset of historical data, using only the LTS or more if requirements were not fulfilled.

Scenario 3 Models were executed with core datasets for TS historical data.

*Scenario 4*The last scenario used only the LTS and core datasets.

3.4.2 Two-step evaluation and validation process

This study employed a robust two-step evaluation and validation methodology, ensuring the accuracy, generalization, and reliability of the ML models used. Starting with the CV [55] initial validation, wherein the train (70%) data were split into multiple folds, in which each fold acted as a validation set. In contrast, the model was trained with the remaining data; this technique allowed for an objective assessment of each algorithm's performance while inhibiting the risk of overfitting. After the CV, the models were evaluated with the corresponding test dataset (20%), separated from the training data. In this way, we ensure the model's ability to generalize and make accurate predictions on new, previously unseen data. Finally, each finalized model was further validated using an unseen dataset (10%), which assessed an additional layer of performance evaluation. This comprehensive two-step evaluation and validation procedure ensured that the selected finalized models were accurate and capable of facing our problem effectively with minor deviations between the two steps.

Moreover, using MAE and MAPE as performance metrics offered distinct advantages in reviewing the models' performance across players and scenarios. Comparing MAE values of different models, we can identify the models that constantly result in small errors and provide us with the ability to select the best performing model for each player based on a common, interpretable metric. Moreover, our comparative study is based on MAPE since it allows us to compare the models' performance across players and scenarios with different magnitudes of player performance values. To conclude, this comprehensive approach guarantees that the selected finalized models would predict valid results as accurately as possible and consistently perform across a range of scenarios.

3.4.3 Daily line-up optimizer (DLO)

The Daily Line-up Optimizer (DLO) is crucial in this research, comprising the final stage. It maximizes the predicted FP for a specific date by the initial modeling processes while adhering to the restrictions followed by betting companies. As its final step, this research aims to provide and benefit all stakeholders of the DFS basketball industry with a profitable end-to-end solution.

The DLO process involves linear optimization techniques to build the best FP-performing line-up for a specific match day. To achieve this, we utilized the Pulp library [56], acquiring extra data related to player salary and position from the [57] website. The DLO process is applied to one of the latest days of the 2021 regular season (May 15, 2021), featuring 26 different NBA teams and 13 events (games).

To produce the optimal lineup, we limited the selection to 53 players who participated in games on the specified date, chosen from our initial pool of 203 candidates. It is essential to note that this pool was relatively small, as this research focused on predicting the performance of the players, who are more likely to participate in an event and perform well, based on each corresponding historical performance.

3.4.4 DLO restrictions

Since the Fantasy Tournaments are designed by companies in the betting industry, like DraftKings, each applies its rules and restrictions. This research's final step is based on DraftKings company tournaments. It aims to build an eight-player line-up with a combination of players by maximizing the potential FP the players will score in their actual game. It uses the generated predictions of each player in our pool who participated in the games.

A short description of the restrictions DraftKings applies, and this research considered follows:

- Buy a player at most once.
- Include players from at least two different NBA games.
- The eight roster positions are:
  - One PG (Point Guard)
  - One SG (Shooting Guard)
  - One SF (Small Forward)
  - One PF (Power Forward)
  - One C (Center)
  - One G (PG, SG)
  - One F (SF, PF)
  - One Util (PG, SG, SF, PF, C)
- Spend no more than $60,000.

4 Findings

In this section, we present the results of this comprehensive research study targeted to identify the optimal ML models for forecasting NBA players' FP. Those results include performance comparison across different scenarios with complete and core datasets and different timespan data TS and LTS. Moreover, we investigate the effectiveness of result optimization that our data modeling methodology provides us with the ability to do by selecting the best performing models for each player across all scenarios. Furthermore, we overview the DLO procedure's outcome, demonstrating a profitable practical implication of our predictive models for selecting the optimal line-up in DFS. The evaluation metrics, as mentioned before, are the MAPE and MAE, computed as averages for all the 203 high-performance player models.

4.1 Performance scenarios comparison

In this subsection, we analyze the performance of the ML models in different scenarios, comparing the impact of using complete and core datasets during different timespans, TS and LTS. This overview aims to determine the best methodological path of features and timespans for valid and most accurate predictions of players' FP in their actual games. The two-step evaluation and validation methodology are followed with data that each corresponding model has not seen before to ensure that models produce well-founded results.

The results of each scenario, based on Table 1, where the bold indicates the lowest value for each evaluation metric, were the following:

In Scenario 1, in which the models were trained with the complete dataset using TS, the average MAPE for the validation set was 30.60% and for the unseen set was 30.70%. The average MAE for the validation set was 7.5, while for the unseen set, it was 8.03.

In Scenario 2, where a complete dataset but LTS is used, the average MAPE for the validation set was 31.60%, and for the unseen set was 31.10%. Moreover, correspondingly, the MAE for validation and unseen set was 7.50 and 8.03.

In Scenario 3, where the core and TS dataset was used on our models, the average MAPE for the validation set was 28.90%, and for the unseen set, it was 28.60%. As follows, for MAE, in the validation set, the score was 7.06, and for the unseen set, it was 7.54.

Lastly, in Scenario 4, where models were trained on the core dataset using LTS data, the average score for MAPE

Table 1 Average model performance metrics across scenarios for validation and unseen data (203 players)

Scenario	MAPE		MAE	
	Validation	Unseen	Validation	Unseen
Scenario 1 (complete, TS)	30.60%	30.70%	7.50	8.03
Scenario 2 (complete, LTS)	31.60%	31.10%	7.98	8.15
Scenario 3 (core, TS)	**28.90%**	**28.60%**	**7.06**	**7.54**
Scenario 4 (core, LTS)	29.70%	30.90%	7.47	8.02

on the validation set was 29.70% and, for the unseen set, 30.90%. Furthermore, the average MAE for the validation set was 7.47, while for the unseen set, it was 8.02.

4.2 Results optimization

In the results optimization section, we go deeper into the performance enhancements, contrasting the results scenarios using either complete or core datasets. The main objective of this section is to follow a model's best performing selection-oriented methodology, ensuring that both datasets have the same number of records and that models have been trained, tested, evaluated, and validated with the same historical data (TS and LTS).

Based on our goals in the first and second research questions, we justified that predictions were accurate in both forecasting final scenarios using the most recent data (LTS) and the whole dataset (TS). Moreover, the differences between our LTS and TS scenarios were not extreme, meaning that our models, even in TS, succeeded in capturing the recent trend based on the game-lag features. For that reason, we created and used a well-generalized models and ensuring good data quality of the constructed dataset in continuation of the data wrangling and feature engineering processes. The results mentioned are presented in Case 1, which involves model selection from scenarios 1 and 3, and Case 2, which involves model selection from scenarios 2 and 4.

The model selection is made by comparing MAPE and MAE scores on test sets and evaluating the performance of this methodology by unseen sets of MAPE and MAE results, ensuring that the proposed method is valid and validated in unseen tests. The generated predictions are used for the DLO procedure and optimal line-up creation, so the final results are unbiased [58]. Concluding, by this framework, we determine the most accurate long-term predictions, and in the end stage, the best combination of models for predicting NBA players' FP is designated. The results are presented in Table 2.

For Case 1, in which TS data were used, the average validation MAPE was 28.30%, and the MAPE in the unseen set was 28.70%, while the average MAE in the validation set was 6.98, and the unseen set's MAE score was 7.54.

