Knowledge and Information Systems (2024) 66:4333–4375
https://doi.org/10.1007/s10115-024-02092-9

REGULAR PAPER

Evaluating the effectiveness of machine learning models for performance forecasting in basketball: a comparative study

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

Received: 9 October 2023 / Revised: 18 December 2023 / Accepted: 28 February 2024 /
Published online: 24 March 2024
© The Author(s) 2024

Abstract

Sports analytics (SA) incorporate machine learning (ML) techniques and models for performance prediction. Researchers have previously evaluated ML models applied on a variety of basketball statistics. This paper aims to benchmark the forecasting performance of 14 ML models, based on 18 advanced basketball statistics and key performance indicators (KPIs). The models were applied on a filtered pool of 90 high-performance players. This study developed individual forecasting scenarios per player and experimented using all 14 models. The models' performance ranking was developed using a bespoke evaluation metric, called weighted average percentage error (WAPE), formulated from the weighted mean absolute percentage error (MAPE) evaluation results of each forecasted statistic and model. Moreover, we employed a comprehensive forecasting approach to improve KPI's results. Results showed that Tree-based models, namely Extra Trees, Random Forest, and Decision Tree, are the best performers in most of the forecasted performance indicators, with the best performance achieved by Extra Trees with a WAPE of 34.14%. In conclusion, we achieved a 3.6% MAPE improvement for the selected KPI with our approach on unseen data.

Keywords Data mining (DM) · Data science · Forecasting · Machine learning (ML) · Sports analytics (SA)

✉ 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 - Moudania, 570 01 Thermi, Greece

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

1 Introduction

Sports analytics (SA) is a vast, developing domain, significant for organizations, teams, and players. Many researchers are developing ideas to provide valuable insights. These can relate to performance evaluation, injury prevention, performance forecasting, or decision-making on tactics and strategies [1]. While basketball is a sport that combines a plethora of statistics, machine learning (ML) and data mining (DM) applications are becoming popular in the data science (DS) research community, with constant research and development, trying to apply and improve their ideas on real cases in NBA and other leagues. However, research requires valid data, collected via various techniques and media (cameras, sensors), to achieve improvements in the SA domain [2].

Each ML and DM technique can be implemented in sports, especially basketball. With the advanced statistics that many basketball leagues offer, there is room for improvement in ML and DM applications in SA. Comprehensive analysis and performance prediction are highly interesting for most prominent clubs, which invest in creating DS and SA department for scraping insights [3].

Researchers have developed statistics to provide a clear view of a player's performance throughout a game; some of the crucial key performance indicators (KPIs) are efficiency (EFF), game score (GMSC), player impact estimate (PIE), player efficiency rating (PER), tendex (TENDEX), (FP), Four Factors (FOUR FACTORS), and Usage Rate [4]. Based on these, teams, technical staff, and organisations rank and evaluate player performance. At the same time, the foresaid metrics are developed as a formula for attaching defensive and teamwork statistics. The analysis of a player or a game is now more straightforward and more apparent for people who need to make decisions [5].

This research aims to provide a clear overview of ML models' performance in 18 different kinds of advanced basketball metrics and KPIs, based on 90 high-performance players' case studies, for basketball player performance forecasting (BPPF). Fourteen models are evaluated in each player's case and his advanced targeted metrics. The list of models used include AdaBoost (AB), K-nearest neighbors (KNN), decision trees (DTs), extra trees (ET), light gradient boosting machine (LGBM), elastic net (EN), random forest (RF), gradient boosting machine (GBM), passive aggressive (PA) Regressor, Bayesian Ridge (BR), least angle regression (LARS), ridge regression (RR), Huber Regression (HR) and least absolute shrinkage and selection operator (LASSO).

To achieve that, we followed an original approach of 381 game-lag features and the application of ML regression models, to predict the upcoming performance of each of 90 high-performance players from the filtered pool in each case, using scraped data characterised as advanced statistics (Base, Advanced, Miscellaneous, Four Factors, Scoring, Opponents, Usage) related to players and teams from season 2019–20 up to season 2021–22. The selected players can be considered as high performance. These are filtered from all active NBA players by their GameScore (GMSC), FOUR FACTORS, TENDEX, FP and Efficiency (EFF) averages and their participation time, excluding the players that did not participated for at least 150 games during the last three seasons (2019–20, 2020–21, 2021–22) and at least 30 games during the last season (2021–22). They should have at least twenty minutes average participation time in the previous three seasons (2019–20, 2020–21, 2021–22). Besides, each player who's foresaid KPIs averages are below the league's average score is excluded.

An additional aim of this study was to rank basketball forecasting models not only based on their forecasting performance, but also on how feasible it is to produce individual predictions for each targeted advanced basketball statistic and overall performance. To achieve this

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

goal, we employed a comprehensive forecasting approach, which involved analyzing different prediction options and presenting an overview of the predictions that can be improved. The methodology for this study included two experiments. The first experiment focused on forecasting Fantasy Points (FP) as a single metric, while the second experiment predicted individual performance metrics such as Points (PTS), Rebounds (REB), Assists (AST), Steals (STL), Blocks (BLK), and Turnovers (TOV), which are used to formulate the FP. These individual predictions were then used to construct the forecasted FP formula. The results of both experiments were compared to assess the effectiveness of the two processes. The study found that by expanding the forecasting options and using a comprehensive forecasting approach, predictions of KPIs can be significantly improved.

This research is a significant contribution to ML applied to sports, as it evaluates the forecasting abilities of various ML models in predicting basketball player performance. The study focuses on a set of 18 advanced basketball statistics and KPIs and applies 14 ML models to a group of 90 high-performance basketball players. The main objective of this investigation is to identify the best ML models for individualized prediction of advanced basketball statistics and to evaluate their overall effectiveness in forecasting basketball player performance, assessing the complexity and dynamism of player performance.

The research is distinctive in its approach, developing individual forecasting scenarios for each player and utilizing a bespoke evaluation metric: weighted average percentage error (WAPE), to evaluate the accuracy of the predictions. This metric takes into account the weighted mean absolute Percentage error (MAPE) of each predicted statistic and model, providing a detailed comparison of different ML models.

By leveraging the latest three seasons of NBA advanced box-scores statistics and applying extensive data preprocessing and feature engineering, the study was conducted not only to evaluate the performance of ML models, but also to introduce an innovative and comprehensive approach to improve KPI forecasting results. This approach includes predicting individual statistics that contribute to a KPI and then modifying the KPI using these forecasts, which resulted in a significant improvement in the accuracy of future predictions.

The research findings have consequential implications for the sports analytics industry, as they offer valuable insights for researchers, coaches, data scientists, and stakeholders. The study sets a new benchmark in predicting player performance, combining sophisticated statistical techniques with practical usefulness in the competitive world of professional basketball.

2 Background

With the constant improvement of ML and DM applications, different industries are on DS and DM chase for evaluation, improvement, forecasting, and optimisation. SA is now an excellent tool for organisations, and professional teams to use to advise decision-making and plan their strategies. ML and DM use, especially in basketball, has been beneficial until now [6]. However, as professional leagues offer data, there is plenty of room for improvement and testing. Such applications include overall player performance evaluation and predictions, injury prediction, play style strategies and line-up combos. The recent years, researchers tried to get the best results with innovations and approached each case differently.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

2.1 Basketball players' performance prediction literature overview

All major sports organizations and professional teams use SA to assemble their teams, improve each player's performance, and pinpoint problems difficult for coaches and staff to detect. SA is a constantly involving domain, so technological advancements have made it possible and essential for coaches, staff, and corresponding teams [7]. Relying on decision-making on SA and predictive analytics provide teams and organizations with the decisiveness that their actions are taken based on valid data. Furthermore, with ML and DM techniques, predictive analytics development allows researchers to extend their experiments with SA, propose new approaches, and evaluate their findings [8].

For the first time, researchers in [9] forecasted the NBA player's performance using sparse functional data, providing a competitive method in contrast with the other traditional methods. Also, in the study [10], a unique network with ML and graph theory is developed to predict the performance of an NBA line-up anytime based on a founded metric called Inverse Square Metric, using an edge-centric method achieved 80% average accuracy and with graph-theory, performance prediction results yield 10% in comparison with baseline methods. Additionally, researchers in [11] claimed to determine the key factors and statistics for a team to win the game. Their case study of Golden State Warriors claimed that the winning success factors related firstly to shooting and after to defensive rebounds and opponent turnovers. Furthermore, the study [12] uses a graph theory neural network-based model for injury prediction.

In contrast, in the study [13], validation based on versatility or specialisation is done for basketball players, claiming that by filtering only the best players, a trend of higher numbers of versatility is shown compared to the specialisation. In addition, researchers [14] correlate NBA players' performance with their personality features. Comparing All-Star players with the rest of the league, they concluded that the traits of conscientiousness and agreeableness had the biggest significant positive difference. With a different approach [15], data envelopment analysis (DEA), researchers investigated the correlation between winning probabilities and game outcomes for NBA teams, claiming that the DEA-based approach successfully predicts team performance.

The researchers [4] correctly predicted the NBA MVP for the 2017–18, 2018–19, and 2019–20 NBA seasons. In addition, based on verified data from seasons 2017–18 up to 2019–20, they forecasted the best Defender for the aforesaid NBA seasons. Each season's dataset comprised 82 game events in each forecast scenario split into four groups(Q1-Q4). They selected a pool of twenty NBA players filtered by the number of games (at least thirty games per season) and their participation time (fifteen minutes per game-event). With extended analysis, they created two metrics, the Aggregated Performance Indicator (API) and the Defensive Performance Indicator (DPI). Based on these two metrics, using API, which is constructed by advanced statistics that illustrate the player's general performance, they successfully predicted the NBA MVP for seasons 2017–18 up to 2019–20. With the use of DPI, a composition of advanced analytics variables focused on player contribution to Defence, they successfully predict the Best Defender for seasons 2017–18 up to 2019–20.

The study in [16] presents an approach to determining the critical factor on which each player's shooting performance accuracy in the NBA depends. Researchers experimented with seven different models based on ten statistics related to shooting to predict whether a player could make the shot. Their results show that shot distance, the distance of the closest defence player and touch time are the three most crucial variables impacting a player's successful field goal accuracy. Their results concluded that KNN (KNN) had performed best with 67.6% classification accuracy.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Furthermore, researchers at [17] also implement ML models to predict the potential shooting accuracy of NBA players, stating that someone must focus on the variables that this metric depends on, targeting to indicate a key performance metric like successful shooting points. For this reason, they tried to classify each player's efficiency at shooting from various ranges and frequently employed defensive tactics. To reach their targets, they used eXtreme GBM (XGBoost) and RF, figuring that XGBoost was the best choice scoring 68% accuracy with parameter tuning and 60% without tuning. However, they claimed that RF is also a good choice scoring 57% in their experiment.

Considering basketball players' performance evaluation, in this study [18], two methods were employed to determine the crucial variables for each player's position and construct an alternative performance evaluation system similar to the Performance Index Rating (PIR). Firstly, they clustered the players based on their position for their research on data from Euroleague 2017–18. Secondly, DT and one-way ANOVA tests determine the critical variables for each position, and TOPSIS results are compared with PIR for indexing players into a ranking system. They claimed that it is possible with this alternative way to determine player performances finally.

The authors in [19] identified if a player belongs to All-Stars after the end of each regular season in the NBA, based on his advanced box score statistics; additionally, they targeted to identify the most important characteristics that make a player an All-Star player—started with the employment of RF model on data from seasons 1936–37 up to 2010–11, for classification. To continue, while they succeeded in creating an ML model capable of classifying correctly with an accuracy of 92.5%, they built up an application with Apache Spark to simplify the process. To conclude, even if the selection of players for the NBA All-Star game purely depends on votes, their approach can predict the potential NBA All-Star players.

However, since the previous work in performance prediction for the past years mainly focused on NCAAB, the study [20] tried to identify if there is a possibility to use data from NCAAB for ML and DM applications for performance prediction in NBA or the opposite. Across their research, several representations, training settings, and classifiers for comparing their results on NCAAB and NBA data. Additionally, they used three different metrics to evaluate and predict the team's performance, adjusted EFF and Adjusted FOUR FACTORS. They discovered that adjusted efficiencies work well for the NBA; besides, for predicting the NCAAB post-season period, the regular season for training is not the best choice. Also, they claimed that to predict as better as possible team" performance, different classifiers with different bias needed for each league. Finally, based on their findings, they conclude that the best classifier for predicting the outcome of the NBA playoff series is the naïve Bayes.

Players' Performance predictions can be based on different metrics and KPIs; one of them that also has many applications in the betting domain is FP. Advanced box score statistics structure this KPI, which can show an overview of the attacking, defensive and teamwork performance of each player participating in a game. In recent years many researchers tried to predict players' performance with FP and, in many cases, use their findings or Fantasy Tournaments case studies for betting applications [21]. The researchers [22] tried to predict the potential FP and develop a system capable of predicting the best combination of players for the Daily Fantasy Line-ups application. Firstly, they used Bayesian random-effects model and data from season 2013–14 up to season 2015–16, in which they conducted their experiments. After the results were acquired, they compared two methods of constructing the forecasted line-up with a Bayesian random-effects model and a KNN model. Finally, they conclude that both approaches have successful results, with KNN coming first in generating profits in Fantasy Tournaments.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

The study by [23] tried to predict the final score of an NBA game using data from seasons 2017–18. They experiment with a hybrid data-mining-based scheme using five data mining models, Extreme Learning Machine (ELM), Multivariate Adaptive Regression Spline (MARS), XGBoost and a KNN approach and game-lag features. The empirical results proved that the XGBoost mode achieved the best performance, using game-lag = 4. At the same time, they also presented the most critical vital statistics features for their forecast. Aiming for the same results, researchers [24] proposed a new intelligent ML framework that claimed to predict the results of a game played in the NBA. Nevertheless, based on this, they also experiment with the key factors and statistics that are critical for their forecast. Using Naïve Bayes, Artificial Neural Networks (ANNs), and DT, they were confident that defensive rebound is one of the essential features with others to follow, concluding on with the proper feature selection, models' performance increased from 2% up to 4%.

The authors of [25] experiment with data mining methods targeting to predict the correct NBA GMSC. Their applications involved the five most-known data mining methods, multivariate adaptive regression splines (MARS), KNN, extreme learning machine (ELM), extreme GBM (XGBoost) and stochastic GBM (SGB), finalising their research on creating a successful GMSC prediction model. While in [26], the authors tried to predict the outcome of NBA playoffs by creating a scheme with k-means clustering and the maximum entropy principle.

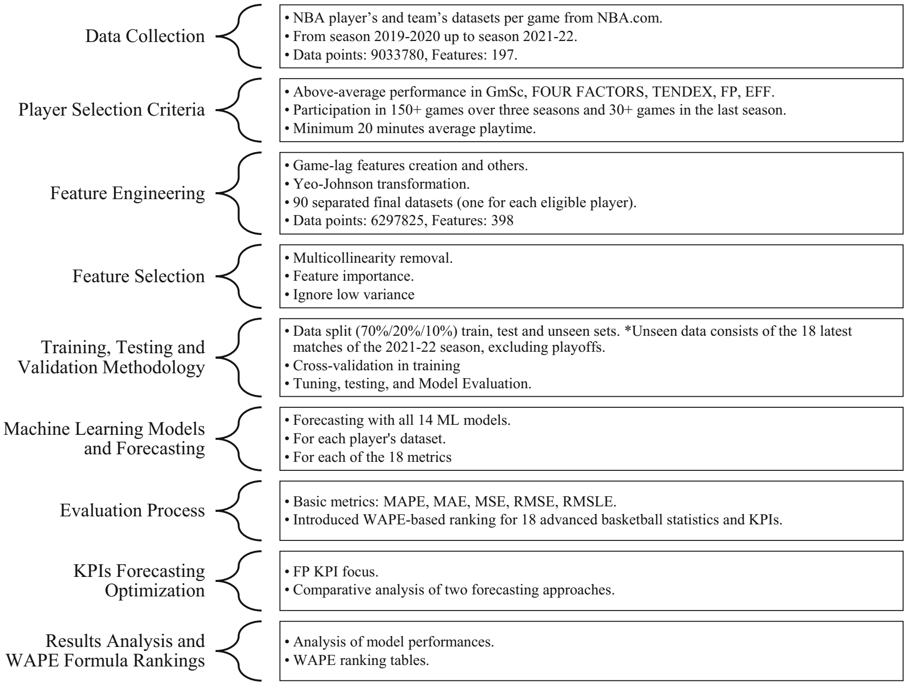
3 Methodology

This section outlines the methodology that followed. Starting with basketball's data availability, scrapped data are from the official NBA website [27] from the Seasons 2019–20, 2020–21, and 2021–22. Including plenty of evaluation and performance statistics, Player's and Team's Box Scores for each game, related to the attack, defence, teamwork and advanced KPIs, which overview total each player's performance [28]. Continuously, cleansing and transformations are performed on the data and the essential pre-processing on both Player's and Team's Box Scores related to each recorded game and merging them to continue with feature engineering. In the next stage, 1,3,5,7 and 10 game-lag features are created from base data for implementing regression ML model forecasting. In the forecasting phase, 18 different advanced basketball performance statistics and KPIs with 14 different types of ML models used, AB, KNN, DT, ET, LGBM, EN, RF, GBM, PA, BR, LARS, HR, RR and LASSO.

Furthermore, in each case study, per player of the selected pool, 18 advanced statistics and KPIs are tested, forecasted and evaluated with each of the 14 different ML models. As mentioned earlier, the goal was to create a performance ranking table for the trained models to assess which model or type of model performs better for forecasting each statistic and KPI that overview player performance [28]. The Ranking Table is based on MAPE results per model and metric, introducing also the WAPE metric. The created key indicator will be analysed in the following sections.

Finally, the last experiment is conducted to yield KPIs results. We are introducing a selective and models' comprehensive approach for calculating the average of KPIs performance prediction evaluation scores. The KPI for the last experiment is based on Fantasy Points (FP). This formula is constructed on different player statistics, evaluating the total players' performance from different perspectives. Per statistics results of the last experiment will be analysed, considered, and constructed with results as a formula. The summarized workflow of the methodology utilized is illustrated in Fig. 1. It outlines the progression from data

Content courtesy of Springer Nature, terms of use apply. Rights reserved.



Data Collection
- NBA player's and team's datasets per game from NBA.com.
- From season 2019-2020 up to season 2021-22.
- Data points: 9033780, Features: 197.

Player Selection Criteria
- Above-average performance in GmSc, FOUR FACTORS, TENDEX, FP, EFF.
- Participation in 150+ games over three seasons and 30+ games in the last season.
- Minimum 20 minutes average playtime.

Feature Engineering
- Game-lag features creation and others.
- Yeo-Johnson transformation.
- 90 separated final datasets (one for each eligible player).
- Data points: 6297825, Features: 398

Feature Selection
- Multicollinearity removal.
- Feature importance.
- Ignore low variance

Training, Testing and Validation Methodology
- Data split (70%/20%/10%) train, test and unseen sets. *Unseen data consists of the 18 latest matches of the 2021-22 season, excluding playoffs.
- Cross-validation in training
- Tuning, testing, and Model Evaluation.

Machine Learning Models and Forecasting
- Forecasting with all 14 ML models.
- For each player's dataset.
- For each of the 18 metrics

Evaluation Process
- Basic metrics: MAPE, MAE, MSE, RMSE, RMSLE.
- Introduced WAPE-based ranking for 18 advanced basketball statistics and KPIs.

KPIs Forecasting Optimization
- FP KPI focus.
- Comparative analysis of two forecasting approaches.

Results Analysis and WAPE Formula Rankings
- Analysis of model performances.
- WAPE ranking tables.

Fig. 1 Performance Forecasting Methodology Workflow

collection to the final stages of forecasting and evaluation.

3.1 Research questions (RQs)

1. Which ML Model is best for predicting individually Advanced Basketball Statistics? (RQ1)
2. Which ML Models are the better performers in Basketball Player Performance Forecasting? (RQ2)
3. How can the Basketball Player Performance be improved using a comprehensive forecasting approach for KPIs? (RQ3)

These questions are essential for organisations, teams and especially SA and DS departments, which provide executive advice for player decision-making and improvement at multiple levels. In addition, they offer a clear view of the critical contribution and its uses of KPIs in SA on how these can be useful for prediction-making and evaluating each athlete's existing or potential performance [29].

3.2 Aim and objectives

This study aims to accurately predict the key indicators that overview an NBA player's performance and identify and benchmark the available ML models forecasting performance for each key metric and KPIs contrasting 90 high-performance player cases. Resulting in an accurate and validated models' performance ranking table based on 90 forecast case studies

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

that previews which ML model's type is the best suitable for SA performance forecasting under the methodology followed.

Additionally, based on the models' performance ranking table, a forecasting approach for the selected KPI, FP, will be constructed. Since FP is one of the KPIs built by players' key metrics related to the attack, defence, and teamwork, it is one of the preferable ones that overviews and evaluates each player's performance. For this reason, based on the models' performance ranking, a comprehensive ML models' approach is followed to forecast and assess each of the 90 selected pool players, but as an average of the whole pool.

3.3 Data acquisition & pre-processing

Official NBA's website offers plenty of statistics and information, including Box Scores for players and teams [27]. The retrieved and used data were from the season 2019–20 up to 2021–22 for regular seasons and playoffs, targeting to find the latest trend for players' and teams' performance. Scraped data are referred to as advanced statistics, with different types; base, advanced, miscellaneous and scoring data for all players that participated in the referred seasons, same with base, advanced, miscellaneous, scoring, four factors and opponents type of data for all NBA teams. Based on performance and participation criteria, the study focused on only high-performance players during the pre-processing and cleaning. Long-time injured players, rookies and players who do not still participate in NBA are excluded. It started with banning the players that had the selected KPIs, GmSc, FOUR FACTORS, TENDEX, FP and EFF averages below the average of the league for the three seasons. Also, we excluded the players that did not participate in at least one hundred fifty games in the last three seasons (2019–20, 2020–21, 2021–22) and at least thirty games in the previous season (2021–22), additionally, players that had less than twenty minutes average participation time in last three seasons were excluded.