Table 2 Model performance comparison for TS and LTS

Case	MAPE		MAE	
	Validation	Unseen	Validation	Unseen
Case 1 (TS)	28.30%	28.70%	6.98	7.54
Case 2 (LTS)	28.90%	29.50%	7.33	7.74

For Case 2, with LTS data, the average MAPE for the validation set was 28.90% and for the unseen set was 29.50%. Additionally, the MAE for validation and unseen set was 7.33 and 7.74, respectively.

All models underwent rigorous assessment for stability and generalization to guarantee the reported average evaluation and validation results accurately represent model performance, ensuring that the models remained unbiased and not prone to overfitting.

4.3 DLO results

In the DLO results section, we will present the outcome of the actual application of this research by applying DLO with our generated predictions from Case 2 to create an optimal line-up for a specific matchday (May 15, 2021), targeting to maximize the potential FP and predict one of the winning line-ups. Even though the Case 1 results produced better results on average for the DLO procedure, the Case 2 models and data were used; this choice will be further discussed and reasoned in the next section (Discussion & Implications).

It is important to note that the optimal line-up is determined by comparing all possible potential line-up combinations, indicating that the primary objective at this stage was not to predict the total FP with absolute accuracy, however, to create a potentially high FP productive and profitable line-up. Finally, this line-up would be evaluated based on the results of an actual tournament's benchmark table on the specific matchday, challenged with actual uses that took part.

Results presented in Table 3 showcase that DLO produces an optimal line-up, with a total predicted FP of 358.7, while the real FP of this line-up was 297.5. As shown, the DLO follows the restrictions set by [57] taking under consideration the position restrictions (two SFs, two SGs, two PGs, one PF,

Table 3 Optimal line-up for May 15, 2021, matchday

Player	Position	Salary
Andre Iguodala	SF	2400
Bruce Brown	SF	4300
Caris LeVert	SG	8600
Devin Booker	SG	8500
James Harden	PG	10,800
Karl Anthony Towns	C	10,100
LeBron James	PG	9600
Thaddeus Young	PF	5600
Position restrictions		Count
PG		2
SG		2
SF		2
PF		1
C		1
Spending $:		59,900
Total predicted FP:		358.7
Total actual FP:		297.5

and one C), the salary cap (59,900$ used), players of more than 2 two teams were included, and there are no repetitions in player selection.

5 Discussion and implications

In this section, we dive into the implications and interpretations of our research findings on how individual ML models for each NBA player can be efficient and stable to predict accurately their FP performance daily. We investigate the results of using standard (core) and advanced features (complete) in different periods, identifying their pros and cons. Additionally, we discuss the real-life applications of the last stage of our research related to linear optimization and line-up generation, focusing on producing profit, considering all restrictions the tournament's provider imposed, and benchmarking our results against the line-ups of actual users.

5.1 Performance scenarios analysis

Based on the results in Sect. 4.1 and our goals in research questions 1 and 2, all scenarios provide satisfactory outcomes with standard deviations of 9.47 and 8.93 for LTS and 9.73 and 9.07 for TS in the unseen sets, respectively. Forecasting using both the most recent data (LTS) and the complete dataset (TS) provides accurate results. The two-step evaluation and validation process give stable and well-generalized models, preventing overfitting and ensuring the validity of constructed models. The differences between LTS and TS scenarios were not extreme, indicating that even the TS models capture recent trends based on the game-lag features we used, further ensuring good data quality after data wrangling and feature engineering. The results from Table 1 show that core features are more effective in predicting player performance in both (TS and LTS), outperforming complete. Furthermore, models trained in TS appear to provide slightly better scores. However, those differences are not significant, as results are computed as averages, and each player's performance trend could be captured better in different scenarios. This answers both our first and second research questions.

Moreover, based on Figs. 3, 4, 5, and 6 in the appendix, which showcase the proportion of different finalized models in each scenario, there is no unique methodology for best fitting each player's case with one type of model. Nevertheless, Voting Meta-Model consistently ranks as the best or top performing model across all scenarios except for Scenario 1 (TS complete). Voting Meta-Model performance demonstrated its robustness and adaptability when handling both complete and core datasets; moreover, its ability to combine already well performing models, excelling overall their overall predictive performance. Additionally, Random Forest performs exceptionally in Scenario 1 and fits best in other scenarios, effectively handling larger datasets with various advanced statistics capturing complex interactions and relationships. Continuously, Bayesian Ridge was also best fitting in many players' cases and datasets in all scenarios, indicating its capability to adapt to various datasets, making it a reliable and steady model.

Furthermore, AdaBoost was also a favorite in many cases, especially in Scenarios 3 and 4, suggesting that it is more effective in handling simple datasets with standard features. Lastly, Elastic Net was constantly picked as best performing across all scenarios, handling exceptionally complete and core datasets in different timespans, however, not the favorite in any scenario. In summary, this research proves that no model outperforms others, and the approach of individual modeling is efficient. Moreover, those results offer us valuable insight that the best performing models exhibit a combination of adaptability, robustness, and effectiveness in handling different types of datasets and timespans. This is related to both our first and second research questions.

5.2 Optimization interpretations

Most importantly, it is essential to highlight that each forecasting scenario is equally important in understanding the problem we are trying to solve. This research provides a comparative overview of forecasting high performers in NBA,

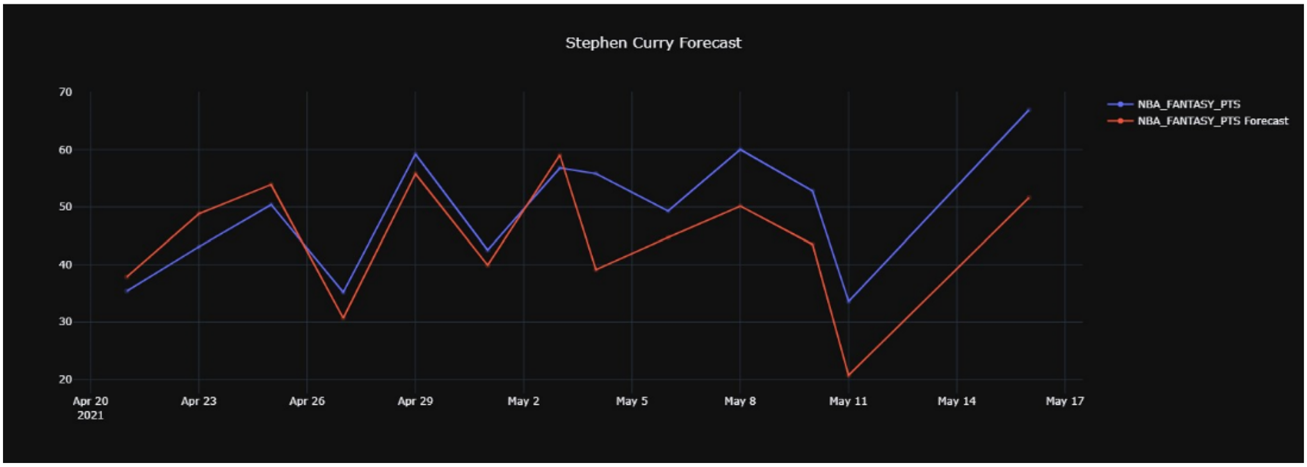


Fig. 2 S. Curry forecast results with LTS (unseen data)

taking advantage of individual players' modeling to optimize the results and provide the most efficient approach to predicting each player's performance.