In the pre-processing phase, each player's dataset is merged with their corresponding team, creating new features about per-game opponent performance. Furthermore, statistics about final rankings are excluded because those could not provide any information about players' potential performance in each upcoming game. The required 90 datasets included one hundred 90-seven features and statistics related to attack, defence, and teamwork, which are referred to as basic, advanced, and informative KPIs.

3.4 Feature engineering

This study uses game-lag features in feature engineering, creating 1, 3, 5, 7 and 10 games-lag features. Those game-lag features are designed as averages of the previous performance for each player in each statistic, except for the categorical features, which are calculated as sums. It is worth mentioning that all primary features transformed into game-lag, but the averages are applied only in the following advanced metrics [30] Plus-Minus (PM), FOUR FACTORS, Net Rating (NETRTG), EFF, TENDEX, GMSC, PIE, Effective Field Goal Percentage (EFG%), FP, Usage Percentage (USG%), Assists to Turnover (AST/TO), PTS, REB, AST, Assists Ratio (AST RATIO), STL, BLK and TOV. Additionally, 3-game-lag sums are created and used for the categorical features; Win/Lose, Double- Doubles, Triple-Doubles and Minutes of participation time. Lastly, to avoid players' participation in games in which they were injured, and their participation time was limited, causing outliers in the dataset, appearances with under twelve minutes of participation time are excluded. However,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 1 Summary of Dataset Characteristics and Player-Specific Records

Description	Value
Total records in the dataset	16025
Number of features in the dataset	398
Total number of separate player datasets	90
Mean records per player	178
Minimum records for a single player	126
Maximum records for a single player	235

their historical information is kept under the game-lag features. After feature engineering, each of the 90 datasets contained 398 features, as presented in Table 1, with datasets structures.

Before starting each of the 90 case studies, the transformation of each target variable is done with the Yeo-Johnson method. That method is selected because basketball performance can be influenced by various reasons, like how many minutes the coach will decide that the player will participate in the game. Those reasons cause the distribution of each statistic, or KPI, to be non-symmetrical. With the Yeo-Johnson method [31], we made the distribution of those more symmetric. Additionally, even if basketball is full of statistics and metrics, the quantity of those is limited. Each NBA season has 82 games plus the playoffs, but not all players participate in all games each year. Across our study, the last three seasons were selected, and the ratio of the records and the generated features was not ideal. For this reason, three feature selection [32] methods are followed; in each experiment, different features did not add to the explained variance of the model and were removed.

The first method, removing multicollinearity between features [33], is applied because datasets contain highly correlated features. After all, a player's performance could be defined on specific levels (attack, defence, teamwork), and the variance of the coefficients is increased, generating noise. This method drops each feature highly linearly correlated with another feature and less correlated with the target variable. The threshold is set to 0.50, causing the features with inter-correlations higher than 0.50 to be dropped. At the next stage, the feature importance method is applied [34], aiming to constrain the feature space and improve modelling efficiency, using a mix of permutation importance approaches of linear correlation with the target variable, RF and AB. With the feature selection threshold set to 0.9, the model keeps only the features that explain at least 90 per cent of the dataset's variance. Lastly, the ignore low variance method [35] focuses on the categorical features, playoffs, opponent, and season year. According to this method, features with statistically insignificant variances are removed, and their variance is calculated by dividing the number of samples by the number of their unique values. The two conditions set were, firstly, the count of the unique values of each feature divided by the sample size to be less than ten per cent. Secondly, the count of the most common values divided by the count of the second most common values is greater than twenty. Data are available at each of the three stages throughout this research—initially as raw collections, then pre-processed, and finally after feature engineering—all of which are accessible in the linked GitHub repository.

3.5 Modelling

Across this research, the main target is to predict and compare the results of each of the selected ML models for each basketball performance statistics, KPIs and overall performance.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

By this, a ranking table based on the bespoke metric WAPE overviews and benchmarks the performance of each ML model. After data preparation, each of the 90 different datasets, one for each player, 18 BPPF metrics with 14 different ML models, will be tested.

The Pycaret library allows us to train, tune, test and evaluate the models simultaneously. Pycaret library has, except for regression applications, uses in classification, clustering, anomaly detection, natural language processing, association rules mining and time series [36]. Additionally, it offers automated feature selection and model ensembling (bagging, boosting, blending) methods, providing optimised model performance [37, 38].

3.6 Linear, tree based, non-parametric and online learning models in SA

In this study, an extending forecasting performance analysis has been done with a pool of 90 players with their corresponding nineteen performance metrics and KPIs. In addition, 14 different models have been trained in the case studies and will be evaluated based on their performance in SA applications. Those models can be separated into four categories; Linear-Based [39], Tree-based [40], Non-Parametric [41], Online-Learning [42].

3.6.1 Linear models

The models used for ML model creation are based on a linear combination of features, and the target value belongs in the Linear-based category. It refers to a linear approach for modelling between a scalar and explanatory variable, while the assumption that dependent and independent variables are linearly related, and the model works by finding the line which fits linearly better between the variables. Because of this, it performs excellently for linear separating data. Based on the methodology, the Linear-based models follow for regression applications; an execution formula is applied to find the best-fit line throughout the set of training data [43]. These models, foundation based on statistical modeling, highlight the intersection of data-driven prediction and statistical theory [44].

Least Absolute Shrinkage and Selection Operator (LASSO): A linear regression and regularisation technique from statistical learning [44], using shrinkage to determine the coefficients. Additionally, called as penalized regression method because it penalizes the less essential features and automatically performs feature selection. Therefore, it usually is preferred when data have high dimensionality and multicollinearity [45].

Elastic Net (EN): A penalized linear regression model that can consider a hybrid of RR and LASSO regularization. Similar to LASSO, it generates reduced models using zero-valued coefficients. Most cases are used on data in which predictors are highly correlated [46].

Bayesian Ridge (BR): with Bayesian, the regularization parameters are used in estimation, while RR is applied to the BR estimator and its coefficients to determine a posteriori estimator. Instead of other models' applications with point estimation, BR works with probability distributors [47].

Least Angle Regression (LARS): A model similar to forward stepwise regression because in datasets with many attributes, at each stage, identify the highest correlated attribute with the target value. If there is more than one variable with the same highest correlation value, average the attributes and proceed to the same angle. Continuously, keep directing the line regression to those mentioned above until it reaches another same or higher correlated variable [48].

Huber Regression (HR): A powerful regression technique with data including outliers. A difference from the least squares loss function is used with HR. With small residuals,

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

penalties are the same with the least squares loss function, but with large residuals, HR's penalty is lower and increases linearly instead of increasing quadratically [49].

Ridge Regression (RR): A linear regression model in which the coefficients are not estimated by least squares (OLS) but by a RR estimator. Which is biased, and its variance is lower than the OLS estimator. The RR method is usually applied when multicollinearity problems occur, reducing the standard error [50].

3.6.2 Tree-based models

With the different constructed approaches, Tree-based models use rules of conditional statements on training data to generate predictions [51]. These models, deeply rooted in statistical decision theory, exemplify the fusion of statistical methods and ML [44]. The Tree-based models and Decision Trees (DT) structure starts with a node and, in the next level, split into branches, concluding that decisions are the final leaves. In addition, Ensemble methods, and Bagging, in which different Tree-based models are trained simultaneously and individually, and predictions are decided by voting or averages of the individual predictions [52]. In the next stage, boosting method based on Tree-based models uses a different method, which in contrast with bagging, creates and trains each selected model one by one and the next with the selected train dataset and mistakes from the previous model [53].

Gradient Boosting Machine (GBM): An ensemble technique, Boosting, uses DT for weak learners. Sequentially, train the weak learners and fit the negative gradient of the given loss function. Across this methodology, GBM Regressor performs excellently in finding a nonlinear relationship between features and the target variable [54, 55].

Random Forest (RF): An ensemble method, called bagging, using DT and techniques of bootstrap and aggregation. Works by combining several DT as base learning models to determine the final output. [56].

Light Gradient Boosting Machine (LGBM): A GBM technique constructed by DTs. Unlike other boosting methods, LGBM splits/grows the tree leaf-wise(horizontally). It uses two more techniques, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), resulting faster training process, lower memory usage and, at most times, higher accuracy [57].

Extra (Randomized) Trees (ET): An ensemble method related to bagging and RF that creates several DT and, without replacement, samples them randomized, resulting in unique datasets for each constructed tree. The difference between the ET method is that random select the split value of each feature [58].

Decision Tree (DT): A tree structure regressor is one of the most used techniques in ML. It is constructed with the root node, representing the whole dataset, and splits into other nodes (interior). Each interior node represents the dataset's features, and each branch represents the decision rules. Finally, each leaf node includes the outcome of the model. It is worth mentioning that DT is incrementally produced. In supervised learning, DT is the base of the creation of many ensemble ML models, and their application outperforms linear-based models on nonlinear datasets [59].

AdaBoost (AB): A competitive ensemble method, boosting model. Across his application, fits the primary regressor on the original dataset and, based on the prediction error, creates copies of the regressor on the same data but adjusts the weights of the instances based on the first error [60].

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

3.6.3 Non-parametric model

The Non-Parametric method works opposite to all the other Parametric methods, in which the assumptions are made on best-fitting training data. The model does not make primary assumptions transforming the based function, keeping it unidentified [61]. This approach demonstrates the statistical modeling principle of allowing data to guide model structure, providing flexibility in handling complex datasets [44].

K- Neighbors (KNN): Considered a non-parametric method that includes associations' calculations between the target variable and features by averaging each observation in the same "neighbourhood". In cases, neighbourhood size is set across cross-validation to minimize the mean-squared error [62].

3.6.4 Online-learning model

Online-Learning can be considered a fundamentally different approach in which the models are not trained at once. At the same time, data come in sequential order, and the best predictor is updated in each learning step [63, 64].

Passive Aggressive Regressor PA: A not commonly used model, which implementation is used for cases in which data is streaming continuously. It works by sequentially feeding individual data or mini batches, while its loss function is similar to a standard hinge loss function [65].

3.6.5 Performance forecasting optimization

The advantages of BPPF are plenty in various industries. The main target always is to predict as accurately as possible each player's potential performance for long or short-term decision-making. With the following in this research approach, there is the opportunity to investigate and evaluate methods for yielding forecasting results [66]. In basketball, formatted KPIs give an overview and evaluate a player's performance. Across this research, except for investigating different ML models forecasting performance in 90 players' case studies, an advanced formatted approach has been created and evaluated for KPIs.

The targeted KPI that has been evaluated with different forecasting approaches is the FP. As mentioned before, FP is a KPI constructed by advanced basketball statistics. The base methodology, in which FP is forecasted as a formatted formula. While alternatively, each statistic that contributes to the formula is forecasted independently, and the KPI is formatted afterwards with the predictions. With this method, we investigate if there is a possibility to yield KPIs forecasting results by comparing the prediction results of ten ML models with the average MAE and MAPE results of 90 players' case studies.

4 Findings

Acquire highly accurate BPPF implements to extend the research in the state-of-the-art ML models. Open SA data is always advantageous in experiments with different DM and ML approaches [67]. At this point of research, the selected forecasting approach allows us to find the possible trends of the player's performance in his latest 10 matches in which he participated and forecast his upcoming performance. The results will provide an overview of how each ML model described above performs with the selected approach to forecasting the

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

upcoming player's performance. In this research, not only 14 ML models are compared, but 18 key performance metrics are forecasted in each of the 90 different player case studies.