Based on the aforementioned, it is crucial to analyze the optimization results from Table 2, where the most efficient and stable model from the scenarios of Table 1 is selected to create a final forecasting approach. Additionally, to ensure the validity of the results, models are benchmarked grouped by their data (TS and LTS), as these are evaluated in a different timespan of historical data in each step of the two-step evaluation and validation process.

Based on these results, both Cases 1 and 2 approaches optimize the scores in both TS and LTS, as each could include models that are trained on complete or core quality of data. While all scenarios and cases score results are close, it is more prudent to select the LTS scenarios and approaches because they are evaluated from the two-step procedure in the most recent results, and we can confidently rely on those predictions. One example is S. Curry's showcase in Figure 2, in which we can identify that the corresponding model in the LTS follows the performance trend closely, even with the difference that occurred between the actual and predicted FP.

This research also suggests a comprehensive datasets and models' approach for optimized FP predictions, indicating that the selected Case 2 approach with LTS data, including models trained on both core and complete data with standard and advanced features, is recommended. Moreover, with this approach, we can efficiently capture the latest player's performance trend and accurately predict each potential performance. This also addresses both our first and second research questions. Some of the best models that were performed in the second step of the validation procedure, with the bold indicating the lowest error metrics, are shown in Table 4.

5.3 DLO results and practical implications

The final objective of this research is to predict DFS performance accurately, supporting the decision-making process in DFS tournaments. The optimized model selection from Case 2, in which results presented in 5.2, serves as a basis for DLO process. Even though Case 1 resulted in better results, its validation set was wider and trained in all available seasons. The Case 2 with LTS data can better capture the recent form and performance trends of the players. Since dataset in which models are trained/test/validated has only recent data, it is better suited for making short-term predictions, such as those that are needed for the DLO.

After selecting the optimal combination of models from Case 2, predictions from unseen sets were used to create optimal line-ups in the DLO using linear optimization techniques. The primary goal of the DLO was achieved and tested against the actual results of a real tournament. The algorithm's results produced a successful and profitable line-up that, if entered the competition [59] of May 15, 2021, would have generated profit with real 297.5 FP scored and ranked within the winning benchmark compared to all other users. In the challenged tournament, 11,764 users participated, and profitable line-ups achieved the top 23.5%, while our DLO ranked in the top 18.4% of the leaderboard.

This comprehensive forecasting approach has a plethora of practical implications and can potentially benefit various stakeholders involved in DFS. Starting by enhancing decision-making for DFS players, using DLO results as valuable consultation to make informed decisions while selecting their line-ups, and increasing the likelihood of winning in tournaments.

Additionally, DLO can provide valuable insights for other stakeholders in the industry, such as sports analysts and DFS platform holders, to make more accurate to refine their salary

Table 4 Top five player forecasts with most accurate models with corresponding data used

Player	Model	MAPE		MAE		Data
		Validation	Unseen	Validation	Unseen	
K. Looney	Elastic Net	26.10%	**14.60%**	**4.44**	**2.49**	Complete
B. Adebayo	Random Forest	26.30%	15.70%	7.87	6.23	Core
S. Curry	Meta-model (voting)	21.60%	16.60%	8.15	7.90	Core
P. Mills	Meta-model (voting)	24.50%	16.60%	4.61	3.28	Core
Lebron James	Meta-model (voting)	**18.40%**	17.20%	8.22	7.57	Core

algorithms and create more competitive and balanced contests. DLO offers a comprehensive solution that not only maximizes winning probabilities of individual users, but also can contribute to DFS sports and sports industry.

5.4 Limitations

In this research, we have undertaken a comprehensive approach to predicting NBA player performance for Daily Fantasy Sports. However, our study does have several limitations that could be considered for future research. Firstly, our analysis does not capture certain external factors that might significantly influence player performance. These factors include player injuries, team dynamics, in-game strategic decisions, real-time game dynamics, or off-court events and trends, which are particularly relevant for Daily Fantasy Sports.

Another limitation lies in the specificity of our ML models, which are based on NBA data and tailored to the specific scenarios of this study. Given the unique aspects of the NBA league, the effectiveness of these models may be limited when applied under different conditions or with alternate datasets from other leagues. Factors such as varying playstyles, points scoring trends, and differences in player quality across leagues could affect the applicability of these models.

Additionally, our research predominantly focuses on short-term predictions, as evidenced by the development and application of the DLO. This approach means that long-term trends in player performance and their broader implications were not the primary focus of our analysis.

Furthermore, while our methodology has proven effective for NBA performance prediction, its direct applicability to other sports is not guaranteed. Sports such as football, baseball, or hockey, each with its unique dynamics and data availability, may require significant methodological adjustments to be effectively applied.

Recognizing these limitations is important as they highlight areas where further research and refinement are needed. Future studies might expand upon these aspects, thereby enhancing the models' applicability and reliability across a broader range of contexts.

5.5 Future work

Future research is needed in this area and could be explored in various directions to enhance the accuracy and applicability of this methodology. Further investigation could be done by adding other data sources [60], such as player injury history, betting odds, or sentiment analysis [61–63], which could improve model performance and excellent forecasting. Additionally, sport-specific or position-specific features could be explored by developing association rules [64] between players and teams, uncovering hidden relationships that could improve modeling. Moreover, different optimization techniques, such as genetic algorithms [65] or particle swarm optimization [66], could be tested and applied with the proposed approach and benchmark the corresponding results.

Moreover, future research and development of this proposed approach could be done in other sports, such as football, baseball, or hockey, which would offer valuable insights into the applicability of ML and linear optimization in the whole DFS industry. Additionally, this methodology could be transformed into a live recommendation system which could provide the DFS users with an up-to-date optimal line-up, considering last-minute changes from teams' squads or players' statuses. In this way, the methodology's performance could be evaluated over time through longitudinal analysis, providing insights into its consistency and reliability over seasons or years, transforming it into a reliable and effective tool for DFS users and other stakeholders in the industry.

6 Conclusion

This study presents several contributions highlighting its novelty in sports analytics and machine learning. While sports analytics with the combination of ML itself is an emerging domain, this research fills a specific gap in the literature by focusing on NBA player performance FP, using a data-driven methodology that incorporates both standard and advanced features in different time-spans, as no prior studies have exhibited the same characteristics. Additionally, the methodology goes beyond predictions and continues with produced findings with linear optimization modeling that considers real-world constraints, such as salary caps and player positions.

An extensive comparative analysis was presented from multiple ML models exploring 14 different approaches to identify the most effective predictive models. Individual modeling demonstrated that there is no one-size-fits-all solution for predicting NBA player performance, and the proposed approach is efficient. Another asset of this research it is the blended approach of data periods, combining long-term data from the last ten seasons and short-term data from the last three seasons, enhancing prediction accuracy and model stability.

Moreover, unlike typical single-step evaluations found in the literature, this study adopts a two-step evaluation and validation process on unseen data, ensuring that the generated models are robust and well-generalized. Furthermore, one more important aspect of the study's novelty is its real-world applicability and financial profitability compared with other studies in the literature. This research validates its models theoretically, enters a real-world tournament, ranked in the top 18.4% of the leaderboard, and demonstrates significant success, proving its usefulness and potential.

This study showed that the use of core (standard statistics) and complete (advanced statistics) datasets, as well as long-term (ten seasons—TS) and short-term (last three seasons—LTS) data, is critical for producing accurate predictions. Additionally, through four scenarios, it was indicated that Voting Meta-Model, Random Forest, Bayesian Ridge, AdaBoost, and Elastic Net were to be more effective models with the approach. Furthermore, in the final optimized model selection in Cases 1 and 2, it was shown that the second proposed methodology taking advantage of the two-step evaluation and validation and performing model selection in different TS and LTS timespans optimizes the overall results, picking the best stable performing model for the final approach.

Moreover, using linear optimization techniques, this research produced an optimal line-up for DFS sports with profitable results, challenged in an actual Tournament on May 15, 2021, validating the practicality and usefulness of this approach. This methodology could be the base and benefit various stakeholders in the DFS industry, such as individual users who participated in that kind of tournaments, sports analysts, and platform holders, by providing accurate predictions.

Regarding the research questions, this research targeted to address, we conclude that individual ML models for each NBA player can produce accurate predictions with well-generalized models, scoring stable scores in a two-step evaluation and validation procedures. With LTS data, the average scoring MAPE for the validation set was 28.90% and for the unseen set was 29.50%. Additionally, the MAE for validation and unseen set was 7.33 and 7.74, correspondingly. Moreover, in TS data, the average scoring MAPE for the validation set was 28.90% and for the unseen set was 29.50%. Additionally, the MAE for validation and unseen set was 7.33 and 7.74, correspondingly. The results profound achieving excelling the FP forecasting when the standard deviation of the LTS datasets was 9.47 for the train/test set and 8.93 for the unseen set, and for TS datasets was 9.73 for the train/test set and 9.07 for the unseen set (RQ1).

In addition, we conclude that in the first stage of the research, in which four different scenarios are considered, the modeling with core features was more effective in predicting player performance in short-term and long-term, improving the results in the test set by 1.7–1.9% in terms of MAPE, and in the unseen set, by 0.2–2.1%. However, we established that a combination of both modeling in core and complete data resulting in the best results, selected by MAPE in test sets, provides more stable and well-generalized models (RQ2).

In the final stage of the research, with linear optimization application to the predictions generated by Case 2 on unseen data, they determined an optimal line-up for a match day. This approach considered the necessary constraints imposed by DFS platform holders, such as player salary caps and positional requirements. The application was successful, while the generated optimal line-up was tested in a real-life scenario using data from a DFS tournament held on May 15, 2021, yielding a profitable solution that outperformed a portion of the actual users' line-ups. Our DLO performed exceptionally well in the tournament and secured a spot in the top 18.4% of the leaderboard, while there were 11,764 participants, and those who managed to create profitable line-ups made it to the top 23.5%. This successful approach showcases the practical applications of this research to various stakeholders in the DFS industry. At the same time, platform users can use this methodology for decision-making and consultancy for their line-up creation, increasing their probability of winning in DFS contests. Moreover, sports analysis and DFS platform holders can leverage these insights from this research's practical application for strategic planning, refining salary algorithms and evaluating performance more accurately (RQ3).

This study contributes to the evolving field of basketball DFS and SA, exhibiting the practical applications and benefits of employing ML for performance prediction and line-up generation and optimization. We introduced a comprehensive and optimized approach that can benefit individuals who want to benefit from the DFS and the DFS platform holders on reformatting more competitive and balanced contests by refining their salary strategies and improving decision-making at various levels in the basketball industry and performance prediction.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by George Papageorgiou. The first draft of the manuscript was written by George Papageorgiou, and all authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by HEAL-Link Greece. N/A.

Data availability Data partly available within the manuscript.

Declarations

Conflict of interest The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval This manuscript is according to the guidelines and comply with the Ethical Standards. Authors ensuring confidentiality and privacy of the research and the data obtained.

Appendix

We include here four figures that illustrate the model allocation per player and scenario during modeling (Figs. 3, 4, 5, 6), and two tables of features for complete and core datasets before feature engineering and post-feature engineering preprocessing with a glossary (Tables 5 and 6, respectively).

Fig. 3 Player models—scenario 1

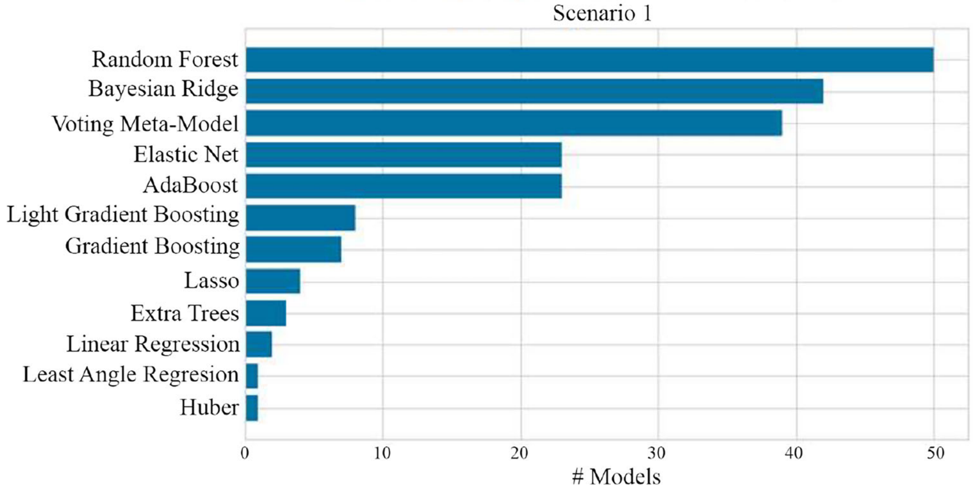
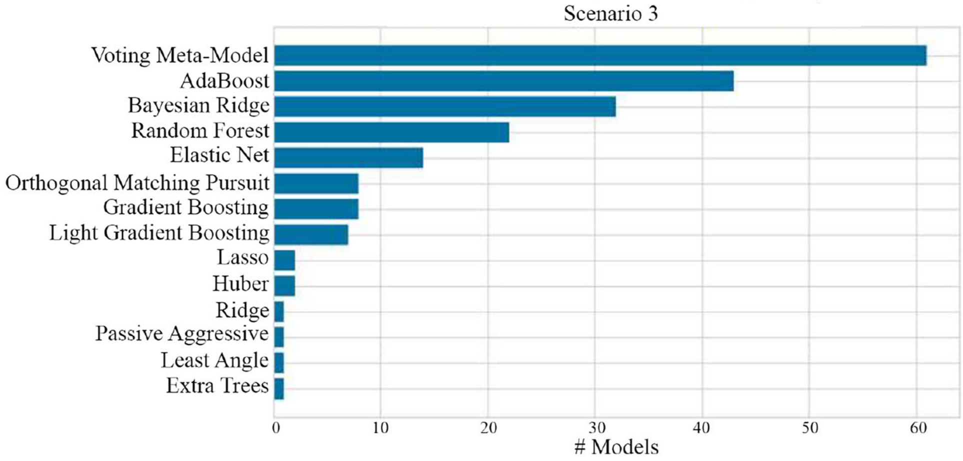