In the first stage, based on the key performance metric's usefulness in overviewing a player's performance, an evaluation ranking metric is constructed with selected weights for each metric's results of MAPE. Based on that, an ML models ranking table is built, overviewing how well each ML model approach can forecast a player's performance via different metrics. In the second stage, we will compare previous results with generated new ones with an experiment following a different forecasting approach. We focused on one basketball player's performance KPI, FP, formulated from basic basketball key performance metrics. In this experiment, based on the ranking table, ten of the 14 models have been selected to be tested and compared on 90 players' case studies and evaluate results averaging the models' evaluation metric, Mean Absolute Percentage Error (MAPE) [68].

4.1 Results scope

This research is based on 90 different high-performance NBA players' case studies. In each case, the dataset follows the same split, 70% for training, 20% for testing and 10% for validation with Unseen data. Data refers to the three latest seasons, including season 2021–22, and unseen has been considered the 18 latest matches of season 2021–22, excluding playoffs. The predicted target key performance basketball metrics are the following: AST/TO, BLK, EFF, FOUR FACTORS, GMSC, NBA, FP, NETRTG, PIE, PM, PTS, REB, STL, TENDEX, TOV, USG%. In addition, the ML models that are tested: are ET Regressor, RF Regressor, DT Regressor, LARS, LASSO Regression, EN, GBM Regressor, KNN Regressor, LGBM, BR, AB Regressor, RR, Passive Aggressive (PA) Regressor, HR Regressor. The selected evaluation metric for each model is MAPE and Mean Absolute Error (MAE) [68].

4.2 Machine learning models ranking score

The purpose of this section is to compare the overall performance scores of ML models in SA, and especially in basketball, WAPE, constructed as a formula (1) of weighted forecasted basketball performance metrics. The weights of each are selected based on the significance of each statistic in evaluating individual basketball players' performance overall for each played match. Each KPI or advanced metric that overviews player performance for one aspect contributes to the formula with a 4% share, statistics and KPIs that overview the domain impact of each player's performance contributes to the formula with a 5% share, and KPIs that evaluate overall players' performance, with considering his attacking, defence, teamwork, and contribution to the win are weighted with 7%.

The WAPE formula (1) is given below:

$$\begin{aligned} WAPE = & 0.04 \times (AST_{MAPE} + AST\_RATIO_{MAPE} + AST/TO_{MAPE} + BLK_{MAPE} \\ & + STL_{MAPE} + TOV_{MAPE}) + 0.05 \times (EFG\%_{MAPE} + PTS_{MAPE} + REB_{MAPE} \\ & + USG\%_{MAPE}) + 0.07 \times (EFF_{MAPE} + 4FACTORS_{MAPE} + GMSC_{MAPE} \\ & + FP_{MAPE} + NETRATING_{MAPE} + PIE_{MAPE} + PLUS/MINUS_{MAPE} \\ & + TENDEX_{MAPE}) \end{aligned} \tag{1}$$

An extended presentation of the occurred results of both test and validation procedures is conducted in Appendices' Tables 7, 8, 9, 10, 11, 12, 13,14, 15, 16, 17, 18 and 19 evaluating

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

each ML model corresponding forecasting performance with MAE, MAPE, Root Squared Error (MSE), Root Mean Squared Error (RMSE) [69] and Root Mean Squared Logarithmic Error (RMSLE) [70].

4.2.1 Cross-validation strategy

The cross-validation strategy followed is the k-fold, with 10 folds in each ML trained/tested, and Appendix Table 5 is an overview of the average MAPE results. Table 2 is an overview of the average MAPE results of only higher performance models, both on the test data for each aforesaid key performance statistic and ML model, with the bold values indicating the lowest MAPE averages and WAPE metrics.

Comparing results on test procedure for individual metrics using BPPF (RQ1) and overall results based on WAPE (RQ2) we conclude that these are similar. However, the three most efficient models were ET with WAPE 35,83%, dominating in three target variables (NETRTG, TOV and USG%). DT, with a WAPE of 35.87%, performing best in five forecasted statistics and KPIs (AST, AST RATIO, 4Factors, REB, STL and TENDEX) and RF, with a WAPE of 35.92% and was more efficient in three players' performance evaluation metrics (FP, PM and PTS). The other five ML-type models of Table 2 scored closer to each other, with LASSO coming first, forecasting better four metrics AST/TO, EFF, EFG% and GMSC). According to Table 5 in Appendices, in the last place on the WAPE table in testing came the HR with

Table 2 WAPE Ranking Table in Testing (Top 8)

Test WAPE	ET	DT (%)	RF (%)	LASSO (%)	LARS (%)	EN (%)	GBM (%)	KNN (%)
AST	50,55	**50,53**	50,79	51,31	51,31	51,53	52,27	52,10
AST RATIO	43,75	**42,91**	44,31	44,56	44,81	45,18	45,47	45,23
AST/TO	68,72	71,69	68,99	**66,15**	66,20	66,94	69,17	67,41
BLK	73,26	72,39	73,92	73,13	73,08	72,97	**68,36**	73,59
EFF	34,97	35,77	34,82	**34,80**	34,82	34,85	35,31	35,38
EFG%	32,05	32,29	32,16	**31,99**	32,01	32,21	32,37	32,47
FOUR_FACTORS	31,88	**29,95**	32,09	32,17	32,19	32,40	32,33	32,01
GMSC	31,11	31,50	31,13	**31,00**	31,05	31,16	31,84	31,70
FP	37,39	37,65	**37,27**	37,58	37,53	37,74	38,42	38,20
NETRTG	**20,05**	20,34	20,18	20,14	20,19	20,32	20,25	20,34
PIE	13,17	13,58	13,24	13,18	**13,16**	13,24	13,34	13,39
PM	23,25	23,90	**23,22**	23,37	23,38	23,46	23,39	23,58
PTS	48,27	48,07	**47,95**	48,02	47,96	48,04	48,94	49,08
REB	51,12	**50,38**	51,21	51,86	51,89	52,21	51,65	53,35
STL	47,03	**44,31**	47,09	53,03	53,06	52,73	51,65	53,10
TENDEX	24,49	**24,06**	24,44	24,51	24,53	24,62	24,83	24,63
TOV	**43,14**	44,98	43,18	46,10	46,09	46,15	47,86	46,54
USG%	**21,16**	21,72	21,45	21,71	21,67	21,53	22,04	21,79
Ranking Score	**35,83**	35,87	35,92	36,22%	36,24	36,36	36,52	36,70
Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

a WAPE of 41,47%. Furthermore, the last seven placed ML types of models and EN could not, on average, outperform all rest in any of the players' performance evaluation statistics.

4.2.2 Forecasting on unseen data

Additionally, for validation purposes for our models' performance, the trained ML models in each player's case study and performance metric are challenged with unseen data (RQ1). Table 5 in Appendices is an overview of the average MAPE results, and Table 3 is an overview of the average MAPE results of only higher performance models, both on the unseen data for each aforesaid key performance statistic and ML model, with bolded values highlighting the lowest MAPE averages and WAPE metrics.

Compared with testing, again, the top three performing models (RQ2) were the ET, RF and DT. ET scored higher with WAPE 34,14%, performing best in four forecasted metrics (AST RATIO, PIE, PM and TOV). In the second place, the RF model evaluated WAPE at 34.23% outperforming others in three targets (AST, PTS and USG%), and the DT came third with WAPE at 34,41%, being more efficient in REB, STL and TENDEX forecasting. According to Table 5 in Appendices, HR was the least efficient ML model on forecasting in SA with the selected approach with WAPE 43.04% and the same pattern with testing results

Table 3 WAPE Ranking Table in Validation (Top 8)

WAPE in Validation Procedure	ET (%)	RF (%)	DT (%)	LARS (%)	LASSO (%)	EN (%)	GBM	KNN (%)
AST	48,89	**48,75**	48,76	50,47	50,44	50,74	50,84	51,01
AST RATIO	**45,52**	45,64	45,80	46,48	46,55	46,91	47,05	47,42
AST/TO	65,04	65,88	65,06	63,72	**63,60**	64,05	67,59	65,85
BLK	72,42	73,04	73,96	71,65	71,65	71,18	**66,17**	71,88
EFF	32,29	32,56	32,56	32,11	**32,09**	32,41	32,81	32,86
EFG%	29,42	29,53	29,83%	29,36	**29,32**	29,64	30,03	30,04
FOUR_FACTORS	25,96	25,88	26,29	**25,75**	25,76	25,89	26,73	26,46
GMSC	28,99	29,14	29,81	28,86	**28,86**	29,32	29,55	29,30
FP	33,74	33,86	33,89	**33,39**	33,47	33,40	33,96	33,75
NETRTG	20,37	20,38	20,69	**20,29**	20,32	20,54	20,44	21,01
PIE	**12,30**	12,34	12,55	12,36	12,35	12,36	12,41	12,59
PM	**24,21**	24,25	24,65	24,23	24,25	24,47	24,34	24,72
PTS	43,05	**42,62**	44,09	42,69	42,68	43,00	43,11	43,46
REB	49,57	49,93%	**49,40**	50,09	50,12	49,96	50,39	50,61
STL	45,48	45,56	**44,64**	52,35	52,36	52,56	51,24	52,35
TENDEX	24,06	24,04	**23,46**	23,88	23,93	24,09	24,43	24,32
TOV	**42,71**	43,16	44,22	46,64	46,63	46,91	47,40	47,67
USG%	22,02	**21,37**	21,47	22,16	22,16	22,27	22,32	22,11
Ranking Score	**34,14**	34,23	34,41	34,53	34,54	34,71	34,83	35,11
Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

occurred, in which the bottom seven and EN did not manage to be more efficient in any of the forecasted players' performance evaluation metrics.

4.2.3 Individual results per player's performance evaluation metric

Based on the results of experiments in the validation procedure (RQ1), only six of the 14 models are qualified as best for predicting advanced individual statistics in basketball (based on MAPE). Besides scoring higher on WAPE, ET made better predictions than other models, for AST RATIO at 45.52%, for PIE at 12.3%, for PM at 24.21% and TOV at 42.71%. The RF came second in the ranking, predicting better the AST with 48.75%, the PTS scored with 42.62% and the USG% with 21.37%. In third place, DT outperforms the others in REB with 49.4%, STL with 44.64%, and TENDEX with 23.46%. LARS had superior performance in predicting FOUR FACTORS with 25.75%, FP with 33.39% and NETRTG with 20.29%. Also, LASSO achieved better in AST/TO at 63.6%, EFF at 32.09%, EFG% at 29.32% and GMSC at 28.86%. Concluding, GBM exceeds other models' performance only in BLK (RQ1).

4.3 KPIs forecasting optimization

The selected basketball KPI that extends experiments is the FP. Table 4 depicts the results of the forecasted FP as the primary target variable and the forecasted FP as a constructed formula (2) with the predictions of the metrics built with (PTS, REB, AST, STL, BLK and TOV). The lowest error values for MAPE are highlighted in bold.