Fig. 4 Player models—scenario 2



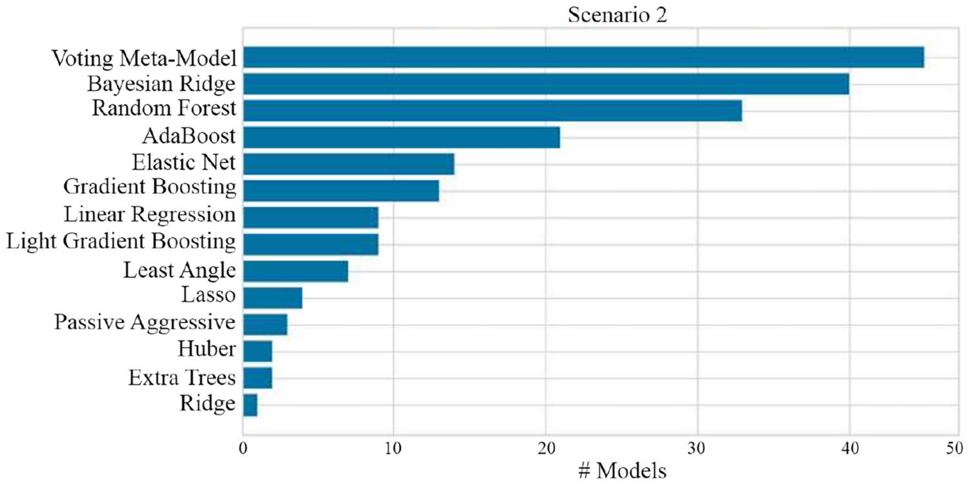


Fig. 5 Player models—scenario 3

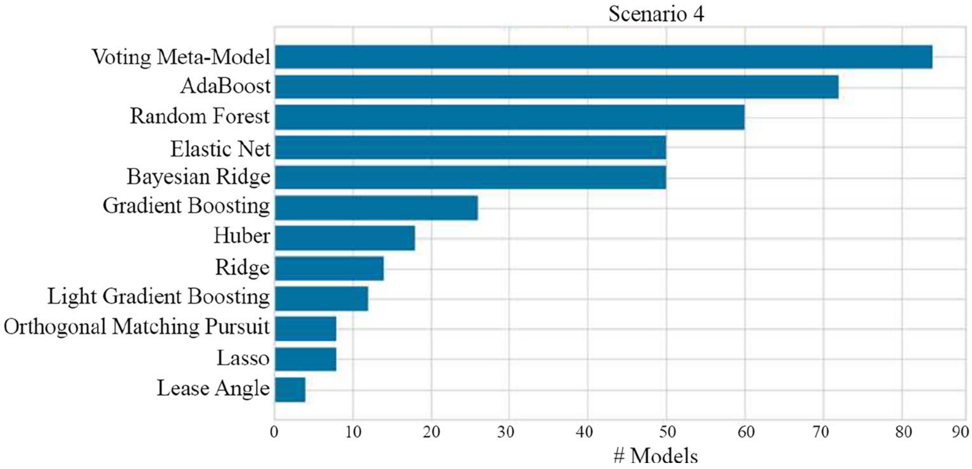


Fig. 6 Player models—scenario 4

Table 5 Complete datasets statistics and glossary: before feature engineering and post-feature engineering pre-processing

Feature	Explanation	Type
SEASON_YEAR	The season year	Traditional
PLAYER_NAME	Player's name	Traditional
TM_NAME	TM's name	Traditional
GAME_DATE	The date of match	Traditional
H/A	Home or away	Traditional
MIN	Minutes participated	Traditional
TM_MIN	TM's minutes played	Traditional
TM_OREB	TM's offensive rebound	Traditional
TM_DREB	TM's defensive rebound	Traditional
TM_REB	TM's rebound	Traditional
TM_AST	TM's assists	Traditional
TM_TOV	TM's turnovers	Traditional
TM_STL	TM's steals	Traditional
TM_BLK	TM's blocks	Traditional
TM_BLKA	TM's blocks against	Traditional

Table 5 (continued)

Feature	Explanation	Type
TM_PF	TM's Personal Foul	Traditional
TM_PFD	TM's Personal Fouls Drawn	Traditional
OPPONENT	Opponent	Traditional
OREB	Offensive Rebounds	Traditional
DREB	Defensive Rebounds	Traditional
REB	Rebounds	Traditional
AST	Assists	Traditional
TOV	Turnovers	Traditional
STL	Steals	Traditional
BLK	Blocks	Traditional
BLKA	Blocks Against	Traditional
PF	Personal Fouls	Traditional
PFD	Personal Fouls Drawn	Traditional
NBA_FANTASY_PTS	Fantasy Points	Traditional
FGM_PG	Field Goals Per Game	Traditional
FGA_PG	Field Goals Attempted Per Game	Traditional
OPP_PTS_OFF_TOV	Opponent Points Off Turnovers	Traditional
OPP_PTS_2ND_CHANCE	Opponent Second Chance Points	Traditional
OPP_PTS_FB	Opponent Fast Break Points	Traditional
OPP_PTS_PAINT	Opponent Points in the Paint	Traditional
TM_E_OFF_RATING	TM's Estimated Offensive Rating	Advanced
TM_OFF_RATING	TM's Offensive Rating	Advanced
TM_E_DEF_RATING	TM's Estimated Defensive Rating	Advanced
TM_DEF_RATING	TM's Defensive Rating	Advanced
TM_AST_PCT	TM's Assist Percentage	Advanced
TM_AST_TO	TM's Assist to Turnover	Advanced
TM_AST_RATIO	TM's Assist Ratio	Advanced
TM_DREB_PCT	TM's Offensive Rebound's Percentage	Advanced
TM_REB_PCT	TM's Rebound's Percentage	Advanced
TM_TS_PCT	TM's True Shooting Percentage	Advanced
TM_E_PACE	TM's Estimated Pace	Advanced
TM_PACE	TM's Pace	Advanced
TM_PACE_PER40	TM's Pace per 40 Min	Advanced
TM_POSS	TM's Possessions	Advanced
TM_PIE	TM's Impact Estimate	Advanced
TM_PCT_FGA_2PT	TM's Percent of Field Goals Attempted (2 Pointers)	Advanced
TM_PCT_FGA_3PT	TM's Percent of Field Goals Attempted (3 Pointers)	Advanced
TM_PCT_PTS_2PT	TM's Percent of Points (2 Pointers)	Advanced
TM_PCT_PTS_2PT_MR	TM's Percent of Points (2-Point Field Goals: Mid-Range)	Advanced
TM_PCT_PTS_3PT	TM's Percent of Points (3-Point Field Goals)	Advanced
TM_PCT_PTS_FB	TM's Percent of Points (Fast Break Points)	Advanced
TM_PCT_PTS_FT	TM's Percent of Points (Free Throws)	Advanced
TM_PCT_PTS_OFF_TOV	TM's Percent of Points (Off Turnovers)	Advanced
TM_PCT_PTS_PAINT	TM's Percent of Points (Points in the Paint)	Advanced
TM_PCT_AST_2PM	TM's Percent of Assists 2-Point Field Goals	Advanced

Table 5 (continued)