The FP formula (2) is given as follows:

$$\mathrm{FP} \ = P + 1.2 \times \mathrm{REB} + 1.5 \times \mathrm{AST} + 3 \times \mathrm{STL} + 3 \times \mathrm{BLK}-\mathrm{TOV} \tag{2}$$

In the second experiment, the metrics, as mentioned earlier, were forecasted individually, and the FP KPI was created based on the predictions.

FP was forecasted as a prior-constructed formula in the first approach, and all ML models' performances were close. The best performing model (RQ3) was the LARS and EN with MAPE on forecasting with Unseen data, 33.39% and 33.4%, respectively. All ML models' performances were close in the second approach except for GBM. With a difference of more

Table 4 Reformatted KPI Models Forecasting Performance Comparison

MAPE in Validation Procedure	Reformatted FP (%)	FP (%)
GBM	**29,81**	33,96
AB	32,58	34,60
LGBM	32,62	34,68
LARS	32,92	**33,39**
LASSO	32,95	33,47
RF	33,10	33,86
KNN	33,21	33,75
EN	33,21	33,40
ET	33,27	33,74
BR	33,81	34,49

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

than 3,5%, GBM, with the second forecasting approach, outperformed LARS, with WAPE at 29.81%.

Across this experiment, each forecasted player's performance metric results were rounded as integers to be constructed FP KPI afterwards. Overviewing Forecasting of Reformatted FP for the other models, based on the validation procedure, MAPE results, AB scored 32.58%, the LGBM 32.62%, the LARS 32.92%, the LASSO 32.95%, the RF 33.1%, the K Neighbours and the EN 33.21%, the ET 33.27% and the BR 33.81%. It is worth mentioning that excluding GBM, whose MAPE result was better by 2.77% from second in FP MAPE ranking AB, other models' MAPE results did not vary significantly, with max the max deviation being 1.23%. Additionally, each model's performance yield respectively; GBM by 4.15%, AB by 2.02%, LGBM by 2.06%, LARS by 0.47%, LASSO by 0.52%, RF by 0.76%, KNN by 0.54%, EN by 0.19% and ET by 0.47%. The only case in which performance results did not improve was BR, which worsened by 0.68%. Concluding, each of the individual models' results turned out to be better than the forecasted FP as an already structured KPI (RQ3).

5 Discussion

In this research, we identified, with a variety of forecasting experiments, using 90 players' case studies, which type of model of the 14 selected, resulting in better predictions for 18 different players' performance metrics. Each player's case study was forecasted individually for each of the selected players' performance metrics with the selected ML types of models. This approach was selected to give us a clear overview of the potential performance of ML models in SA. With 1, 3, 5, 7, and 10 game-lag features, this research proves that it is possible to predict each player's next game performance, even if luck is a crucial volatile factor that affects player performance.

5.1 Models performance

Based on the test and validation results, a strong standpoint occurs that each of the Linear, Tree, Non-Parametric and Online-Learning based models, perform differently in SA (RQ1). Using a metric named WAPE, we identified each model's overall performance (RQ2). We extended our research on how it is possible with a different approach to forecast as well as possible an SA KPI (RQ3).

5.1.1 Linear models

The models that belong to this category and are tested are LASSO, EN, BR, LARS HR and RR. Based on Table 5 in the Appendix, the best performers are the Least Eagle in fourth place, LASSO in fifth place and EN in seventh place in the WAPE Ranking table. This research shows that linear-based methods can perform well in predicting each player's upcoming performance; however, since data are not linearly separable, because a player's performance depends on imponderables [71], linear-based models can perform well, but it is not the best methodology.

In Table 13 in Appendices, LASSO in Table 12 in Appendices, reached place four in the ranking table with WAPE 34.53%, predicting best in the average of three players' performance evaluation metrics. LASSO penalized fewer essential features [72], which were crucial in his performance. Each player's corresponding final dataset contained 398 features;

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

even if three different feature selection methods were applied, LASSO, with penalizing less important final features, achieved forecasting best AST/TO, EFF, EFG% and GMSC. EN**,** in Table 11 in Appendices, achieved being ranked in the top six of the WAPE ranking table, with 34.71%; however, EN, constructed as a hybrid of LASSO and RR using zero-valued coefficients, performed worse than LASSO but better than RR. Even if multiple features were highly correlated [73], EN did not manage to outperform other models in any forecasted metric but performed well overall. BR, in Table 10 of the Appendix, as RR is applied to Bayesian probability distribution applications [74], did not manage to forecast each player's performance metric better than other models overall.

However, even if he ranked tenth, its WAPE score was closer to winners with 35.42%. LARS, in Table 9 in Appendices, achieved place four in Ranking Table, with WAPE 34.52%, was the best Linear based model, predicting better than other essential metrics and KPIs which evaluate player's overall performance, the FOUR FACTORS, the FP and the NETRTG. LARS is specialized for feature selection across cross-validation, and its often better performing than other models when the number of features is more than the data instances [75], a phenomenon which, with the followed approach in many players' cases, occurred limited data instances (matches played). HR in Table 16 of the Appendix overviewed the worst of the selected type of models, ranked in last place, with WAPE 41.46%. Occurred that is not the best option with the followed approach even if his method is the same with LARS with small residuals; with large residuals, the penalty applied in features is linearly increasing and not quadratically [76], to be able to fit better in those cases. In Table 18 of the Appendix, RR in Table 17 of Appendix managed to be ranked in twelve places with a WAPE of 40.68%, and worse by more than 5% WAPE from AB in eleventh place. RR does not perform feature selection and reduce standard errors in its applications [77], concluding that it cannot predict better than other models of any of the forecasted players' performance metrics.

5.1.2 Tree based models

GBM, RF, LGBM, ET, DTs and AB are the selected and tested Tree-based models in this research. Based on the results, Tree-based models are the best for predicting players' upcoming performance, with ET being in rank one in the WAPE table, RF in second place, Dissection Tree in third, GBM in seventh, LGBM in ninth, and AB in eleventh place. The performance of Tree-based models reasoning that data of the followed experiments approach were non-linearly separable, with complex, non-linear relationships, and many features compared with the data instances [78].

GBM, in Table 7in Appendices, managed to finish in the top seven best performers, with a WAPE of 34.83%. Additionally, it outperforms others in predicting players' BLK, a performance metric that is difficult to predict because of the variable's low variance [79]. RF, which performance is presented in Table 5 in Appendix, referred to as a bagging ensemble method, ranked in third place with WAPE 34.23%, predicting best three players' performance metrics. Like ET, bagging methods proved that it is an excellent method, with the selected forecasting approach, by handling unbalanced, non-linear data very well [80]. LGBM ranked in place nine, with WAPE 35.2%, and its' performance is in Table 13 of Appendices. It can be characterised as eligible to forecast players' overall performance, even if it does not outperform the others in any forecasted metric. Usually, LGBM performs less well than other ensemble methods when data instances are limited [81]. ET was the best performing model in this research, which performance results are shown in Table 14 of the Appendix, with 34.14%, achieved perform best on average in four players' performance metrics. The following model proved that bagging ensemble methods are slightly performing better for

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

forecasting in SA, with constructed randomised Trees and unique selected non-linear datasets [82]. DT, a fundamental ML method, performed highly in second place on the WAPE table, with 34.23%. Achieved outperforming other ML models in three players' performance metrics forecasting, presenting its' detailed forecasting performance in Table 18 in Appendices. Additionally, it performs feature selection, and unimported features do not influence outputs [83]. AB accomplished eleventh place, whose performance is presented in Table 7 in the Appendix, with WAPE 35.35%, which does not predict better than any of the selected forecasted players' performance metrics. AB is a boosting ensemble Tree-Based method [84], and even if he does not rank in the top ten of the ranking table, his overall performance is close to other higher-ranked ML models.

5.1.3 Non-parametric

The selected non-parametric model tested in this experiment is K-Nearest Neighbours (KNN), ranked eighth in WAPE Ranking Table. Without making any primary assumptions on model function, the non-parametric model best fits with training data.

KNN, with forecasting performance scores in Table 15 of the Appendix, the Non-Parametric selected method was ranked eighth with WAPE 35.11%. While no assumptions are needed, modifications in tuning parameters to fit each case, KNN, was an excellent choice for non-linear data experiments [85]. However, often KNN underperform with spread observations into feature space.

5.1.4 Online learning

At Online Learning category belongs only to the PA of the selected ML models. Online Learning was a high-expectation ML category because it is used for cases in which models are needed to adapt dynamically to new data patterns [86]. After all, each player's performance is not stable or linearly mutative.

PA finished in penultimate place on the WAPE Ranking table, with 40.77%, and the performance score is shown in Table 19 as Appendix. Did not manage to generate forecasts better than other ML models in any of the selected players' performance metrics. Crucial in the performance of this model is the amount of data instances, and it is specialised in big data [87]. With the selected approach of applying each model in individual players' datasets, data instances were limited, causing PA to underperform.

5.2 WAPE ranking table

Forecasting the upcoming performance of NBA players was always challenging, starting by evaluating individual players' performance metrics and continuing with appropriate forecasting and evaluation. The findings were based on results on both test procedures and forecasts in unseen data. This research produced extended forecasting scenarios in 90 high-performance players' cases, with 14 compared ML model types in 18 players' performance evaluation metrics. Those mentioned above overviewed that it is possible with the chosen approach to generate valid predictions in high-performance NBA players, with eleven of the Fourteen selected ML models' types; ET, RF, DT, LARS, LASSO, EN, GBM, KNN, LGBM, BR and AB. Evaluation results were similar on both test and validation procedures, clarifying

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

that the best performers are the Tree-Based Models; however, most of the selected Linear-Based models and KNN performed great. Nevertheless, the Online Learning model's PA and specific Linear Based RR and HR results were not prominent (RQ2).

5.3 KPIs forecasting optimization

The second phase of this research aimed to yield KPI forecasting results by applying different approaches and discovering efficiencies in individual forecasting. The target KPI, FP, was selected because it overviews players' performance with fundamental players' performance evaluation statistics. Its uses are applied directly by performance evaluation aspects in the sports industry, on teams and organisations and in Betting Industry via Fantasy Tournaments. The mentioned above made FP a significant statistic that its results interest everyone that is occupied with basketball and NBA [88].

Finally, the results in Table 4 assure that all selected models are good performers in forecasting FP with both selected approaches. However, improvements can be applied because, at the professional level, every minor or significant improvement can make a difference in Sports and Betting. The Tree-Based models again outperformed others, with the GBM model, producing exceptional results with significant improvement in forecasting with Unseen data. The aforementioned managed to forecast excellent players' upcoming performances with MAPE 29.81% and rank him in the first place of models for forecasting FP (RQ3).

6 Conclusion & future work

6.1 Conclusion

Players' performance forecasting in each Sport is considered a great challenge, also with significant importance [24]. For this reason, this research aimed to conduct an extended analysis and review of forecasting methods with game-lag features. 14 different ML models were applied for performance forecasting in 18 fundamental advanced basketball statistics on a pool of 90 high-performance NBA players. This research targeted to overview each ML model's performance, a ranking table based on a proposed novel KPI, WAPE, a weighted evaluation metric of the forecasted advanced basketball statistics MAPE results. Additionally, a comprehensive approach was introduced to extend research on KPIs forecasting optimization, providing remarkable insights into which approach can produce better forecasting results.

This research was based on NBA's last 3 seasons, 2019–22 Advanced Box-Scores statistics, selecting the 90 most high-performance NBA players, filtered by historical KPIs average results, GMSC, 4Factors, TENDEX, FP and EFF averages and their participation time. Furthermore, after the appropriate data pre-processing and feature engineering, based on our selected players' pool, 90 forecasting experiments with 14 ML models on 18 advanced basketball statistics were accomplished.

Results showed that most ML models have good performance, but the Tree-Based models are the winners. The best performer was ET, with 34.14% WAPE, being the best predictor in three out of 18 statistics. In second place, RF, with 34.23% WAPE, outperformed others in forecasting four of the 18 statics and in third place, the DT, with 34.41% WAPE, predicting better compared to other ML models, three of the 18 target variables. It is worth mentioning that the Linear-Based model, LASSO, also had exceptional performance with 34.54% WAPE and forecasting the best four of the 18 statistics. Additionally, good results were achieved by

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

LARS with WAPE 34.53%, EN with WAPE 34.71%, GBM with WAPE 34.83% and KNN with WAPE 35.11% (RQ1).

This research successfully classifies ML model performance based on the target variables advanced basketball metrics. Based on the results, we conclude that the best approach to forecasting in basketball is when using multiple ML models selected based on the target basketball statistic of each experiment. Additionally, benchmarking the performance of different fundamental ML models to identify the best-performing models for basketball player performance prediction. The results suggest that the 3 Tree-based regression models outperformed the other models, namely ET, RF, and DT. Moreover, LASSO, LARS, EN, GBM, and KNN were identified as promising selections for performance forecasting (RQ2).

The second stage aimed to improve KPIs FP forecasted results, following a different approach. Instead of forecasting FP as it was transformed, the individual statistics that construct FP are forecasted separately, and the FP KPI was transformed afterwards with the generated predictions. Insights were remarkable, while forecasted results for all ten selected ML models were improved, with GBM outperforming overall performance. In the first experiment, where FP was forecasted as a unit, the best predictor was LARS, with 33.39% MAPE; however, in the second approach followed, GBM was evaluated with 29.81% MAPE, improving forecasting by 3.58%. Finally, this research concludes that it is possible to yield significantly better prediction results for the corresponding target basketball KPI, with a different forecasting approach (RQ3).

6.2 Future work

The aforesaid experiments provided significant results, overviewing BPPF with game-lag features [89]. Based on this work, there is room for improvement in predicting individual players' performance as well as possible, using open data provided by a valid source, NBA itself.

Future work can consider forecasting each player individually, for each selected statistic, with all available models. At the same time, as observed, Tree-Based models have the best overall performance, but in each statistic, different results occurred, with no Tree-Based models being the best. Additionally, sentiment analysis [90–94] results could be generated for individual players or teams, and betting odds and transfer market data [95] can all be included as forecasting features. Moreover, motion capture technologies [96] are already used to provide a variety and volume of data that can be used as an additional source for players' performance in-game, capturing the recent playstyles and tactics of players' last games [97].

Furthermore, Association Rules [97–100] and patterns recognition [101, 102] could be applied to basketball statistics, forecasting results and players as individuals or in teams. Those results can also be used as features on fundamental forecasting and analysis in NBA. Moreover, other forecasting techniques can be applied to many basketball statistics, for example, classification [103] for BLK, STL, TOV and AST that a player achieved, a statistic with minor variation. More techniques can be Timeseries [104] for forecasting or feature engineering and creation.

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

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Funding Open access funding provided by HEAL-Link Greece. Open Access funding, as the publication agreement for License to Publish - Open Access has been completed (Springer Nature).

Data availability Data are available at each of the three stages throughout this research—initially as raw collections, then pre-processed, and finally after feature engineering—all of which are accessible in the following GitHub repository: github.com/gpapageorgiouedu/Evaluating-the-Effectiveness-of-Machine-Learning-Models-for-Performance-Forecasting-in-Basketball.

Declarations

Conflict of interest The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This manuscript is according to the guidelines and comply with the Ethical Standards.

Ethical approval Consent given above. Authors ensuring confidentiality and privacy of the research and the data obtained.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendices

Tables 5 depicts all models' % MAPE results on CV and Validation forecasting procedures, ranked by WAPE. Models validating the forecasting procedure based on MAPE and Ranked with WAPE. Tables 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19 illustrate model performance as averages of 90 players' case studies, with experiments on 18 basketball performance metrics.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 5 Models Performance and Ranking in Validation vs CV procedures

%MAPEs' Average Results on Validation | CV	ET	RF	DT	LARS	LASSO	EN	GBM	KNN	LGBM	BR	AB	RR	PA	HR
AST	49 | 51	49 | 51	49 | 51	51 | 51	50 | 51	51 | 52	51 | 52	51 | 52	50 | 52	51 | 52	52 | 52	61 | 63	63 | 60	63 | 65
AST RATIO	46 | 44	46 | 44	46 | 43	47 | 45	47 | 45	47 | 45	47 | 46	47 | 45	48 | 47	48 | 47	49 | 46	55 | 55	57 | 51	56 | 58
AST TO	65 | 69	66 | 69	65 | 72	64 | 66	64 | 66	64 | 67	68 | 69	66 | 67	68 | 70	67 | 72	67 | 69	80 | 84	80 | 82	83 | 87
BLK	72 | 73	73 | 74	74 | 72	72 | 73	72 | 73	71 | 73	66 | 68	72 | 74	70 | 73	71 | 73	74 | 76	70 | 72	67 | 69	75 | 78
EFF	32 | 35	33 | 35	33 | 36	32 | 35	32 | 35	32 | 35	33 | 35	33 | 35	33 | 35	33 | 37	33 | 35	38 | 40	36 | 38	39 | 40
EFG PCT	29 | 32	30 | 32	30 | 32	29 | 32	29 | 32	30 | 32	30 | 32	30 | 33	30 | 33	31 | 33	30 | 32	36 | 37	36 | 36	36 | 37
FOUR FACTORS	26 | 32	26 | 32	26 | 30	26 | 32	26 | 32	26 | 32	27 | 32	27 | 32	26 | 33	26 | 33	26 | 32	31 | 36	30 | 35	30 | 36
GMSC	29 | 31	29 | 31	30 | 32	29 | 31	29 | 31	29 | 31	30 | 32	29 | 32	30 | 32	30 | 32	30 | 32	34 | 36	35 | 35	34 | 36
FP	34 | 37	34 | 37	34 | 38	33 | 38	34 | 38	33 | 38	34 | 38	34 | 38	35 | 38	35 | 39	35 | 38	39 | 41	38 | 42	40 | 42
NETRTG	20 | 20	20 | 20	21 | 20	20 | 20	20 | 20	21 | 20	20 | 20	21 | 20	21 | 20	21 | 21	21 | 20	24 | 23	24 | 23	24 | 23
PIE	12 | 13	12 | 13	13 | 14	12 | 13	12 | 13	12 | 13	12 | 13	13 | 13	12 | 13	13 | 14	13 | 13	15 | 16	15 | 15	15 | 16
PM	24 | 23	24 | 23	25 | 24	24 | 23	24 | 23	25 | 24	24 | 23	25 | 24	25 | 24	25 | 24	25 | 24	28 | 26	29 | 26	28 | 27
PTS	43 | 48	43 | 48	44 | 48	43 | 48	43 | 48	43 | 48	43 | 49	44 | 49	43 | 48	45 | 50	44 | 48	50 | 54	52 | 56	50 | 55
REB	50 | 51	50 | 51	49 | 50	50 | 52	50 | 52	50 | 52	50 | 52	51 | 53	51 | 52	50 | 52	51 | 53	58 | 59	57 | 56	59 | 60
STL	46 | 47	46 | 47	45 | 44	52 | 53	52 | 53	53 | 53	51 | 52	52 | 53	53 | 53	53 | 53	55 | 54	62 | 63	63 | 62	65 | 67
TENDEX	24 | 25	24 | 24	24 | 24	24 | 25	24 | 25	24 | 25	24 | 25	24 | 25	24 | 25	25 | 26	25 | 25	28 | 29	29 | 29	29 | 29
TOV	43 | 43	43 | 43	44 | 45	47 | 46	47 | 46	47 | 46	47 | 48	48 | 47	47 | 47	47 | 46	49 | 48	64 | 61	62 | 57	66 | 63

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 5 (continued)

%MAPEs' Average Results on Validation | CV	ET	RF	DT	LARS	LASSO	EN	GBM	KNN	LGBM	BR	AB	RR	PA	HR
USG PCT	22 | 21	21 | 21	22 | 22	22 | 22	22 | 22	22 | 22	22 | 22	22 | 22	22 | 22	23 | 23	22 | 22	25 | 25	25 | 24	25 | 25
WAPE	34 | 36	34 | 36	34 | 36	35 | 36	35 | 36	35 | 36	35 | 37	35 | 37	35 | 37	35 | 37	36 | 37	41 | 42	41 | 41	42 | 43
Ranking	1 | 1	2 | 3	3 | 2	4 | 5	5 | 4	6 | 6	7 | 7	8 | 8	9 | 9	10 | 11	11 | 10	12 | 13	13 | 12	14 | 14

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 6 RF Forecasting Performance

RF	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.679	0.508	5.222	0.499	2.160	1.808	0.487	6.048	2.286	0.495
AST RATIO	7.351	0.443	93.966	0.797	9.376	7.534	0.456	97.982	9.424	0.721
AST TO	1.172	0.690	3.264	0.559	1.702	1.278	0.659	3.941	1.802	0.567
BLK	0.586	0.739	0.961	0.472	0.909	0.574	0.730	0.863	0.831	0.453
EFF	6.985	0.348	1.479	0.273	1.203	7.976	0.326	1.621	1.250	0.267
EFG PCT	0.144	0.322	0.035	0.120	0.180	0.136	0.295	0.031	0.170	0.111
REFORMATTED FP	9.330	0.349	140.011	11.680	0.148	10.191	0.331	169.533	12.577	0.155
FOUR FACTORS	0.086	0.321	0.012	0.078	0.108	0.083	0.259	0.011	0.102	0.073
GMSC	5.653	0.311	1.359	0.257	1.150	6.574	0.291	1.512	1.206	0.254
FP	8.645	0.373	119.721	0.362	10.812	8.927	0.339	128.056	11.056	0.336
NETRTG	17.244	0.202	1.186	0.193	1.075	19.190	0.204	1.299	1.119	0.198
PIE	0.050	0.132	0.594	0.133	0.761	0.051	0.123	0.572	0.737	0.126
PM	10.489	0.232	1.661	0.219	1.278	12.288	0.242	1.888	1.355	0.230
PTS	5.521	0.479	48.764	0.466	6.867	5.912	0.426	57.461	7.345	0.432
REB	2.288	0.512	8.833	0.449	2.893	2.217	0.499	8.269	2.771	0.424
STL	0.768	0.471	1.152	0.522	1.050	0.757	0.456	1.157	1.037	0.513
TENDEX	0.713	1.984	0.810	0.212	0.892	0.736	0.240	0.878	0.917	0.206
TOV	1.114	0.432	2.202	0.507	1.436	1.089	0.432	2.140	1.399	0.489
USG PCT	0.040	0.214	0.003	0.041	0.050	0.042	0.214	0.003	0.053	0.042