Feature	Explanation	Type
TM_PCT_UAST_2PM	TM's Percent of Unassisted 2-Point Field Goals	Advanced
TM_PCT_AST_3PM	TM's Percent of Assists 3-Point Field Goals	Advanced
TM_PCT_UAST_3PM	TM's Percent of Unassisted 3-Point Field Goals	Advanced
TM_PCT_AST_FGM	TM's Percent of Assists Field Goals	Advanced
TM_PCT_UAST_FGM	TM's Percent of Unassisted Field Goals	Advanced
E_OFF_RATING	Estimated Offensive Rating	Advanced
OFF_RATING	Offensive Rating	Advanced
sp_work_OFF_RATING	Sp Work Offensive Rating	Advanced
E_DEF_RATING	Estimated Defensive Rating	Advanced
DEF_RATING	Defensive Rating	Advanced
sp_work_DEF_RATING	Sp Work Defensive Rating	Advanced
AST_PCT	Assists Percentage	Advanced
AST_TO	Assist to Turnover Ratio	Advanced
AST_RATIO	Assist Ratio	Advanced
DREB_PCT	Defensive Rebound Rating	Advanced
REB_PCT	Rebound Percentage	Advanced
E_TOV_PCT	Estimated Turnover Percentage	Advanced
TS_PCT	True Shooting Percentage	Advanced
E_PACE	Estimated Pace	Advanced
PACE	Pace	Advanced
PACE_PER40	Pace per 40 Min	Advanced
sp_work_PACE	Sp Work Pace	Advanced
PIE	Player Impact Estimate	Advanced
POSS	Possessions	Advanced
PCT_FGA_2PT	Percentage Of Field Goals Attempted that are two-point field goal attempts	Advanced
PCT_FGA_3PT	Percentage Of Field Goals Attempted that are three-point field goal attempts	Advanced
PCT_PTS_2PT	Percentage Of Points that are from two-point field goals	Advanced
PCT_PTS_2PT_MR	Percentage Of Points that are from two-point field goals from mid-range field goals	Advanced
PCT_PTS_3PT	Percentage Of Points that are from three-point field goals	Advanced
PCT_PTS_FB	Percentage Of Points that are from fast break opportunities	Advanced
PCT_PTS_FT	Percentage Of Points that are from free throws	Advanced
PCT_PTS_OFF_TOV	Percentage Of Points that are off turnovers	Advanced
PCT_PTS_PAINT	Percentage Of Points that are from the paint	Advanced
PCT_AST_2PM	Percentage Of two-point field goals that are assisted	Advanced
PCT_UAST_2PM	Percentage Of two-point field goals that are unassisted	Advanced
PCT_AST_3PM	Percentage Of three-point field goals that are assisted	Advanced
PCT_UAST_3PM	Percentage Of three-point field goals that are unassisted	Advanced
PCT_AST_FGM	Percentage Of field goals that are assisted	Advanced
PCT_UAST_FGM	Percentage Of field goals that are unassisted	Advanced
PCT_FGM	Percentage Of Field Goal while on court	Advanced
PCT_FGA	Percentage Of Field Goal Attempts while on court	Advanced
PCT_FG3M	Percentage Of three-point field goal while on court	Advanced
PCT_FG3A	Percentage Of three-point field goal attempts while on court	Advanced
PCT_FTM	Percentage Of free throws while on court	Advanced
PCT_FTA	Percentage Of free throw attempts while on court	Advanced
PCT_OREB	Percentage Of Offensive rebounds while on court	Advanced

Table 5 (continued)

Feature	Explanation	Type
PCT_DREB	Percentage Of defensive rebounds while on court	Advanced
PCT_REB	Percentage Of Rebounds while on court	Advanced
PCT_AST	Percentage Of Assists while on court	Advanced
PCT_TOV	Percentage Of Turnovers while on court	Advanced
PCT_STL	Percentage Of Steals while on court	Advanced
PCT_BLK	Percentage Of Blocks while on court	Advanced
PCT_BLKA	Percentage Of Blocks Attempted while on court	Advanced
PCT_PF	Percentage Of Personal Fouls while on court	Advanced
PCT_PFD	Percentage Of Personal Fouls Drawn while on court	Advanced
PLAYOFFS	Playoff Match	Miscellaneous
OPPONENT	The Opponent that TM/Player faces	Miscellaneous
REST_DAYS	Days brake before last match (max "5")	Miscellaneous
WL	Result	Miscellaneous
DD2	Double–double	Miscellaneous
TD3	Triple–double	Miscellaneous
PLAYOFFS	Type of game	Miscellaneous
PCT_PTS	Percentage Of Points while on court	Scoring
NBA_FANTASY_PTS	The Fantasy Points scored	Scoring
TM_FGM	TM's field goals	Scoring
TM_FGA	TM's field goals attempted	Scoring
TM_FG_PCT	TM's field goals percentage	Scoring
TM_FG3M	TM's 3-point field goal	Scoring
TM_FG3A	TM's 3-point field goal attempted	Scoring
TM_FG3_PCT	TM's 3-point field goal percentage	Scoring
TM_FTM	TM's free throws	Scoring
TM_FTA	TM's free throws attempted	Scoring
TM_FT_PCT	TM's free throws percentage	Scoring
TM_PTS	TM's points scored	Scoring
TM_PTS_OFF_TOV	TM's points off turnovers	Scoring
TM_PTS_2ND_CHANCE	TM's 2nd chance points	Scoring
TM_PTS_FB	TM's fast break points	Scoring
TM_PTS_PAINT	TM's paint touch points	Scoring
FGM	Field goals	Scoring
FGA	Field goals attempted	Scoring
FG_PCT	Percent field goals	Scoring
FG3M	Field goals (3 pointers)	Scoring
FG3A	Field goals attempted (3 pointers)	Scoring
FG3_PCT	Field goals percentage	Scoring
FTM	Free throws	Scoring
FTA	Free throws attempted	Scoring
PTS	Points	Scoring
PTS_OFF_TOV	Points off turnovers	Scoring
PTS_2ND_CHANCE	Second chance points	Scoring
PTS_FB	Fast break points	Scoring
PTS_PAINT	Points in the paint	Scoring

Table 5 (continued)

Feature	Explanation	Type
USG_PCT	Usage percentage	Usage
E_USG_PCT	Estimated usage percentage	Usage
E_USG_PCT	Estimated usage percentage	Usage
TM_OREB_PCT	TM's defensive rebound's percentage	Four factors
TM_TM_TOV_PCT	TM's turnover percentage	Four factors
TM_EFG_PCT	TM's effective field goal percentage	Four factors
TM_FTA_RATE	TM's free throw attempt rate	Four factors
FT_PCT	Free throws percentage	Four factors
OREB_PCT	Offensive rebound rating	Four factors
TM_TOV_PCT	TM's turnover percentage	Four factors
EFG_PCT	Effective field goal percentage	Four factors
OPP_TM_OFF_RATING	Opponent's last match offensive rating	Opponent
OPP_TM_DEF_RATING	Opponent's last match defensive rating	Opponent
OPP_TM_NET_RATING	Opponent's last match net rating (difference of OFF/DEF)	Opponent
OPP_TM_NBA_FANTASY_PTS	Opponent's last match sum Fantasy Points scored	Opponent
OPP_PTS_OFF_TOV	Opponent points off turnovers	Opponent
OPP_PTS_2ND_CHANCE	Opponent second chance points	Opponent
OPP_PTS_FB	Opponent fast break points	Opponent
OPP_PTS_PAINT	Opponent points in the paint	Opponent

Table 6 Core datasets statistics and glossary: before feature engineering and post-feature engineering pre-processing

Feature	Explanation	Type
SEASON_YEAR	The season year	Traditional
PLAYER_NAME	Player's name	Traditional
GAME_DATE	The date of match	Traditional
H/A	Home or away	Traditional
NBA_FANTASY_PTS	Fantasy points	Traditional
TM_NAME	TM's name	Traditional
OPPONENT	Opponent	Traditional
OPP_TM_NBA_FANTASY_PTS	Opponent's last match sum Fantasy Points scored	Opponent
OPP_TM_OFF_RATING	Opponent's last match offensive rating	Opponent
OPP_TM_DEF_RATING	Opponent's last match defensive rating	Opponent
OPP_TM_NET_RATING	Opponent's last match net rating (difference of OFF/DEF)	Opponent
PLAYOFFS	Playoff match	Miscellaneous
REST_DAYS	Days brake before last match (over "5" assigned as "5")	Miscellaneous
MIN	Minutes participated	Traditional
PTS	Points	Scoring
FG3M	Field goals (3 pointers)	Scoring
REB	Rebounds	Traditional
AST	Assists	Traditional
STL	Steals	Traditional
BLK	Blocks	Traditional
TOV	Turnovers	Traditional

Table 6 (continued)

Feature	Explanation	Type
DD2	Double–double	Miscellaneous
TD3	Triple–double	Miscellaneous
NET_RATING	Difference between offensive rating and defensive rating	Advanced
USG_PCT	Usage percentage	Usage
PIE	Player impact estimate	Advanced
WL	Result	Miscellaneous
TM_OFF_RATING	TM's offensive rating	Advanced
TM_DEF_RATING	TM's defensive rating	Advanced
TM_NBA_FANTASY_PTS	TM's Fantasy Points scored	Scoring
NBA_FANTASY_PTS	fantasy points	Traditional
OPPONENT	Opponent	Traditional

References

1. Drazan, J.F., Loya, A.K., Horne, B.D., Eglash, R.: From Sports to Science: Using Basketball Analytics to Broaden the Appeal of Math and Science Among Youth (2020)
2. Szymanski, S.: Sport analytics: Science or alchemy? Kinesiol. Rev. **9**, 57–63 (2020). https://doi.org/10.1123/KR.2019-0066
3. Vinué, G., Epifanio, I.: Archetypoid analysis for sports analytics. Data Min. Knowl. Discov. **31**, 1643–1677 (2017). https://doi.org/10.1007/s10618-017-0514-1
4. Sarlis, V., Chatziilias, V., Tjortjis, C., Mandalidis, D.: A Data science approach analysing the impact of injuries on basketball player and team performance. Inf. Syst. **99**, 101750 (2021). https://doi.org/10.1016/J.IS.2021.101750
5. Shah, R., Romijnders, R.: Applying Deep Learning to Basketball Trajectories (2016)
6. Radovanovic, S., Radojicic, M., Jeremic, V., Savic, G.: A novel approach in evaluating efficiency of basketball players. Manag. J. Theory Pract. Manag. **18**, 37–46 (2013). https://doi.org/10.7595/management.fon.2013.0012
7. Thabtah, F., Zhang, L., Abdelhamid, N.: NBA game result prediction using feature analysis and machine learning. Ann. Data Sci. **6**, 103–116 (2019). https://doi.org/10.1007/s40745-018-00189-x
8. Georgievski, B., Vrtagic, S.: Machine learning and the NBA game. J. Phys. Educ. Sport **21**, 3339–3343 (2021). https://doi.org/10.7752/jpes.2021.06453
9. Singh, N.: Sport analytics: a review. Int. Technol. Manag. Rev. **9**, 64 (2020). https://doi.org/10.2991/itmr.k.200831.001
10. Morgulev, E., Azar, O.H., Lidor, R.: Sports analytics and the big-data era. Int. J. Data Sci. Anal. **5**, 213–222 (2018). https://doi.org/10.1007/s41060-017-0093-7
11. Wanless, L.A., Naraine, M.: Sport analytics education for future executives, managers, and nontechnical personnel. Sport Manag. Educ. J. **15**, 34–40 (2021). https://doi.org/10.1123/SMEJ.2019-0070
12. Van Haaren, J., Van Haaren, J., Zimmermann, A., et al.: Machine learning and data mining for sports analytics. In: 8th International Workshop, MLSA 2021, Virtual Event, Revised Selected Papers, p. 1571 (2022)
13. Sun, H.-C., Lin, T.-Y., Tsai, Y.-L.: Performance prediction in major league baseball by long short-term memory networks. Int. J. Data Sci. Anal. **15**, 93–104 (2023). https://doi.org/10.1007/s41060-022-00313-4
14. Hamdad, L., Benatchba, K., Belkham, F., Cherairi, N.: Data Mining for Acquiring Performances, pp. 13–24 (2018). https://doi.org/10.1007/978-3-319-89743-1_2ï
15. Ahmadalinezhad, M., Makrehchi, M.: Basketball lineup performance prediction using edge-centric multi-view network analysis. Soc. Netw. Anal. Min. (2020). https://doi.org/10.1007/s13278-020-00677-0
16. Casals, M., Martinez, J.A.: Modelling player performance in basketball through mixed models. Int. J. Perform. Anal. Sport **13**, 64–82 (2013). https://doi.org/10.1080/24748668.2013.11868632
17. Sarlis, V., Tjortjis, C.: Sports analytics—evaluation of basketball players and team performance. Inf. Syst. (2020). https://doi.org/10.1016/j.is.2020.101562
18. Evans, B.A., Roush, J., Pitts, J.D., Hornby, A.: Evidence of skill and strategy in daily fantasy basketball. J. Gambl. Stud. **34**, 757–771 (2018). https://doi.org/10.1007/s10899-018-9766-y
19. Earl, J.: Optimization of Fantasy Basketball Lineups via Machine Learning. Senior Honors Theses (2019)
20. Nelson, S.E., Edson, T.C., Grossman, A., et al.: Time out: prediction of self-exclusion from daily fantasy sports. Psychol. Addict. Behav. **36**, 318–332 (2022). https://doi.org/10.1037/adb0000829
21. Beal, R., Norman, T.J., Ramchurn, S.D.: Optimising daily fantasy sports teams with artificial intelligence. Int. J. Comput. Sci. Sport **19**, 21–35 (2020). https://doi.org/10.2478/ijcss-2020-0008
22. Bonomo, F., Durán, G., Marenco, J.: Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game. Int. Trans. Oper. Res. **21**, 399–414 (2014). https://doi.org/10.1111/itor.12068
23. Štrumbelj, E., Šikonja, M.R.: Predictive power of fantasy sports data for soccer forecasting. Int. J. Data Min. Model. Manag. **7**, 154 (2015). https://doi.org/10.1504/IJDMMM.2015.069247
24. South, C., Elmore, R., Clarage, A., et al.: A starting point for navigating the world of daily fantasy basketball. Am. Stat. **73**, 179–185 (2019). https://doi.org/10.1080/00031305.2017.1401559
25. van Bommel, M., Bornn, L.: Adjusting for scorekeeper bias in NBA box scores. Data Min. Knowl. Discov. **31**, 1622–1642 (2017). https://doi.org/10.1007/s10618-017-0497-y
26. National Basketball Association: NBA.com. In: NBA - https://www.nba.com. https://www.nba.com (2022). Accessed 1 Jul 2021
27. García, J., Ibáñez, S.J., Martinez De Santos, R., et al.: Identifying basketball performance indicators in regular season and playoff Games. J. Hum. Kinet. **36**, 161–168 (2013). https://doi.org/10.2478/hukin-2013-0016