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 7 AB Forecasting Performance

AB	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.706	0.521	5.291	0.502	2.173	1.856	0.515	6.231	2.317	0.497
AST RATIO	7.498	0.457	97.445	0.814	9.529	7.669	0.485	99.817	9.539	0.730
AST TO	1.184	0.690	3.320	0.565	1.715	1.295	0.670	3.974	1.810	0.575
BLK	0.611	0.757	0.958	0.462	0.906	0.579	0.743	0.816	0.805	0.427
EFF	7.099	0.353	1.531	0.277	1.223	8.272	0.331	1.668	1.267	0.271
EFG PCT	0.145	0.324	0.035	0.121	0.182	0.139	0.301	0.032	0.173	0.113
REFORMATTED FP	9.426	0.350	142.805	11.807	0.152	10.171	0.326	168.737	12.585	0.156
FOUR FACTORS	0.087	0.322	0.013	0.079	0.110	0.084	0.258	0.011	0.103	0.073
GMSC	5.730	0.316	1.407	0.262	1.170	6.249	0.297	1.582	1.232	0.260
FP	8.737	0.376	122.038	0.365	10.913	9.046	0.346	131.306	11.184	0.341
NETRTG	17.473	0.203	1.210	0.195	1.087	18.848	0.208	1.350	1.138	0.202
PIE	0.050	0.134	0.609	0.135	0.771	0.053	0.125	0.584	0.746	0.127
PM	10.634	0.235	1.707	0.221	1.295	11.568	0.245	1.955	1.380	0.234
PTS	5.590	0.481	50.247	0.470	6.959	5.975	0.435	58.541	7.421	0.438
REB	2.348	0.527	9.150	0.456	2.944	2.250	0.514	8.289	2.778	0.430
STL	0.811	0.536	1.178	0.520	1.063	0.797	0.547	1.169	1.038	0.501
TENDEX	0.723	1.982	0.833	0.215	0.904	0.752	0.246	0.906	0.932	0.209
TOV	1.151	0.479	2.218	0.511	1.441	1.140	0.491	2.174	1.416	0.495
USG PCT	0.041	0.220	0.003	0.041	0.051	0.043	0.216	0.003	0.053	0.043

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 8 GBM Forecasting Performance

GBM	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.720	0.523	5.353	0.503	2.187	1.839	0.508	6.112	2.290	0.492
AST RATIO	7.493	0.455	98.329	0.799	9.562	7.754	0.470	103.115	9.665	0.727
AST TO	1.202	0.692	3.338	0.567	1.726	1.320	0.676	4.034	1.833	0.580
BLK	0.680	0.684	1.035	0.466	0.944	0.649	0.662	0.884	0.860	0.446
EFF	7.121	0.353	1.535	0.277	1.225	7.760	0.328	1.699	1.277	0.272
EFG PCT	0.145	0.324	0.035	0.121	0.182	0.138	0.300	0.032	0.173	0.113
REFORMATTED FP	9.536	0.358	146.248	11.953	0.151	10.230	0.298	172.606	12.674	0.148
FOUR FACTORS	0.087	0.323	0.012	0.078	0.109	0.085	0.267	0.011	0.105	0.074
GMSC	5.791	0.318	1.443	0.264	1.182	6.019	0.295	1.587	1.232	0.260
FP	8.872	0.384	125.617	0.369	11.056	9.195	0.340	137.233	11.435	0.346
NETRTG	17.322	0.202	1.206	0.195	1.084	19.573	0.204	1.314	1.124	0.199
PIE	0.050	0.133	0.603	0.134	0.767	0.049	0.124	0.583	0.745	0.127
PM	10.563	0.234	1.689	0.220	1.288	12.086	0.243	1.920	1.365	0.232
PTS	5.627	0.489	50.788	0.475	6.996	6.065	0.431	60.270	7.541	0.445
REB	2.307	0.516	8.875	0.450	2.900	2.234	0.504	8.260	2.768	0.424
STL	0.810	0.517	1.156	0.506	1.050	0.801	0.512	1.164	1.035	0.499
TENDEX	0.726	1.922	0.840	0.215	0.908	0.749	0.244	0.915	0.935	0.210
TOV	1.153	0.479	2.258	0.511	1.457	1.119	0.474	2.154	1.406	0.490
USG PCT	0.041	0.220	0.003	0.042	0.051	0.044	0.223	0.003	0.055	0.044

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 9 LARS Forecasting Performance

LARS	Forecasting in test environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.701	0.513	5.300	0.497	2.172	1.853	0.505	6.219	2.306	0.493
AST RATIO	7.401	0.448	95.614	0.791	9.446	7.657	0.465	102.426	9.607	0.718
AST TO	1.172	0.662	3.229	0.552	1.692	1.288	0.637	4.009	1.815	0.566
BLK	0.613	0.731	0.907	0.437	0.876	0.585	0.717	0.778	0.785	0.412
EFF	6.999	0.348	1.477	0.272	1.201	7.685	0.321	1.658	1.259	0.268
EFG PCT	0.143	0.320	0.034	0.119	0.179	0.135	0.294	0.031	0.169	0.111
REFORMATTED FP	9.489	0.355	143.553	11.833	0.153	10.320	0.329	174.882	12.730	0.158
FOUR FACTORS	0.086	0.322	0.012	0.078	0.108	0.082	0.257	0.011	0.101	0.072
GMSC	5.649	0.311	1.354	0.256	1.147	6.482	0.289	1.546	1.216	0.255
FP	8.677	0.375	119.693	0.362	10.814	9.100	0.334	134.819	11.304	0.341
NETRTG	17.246	0.202	1.186	0.193	1.075	18.768	0.203	1.288	1.114	0.198
PIE	0.049	0.132	0.589	0.132	0.758	0.051	0.124	0.575	0.738	0.126
PM	10.543	0.234	1.679	0.220	1.284	11.971	0.242	1.877	1.352	0.230
PTS	5.496	0.480	48.203	0.465	6.830	5.991	0.427	59.239	7.451	0.438
REB	2.310	0.519	8.906	0.451	2.903	2.234	0.501	8.285	2.771	0.424
STL	0.805	0.531	1.134	0.494	1.037	0.796	0.523	1.141	1.021	0.486
TENDEX	0.717	1.958	0.814	0.212	0.893	0.732	0.239	0.876	0.915	0.205
TOV	1.135	0.461	2.186	0.502	1.431	1.115	0.466	2.127	1.394	0.485
USG PCT	0.040	0.217	0.003	0.041	0.050	0.044	0.222	0.003	0.055	0.044

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 10 BR Forecasting Performance

BR	Forecasting in test environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.710	0.520	5.343	0.500	2.181	1.855	0.512	6.268	2.315	0.497
AST RATIO	7.667	0.470	101.726	0.804	9.753	7.762	0.481	104.154	9.719	0.730
AST TO	1.220	0.718	3.407	0.572	1.733	1.300	0.668	3.975	1.812	0.572
BLK	0.617	0.729	0.914	0.439	0.880	0.580	0.708	0.774	0.782	0.409
EFF	7.404	0.367	1.660	0.289	1.276	7.949	0.334	1.763	1.301	0.278
EFG PCT	0.149	0.332	0.037	0.125	0.188	0.141	0.307	0.033	0.176	0.115
REFORMATTED FP	9.629	0.361	148.172	12.023	0.160	10.397	0.338	175.540	12.815	0.162
FOUR FACTORS	0.091	0.329	0.013	0.082	0.114	0.085	0.264	0.011	0.104	0.074
GMSC	5.915	0.324	1.493	0.269	1.207	6.309	0.300	1.634	1.251	0.264
FP	9.088	0.386	131.574	0.377	11.330	9.287	0.345	138.112	11.472	0.348
NETRTG	18.104	0.211	1.304	0.201	1.126	19.274	0.210	1.377	1.150	0.204
PIE	0.052	0.139	0.657	0.139	0.800	0.051	0.128	0.608	0.759	0.130
PM	11.053	0.243	1.852	0.230	1.351	11.998	0.250	2.024	1.405	0.238
PTS	5.808	0.496	53.933	0.485	7.211	6.248	0.452	62.995	7.711	0.456
REB	2.334	0.522	9.119	0.455	2.935	2.222	0.500	8.240	2.766	0.424
STL	0.813	0.530	1.185	0.502	1.057	0.797	0.525	1.146	1.025	0.489
TENDEX	0.750	1.736	0.888	0.221	0.934	0.753	0.246	0.913	0.935	0.210
TOV	1.141	0.463	2.207	0.506	1.439	1.121	0.471	2.159	1.404	0.489
USG PCT	0.042	0.226	0.003	0.043	0.053	0.045	0.226	0.003	0.056	0.045

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 11 EN Forecasting Performance

EN	Forecasting in test environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.708	0.515	5.352	0.499	2.182	1.856	0.507	6.264	2.314	0.495
AST RATIO	7.458	0.452	96.995	0.796	9.509	7.701	0.469	103.209	9.651	0.722
AST TO	1.176	0.669	3.234	0.555	1.694	1.285	0.640	3.969	1.807	0.567
BLK	0.613	0.730	0.907	0.438	0.877	0.584	0.712	0.780	0.786	0.412
EFF	7.025	0.348	1.495	0.274	1.209	7.766	0.324	1.668	1.265	0.270
EFG PCT	0.144	0.322	0.035	0.120	0.181	0.136	0.296	0.031	0.170	0.112
REFORMATTED FP	9.449	0.352	142.804	11.805	0.151	10.329	0.332	174.082	12.710	0.158
FOUR FACTORS	0.087	0.324	0.013	0.079	0.109	0.082	0.259	0.011	0.102	0.072
GMSC	5.671	0.312	1.375	0.259	1.157	6.474	0.293	1.558	1.221	0.257
FP	8.767	0.377	121.736	0.365	10.905	9.131	0.334	134.936	11.314	0.342
NETRTG	17.397	0.203	1.205	0.194	1.084	18.742	0.205	1.319	1.127	0.199
PIE	0.050	0.132	0.599	0.133	0.764	0.049	0.124	0.580	0.741	0.126
PM	10.612	0.235	1.700	0.221	1.293	11.792	0.245	1.923	1.368	0.232
PTS	5.532	0.480	48.925	0.467	6.876	6.015	0.430	59.560	7.473	0.440
REB	2.331	0.522	9.045	0.453	2.926	2.225	0.500	8.251	2.768	0.424
STL	0.804	0.527	1.137	0.497	1.039	0.796	0.526	1.140	1.022	0.487
TENDEX	0.720	1.959	0.822	0.213	0.898	0.736	0.241	0.882	0.919	0.206
TOV	1.140	0.461	2.209	0.505	1.439	1.119	0.469	2.142	1.399	0.488
USG PCT	0.040	0.215	0.003	0.041	0.050	0.044	0.223	0.003	0.055	0.044

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 12 LASSO Forecasting Performance