28. de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F.: Mean absolute percentage error for regression models. Neurocomputing **192**, 38–48 (2016). https://doi.org/10.1016/j.neucom.2015.12.114
29. Willmott, C., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82 (2005). https://doi.org/10.3354/cr030079
30. Swar. NBA API: An API Client package to access the APIs for NBA.com. GitHub repository. Available at: https://github.com/swar/nba_api. Accessed 1 Jul 2021
31. Fürnkranz, J.: Web mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 899–920. Springer-Verlag, New York (2006)
32. Loeffelholz, B., Bednar, E., Bauer, K.W.: Predicting NBA games using neural networks. J. Quant. Anal. Sports (2009). https://doi.org/10.2202/1559-0410.1156
33. Shon, T., Moon, J.: A hybrid machine learning approach to network anomaly detection. Inf. Sci. (N Y) **177**, 3799–3821 (2007). https://doi.org/10.1016/J.INS.2007.03.025
34. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the ACM Conference on Computer and Communications Security, pp. 587–601 (2017). https://doi.org/10.1145/3133956.3134077
35. He, Y., Zheng, Y.: Short-term power load probability density forecasting based on Yeo–Johnson transformation quantile regression and Gaussian kernel function. Energy **154**, 143–156 (2018). https://doi.org/10.1016/j.energy.2018.04.072
36. Katrutsa, A., Strijov, V.: Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst. Appl. **76**, 1–11 (2017). https://doi.org/10.1016/j.eswa.2017.01.048
37. Imaam, F., Subasinghe, A., Kasthuriarachchi, H., et al.: Moderate automobile accident claim process automation using machine learning. In: 2021 International Conference on Computer Communication and Informatics (ICCCI), pp, 1–6. IEEE (2021)
38. Ali, A.: PyCaret: an open source, low-code machine learning library in Python. In: PyCaret version 1.0.0. https://www.pycaret.org (2020). Accessed 15 Apr 2022
39. Larsen, E., MacVittie, K., Lilly, J.: A Survey of Machine Learning Algorithms for Detecting Malware in IoT Firmware (2021)
40. Massaoudi, M., Refaat, S.S., Abu-Rub, H., et al.: A Hybrid Bayesian ridge regression-CWT-Catboost model for PV power forecasting. In: 2020 IEEE Kansas Power and Energy Conference (KPEC), pp. 1–5. IEEE (2020)
41. Maulud, D., Abdulazeez, A.M.: A review on linear regression comprehensive in machine learning. J. Appl. Sci. Technol. Trends **1**, 140–147 (2020). https://doi.org/10.38094/jastt1457
42. Eyraud, R., de la Higuera, C., Janodet, J.-C., et al.: LARS: a learning algorithm for rewriting systems. Mach. Learn. **66**(1), 7–31 (2006). https://doi.org/10.1007/S10994-006-9593-8
43. Kapelner, A., Bleich, J.: Bartmachine: machine learning with bayesian additive regression trees. J. Stat. Softw. (2013). https://doi.org/10.48550/arxiv.1312.2171
44. Cai, T.T., Wang, L.: Orthogonal matching pursuit for sparse signal recovery with noise. IEEE Trans. Inf. Theory **57**, 4680–4688 (2011). https://doi.org/10.1109/TIT.2011.2146090
45. Yin, G., Jameel Ibrahim Alazzawi, F., Mironov, S., et al.: Machine learning method for simulation of adsorption separation: comparisons of model's performance in predicting equilibrium concentrations. Arab. J. Chem. **15**, 103612 (2022). https://doi.org/10.1016/J.ARABJC.2021.103612
46. Solomatine, D.P., Shrestha, D.L.: AdaBoost.RT: a boosting algorithm for regression problems. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), pp. 1163–1168. IEEE (2004)
47. Liu, Y., Wang, Y., Zhang, J. New Machine Learning Algorithm: Random Forest, pp. 246–252 (2012)
48. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Front. Neurorobot. **7**, 21 (2013). https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX
49. John, V., Liu, Z., Guo, C., et al.: Real-Time Lane Estimation Using Deep Features and Extra Trees Regression. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9431, pp. 721–733 (2016). https://doi.org/10.1007/978-3-319-29451-3_57/FIGURES/5
50. Roth, V.: The generalized LASSO. IEEE Trans. Neural Netw. **15**, 16–28 (2004). https://doi.org/10.1109/TNN.2003.809398
51. Chakraborty, D., Elhegazy, H., Elzarka, H., Gutierrez, L.: A novel construction cost prediction model using hybrid natural and light gradient boosting. Adv. Eng. Inform. (2020). https://doi.org/10.1016/J.AEI.2020.101201
52. Rathore, S.S., Kumar, S.: A decision tree regression based approach for the number of software faults prediction. ACM SIGSOFT Softw. Eng. Notes **41**, 1–6 (2016). https://doi.org/10.1145/2853073.2853083
53. Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G.: The Feature Importance Ranking Measure. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5782, pp. 694–709. LNAI (2009). https://doi.org/10.1007/978-3-642-04174-7_45
54. Bhat, P.C., Prosper, H.B., Sekmen, S., Stewart, C.: Optimizing event selection with the random grid search. Comput. Phys. Commun. **228**, 245–257 (2018). https://doi.org/10.1016/j.cpc.2018.02.018
55. Ramezan, C.A., Warner, T.A., Maxwell, A.E.: Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Remote Sens. (Basel) **11**, 185 (2019). https://doi.org/10.3390/rs11020185
56. Mitchell, S.: PuLP: A Linear Programming Toolkit for Python (2011)
57. DraftKings Daily Fantasy: DraftKings. In: DraftKings - https://www.draftkings.com. https://www.draftkings.com (2021). Accessed 15 May 2021
58. Hewamalage, H., Ackermann, K., Bergmeir, C.: Forecast evaluation for data scientists: common pitfalls and best practices. Data Min. Knowl. Discov. **37**, 788–832 (2023). https://doi.org/10.1007/s10618-022-00894-5
59. Fantasy Labs: $20 NBA $200K Shootaround [50K to 1st] (11764 entries). In: RotoGrinders - https://rotogrinders.com/resultsdb/nba. https://rotogrinders.com/resultsdb/nba (2021). Accessed 15 May 2021
60. Grossi, V., Giannotti, F., Pedreschi, D., et al.: Data science: a game changer for science and innovation. Int. J. Data Sci. Anal. **11**, 263–278 (2021). https://doi.org/10.1007/s41060-020-00240-2

61. Tsiara, E., Tjortjis, C.: Using twitter to predict chart position for songs. In: IFIP Advances Information Communication Technology, vol. 583, pp. 62–72. IFIP (2020). https://doi.org/10.1007/978-3-030-49161-1_6/TABLES/2
62. Nousi, C., Tjortjis, C.: a methodology for stock movement prediction using sentiment analysis on twitter and stocktwits data. In: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), pp. 1–7. IEEE (2021)
63. Liu, H., Morstatter, F., Tang, J., Zafarani, R.: The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. Int. J. Data Sci. Anal. **1**, 137–143 (2016). https://doi.org/10.1007/s41060-016-0023-0
64. Wang, C., Tjortjis, C.: PRICES: An Efficient Algorithm for Mining Association Rules, pp. 352–358 (2004)
65. Reed, P., Minsker, B., Goldberg, D.E.: Designing a competent simple genetic algorithm for search and optimization. Water Resour. Res. **36**, 3757–3761 (2000). https://doi.org/10.1029/2000WR900231
66. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95—International Conference on Neural Networks, pp. 1942–1948. IEEE (1995)