LASSO	Forecasting in test environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.700	0.513	5.291	0.497	2.170	1.852	0.504	6.214	2.304	0.493
AST RATIO	7.382	0.446	95.312	0.791	9.431	7.666	0.466	102.636	9.611	0.718
AST TO	1.173	0.661	3.233	0.552	1.693	1.287	0.636	4.003	1.814	0.565
BLK	0.613	0.731	0.903	0.436	0.875	0.585	0.716	0.777	0.785	0.411
EFF	6.997	0.348	1.476	0.272	1.201	7.774	0.321	1.656	1.258	0.268
EFG PCT	0.143	0.320	0.034	0.119	0.179	0.135	0.293	0.031	0.168	0.110
REFORMATTED FP	9.484	0.355	143.505	11.831	0.153	10.332	0.330	175.048	12.734	0.158
FOUR FACTORS	0.086	0.322	0.012	0.078	0.108	0.082	0.258	0.011	0.101	0.072
GMSC	5.642	0.310	1.353	0.256	1.147	6.341	0.289	1.543	1.214	0.255
FP	8.683	0.376	119.623	0.362	10.813	9.104	0.335	134.727	11.307	0.341
NETRTG	17.214	0.201	1.182	0.193	1.073	19.549	0.203	1.292	1.116	0.198
PIE	0.049	0.132	0.589	0.132	0.758	0.051	0.124	0.575	0.738	0.126
PM	10.536	0.234	1.676	0.219	1.283	11.841	0.243	1.878	1.353	0.230
PTS	5.499	0.480	48.201	0.465	6.831	5.987	0.427	59.175	7.446	0.438
REB	2.311	0.519	8.906	0.450	2.902	2.233	0.501	8.278	2.770	0.424
STL	0.805	0.530	1.133	0.494	1.037	0.796	0.524	1.141	1.021	0.486
TENDEX	0.717	1.958	0.812	0.212	0.893	0.734	0.239	0.878	0.916	0.206
TOV	1.135	0.461	2.186	0.502	1.431	1.115	0.466	2.126	1.393	0.485
USG PCT	0.040	0.217	0.003	0.041	0.050	0.045	0.222	0.003	0.055	0.044

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 13 LGBM Forecasting Performance

LGBM	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.704	0.517	5.274	0.499	2.169	1.822	0.502	6.015	2.278	0.488
AST RATIO	7.554	0.467	99.065	0.801	9.631	7.823	0.484	104.674	9.745	0.735
AST TO	1.200	0.700	3.341	0.565	1.719	1.308	0.676	4.031	1.821	0.575
BLK	0.626	0.731	0.938	0.447	0.893	0.592	0.704	0.804	0.798	0.419
EFF	7.111	0.354	1.533	0.278	1.225	7.920	0.331	1.706	1.281	0.274
EFG PCT	0.147	0.327	0.036	0.122	0.184	0.140	0.304	0.033	0.175	0.114
REFORMATTED FP	9.449	0.349	142.535	11.784	0.148	10.089	0.326	165.980	12.487	0.151
FOUR FACTORS	0.088	0.325	0.013	0.080	0.110	0.084	0.262	0.011	0.104	0.074
GMSC	5.777	0.318	1.423	0.263	1.177	6.453	0.296	1.596	1.238	0.262
FP	8.821	0.377	124.358	0.369	11.028	9.242	0.347	136.805	11.429	0.347
NETRTG	17.589	0.204	1.226	0.196	1.092	18.889	0.208	1.362	1.145	0.202
PIE	0.050	0.134	0.605	0.134	0.769	0.053	0.124	0.582	0.742	0.127
PM	10.683	0.235	1.723	0.222	1.301	11.803	0.248	2.009	1.396	0.236
PTS	5.559	0.481	49.472	0.470	6.914	6.032	0.431	59.194	7.466	0.438
REB	2.332	0.522	9.060	0.454	2.928	2.250	0.509	8.468	2.799	0.429
STL	0.810	0.531	1.147	0.501	1.044	0.809	0.532	1.182	1.042	0.500
TENDEX	0.721	1.853	0.829	0.214	0.902	0.745	0.243	0.911	0.931	0.209
TOV	1.140	0.466	2.212	0.506	1.442	1.118	0.470	2.147	1.403	0.489
USG PCT	0.040	0.217	0.003	0.041	0.050	0.044	0.219	0.003	0.054	0.044

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 14 ET Forecasting Performance

ET	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.679	0.506	5.226	0.499	2.163	1.827	0.489	6.161	2.305	0.497
AST RATIO	7.288	0.438	92.363	0.794	9.307	7.582	0.455	98.850	9.470	0.724
AST TO	1.172	0.687	3.265	0.558	1.702	1.279	0.650	3.985	1.811	0.569
BLK	0.589	0.733	0.966	0.473	0.911	0.578	0.724	0.868	0.834	0.455
EFF	7.010	0.350	1.499	0.274	1.209	7.720	0.323	1.645	1.258	0.268
EFG PCT	0.143	0.320	0.034	0.119	0.179	0.135	0.294	0.031	0.169	0.111
REFORMATTED FP	9.325	0.349	140.374	11.699	0.148	10.337	0.333	173.820	12.725	0.159
FOUR FACTORS	0.086	0.319	0.012	0.078	0.108	0.082	0.260	0.011	0.102	0.072
GMSC	5.648	0.311	1.356	0.257	1.149	6.324	0.290	1.532	1.213	0.255
FP	8.709	0.374	120.343	0.362	10.832	9.060	0.337	132.305	11.247	0.341
NETRTG	17.105	0.200	1.170	0.192	1.067	19.868	0.204	1.292	1.116	0.198
PIE	0.049	0.132	0.594	0.133	0.760	0.052	0.123	0.573	0.736	0.126
PM	10.472	0.233	1.654	0.218	1.275	11.495	0.242	1.870	1.350	0.230
PTS	5.525	0.483	48.670	0.466	6.863	6.007	0.431	59.756	7.475	0.438
REB	2.294	0.511	8.846	0.449	2.896	2.217	0.496	8.223	2.763	0.423
STL	0.771	0.470	1.157	0.524	1.053	0.758	0.455	1.156	1.036	0.514
TENDEX	0.714	1.921	0.810	0.212	0.891	0.732	0.241	0.871	0.912	0.205
TOV	1.115	0.431	2.197	0.506	1.434	1.088	0.427	2.136	1.397	0.489
USG PCT	0.039	0.212	0.002	0.040	0.049	0.044	0.220	0.003	0.054	0.043

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 15 KNN Forecasting Performance

KNN	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.730	0.521	5.493	0.502	2.206	1.859	0.510	6.328	2.328	0.497
AST RATIO	7.418	0.452	96.081	0.797	9.509	7.765	0.474	103.616	9.705	0.724
AST TO	1.188	0.674	3.290	0.559	1.705	1.305	0.659	4.057	1.832	0.575
BLK	0.621	0.736	0.922	0.444	0.885	0.587	0.719	0.794	0.793	0.417
EFF	7.094	0.354	1.513	0.276	1.216	7.436	0.329	1.726	1.287	0.275
EFG PCT	0.145	0.325	0.035	0.121	0.182	0.138	0.300	0.032	0.173	0.113
REFORMATTED FP	9.512	0.355	144.958	11.891	0.154	10.440	0.332	176.683	12.818	0.163
FOUR FACTORS	0.087	0.320	0.013	0.079	0.110	0.084	0.265	0.011	0.104	0.074
GMSC	5.753	0.317	1.407	0.262	1.170	5.684	0.293	1.582	1.231	0.259
FP	8.848	0.382	124.278	0.368	11.012	9.208	0.337	138.049	11.441	0.345
NETRTG	17.430	0.203	1.208	0.195	1.085	19.004	0.210	1.373	1.147	0.203
PIE	0.050	0.134	0.606	0.134	0.769	0.052	0.126	0.592	0.750	0.128
PM	10.644	0.236	1.708	0.221	1.295	11.708	0.247	1.963	1.381	0.234
PTS	5.633	0.491	50.352	0.473	6.974	6.150	0.435	62.658	7.648	0.451
REB	2.357	0.533	9.224	0.458	2.950	2.249	0.506	8.427	2.792	0.428
STL	0.809	0.531	1.152	0.502	1.046	0.803	0.524	1.159	1.032	0.495
TENDEX	0.724	2.187	0.830	0.214	0.903	0.743	0.243	0.899	0.927	0.209
TOV	1.145	0.465	2.231	0.509	1.445	1.135	0.477	2.170	1.413	0.495
USG PCT	0.041	0.218	0.003	0.042	0.051	0.045	0.221	0.003	0.056	0.045

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 16 HR Forecasting Performance

HR	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	2.002	0.650	7.202	0.588	2.569	2.074	0.629	7.314	2.564	0.566
AST RATIO	8.921	0.580	157.439	0.903	11.530	8.812	0.560	133.730	11.037	0.828
AST TO	1.471	0.871	4.585	0.658	2.016	1.542	0.826	4.991	2.078	0.659
BLK	0.796	0.778	1.800	0.519	1.159	0.754	0.751	1.537	1.054	0.491
EFF	8.281	0.403	2.098	0.322	1.433	7.958	0.386	2.243	1.469	0.317
EFG PCT	0.172	0.374	0.049	0.142	0.215	0.169	0.360	0.046	0.208	0.136
FOUR FACTORS	0.103	0.358	0.017	0.093	0.128	0.100	0.303	0.016	0.123	0.088
GMSC	6.674	0.362	1.927	0.303	1.370	6.341	0.344	2.067	1.407	0.301
FP	10.162	0.419	163.955	0.417	12.672	10.524	0.397	173.437	12.910	0.398
NETRTG	20.641	0.234	1.663	0.225	1.270	19.286	0.237	1.791	1.315	0.231
PIE	0.059	0.158	0.853	0.158	0.911	0.053	0.151	0.816	0.884	0.152
PM	12.371	0.269	2.338	0.256	1.516	11.932	0.282	2.676	1.610	0.272
PTS	6.460	0.553	67.902	0.535	8.091	6.899	0.499	77.031	8.538	0.512
REB	2.749	0.598	12.428	0.529	3.435	2.666	0.589	11.786	3.323	0.502
STL	0.983	0.666	1.824	0.599	1.303	0.963	0.646	1.805	1.256	0.585
TENDEX	0.862	1.758	1.168	0.252	1.070	0.877	0.285	1.209	1.080	0.245
TOV	1.394	0.634	3.266	0.610	1.755	1.383	0.657	3.214	1.720	0.593
USG PCT	0.048	0.251	0.004	0.049	0.060	0.051	0.251	0.004	0.062	0.050

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 17 RR Forecasting Performance

RR	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.952	0.625	6.816	0.575	2.492	2.046	0.611	7.266	2.547	0.559
AST RATIO	8.655	0.552	129.658	0.885	10.962	8.705	0.553	129.979	10.917	0.813
AST TO	1.465	0.840	4.534	0.651	2.012	1.553	0.803	5.145	2.106	0.653
BLK	0.807	0.717	1.491	0.518	1.106	0.779	0.700	1.387	1.051	0.497
EFF	8.249	0.400	2.067	0.319	1.421	7.697	0.382	2.204	1.456	0.315
EFG PCT	0.170	0.370	0.048	0.140	0.213	0.170	0.362	0.047	0.208	0.137
FOUR FACTORS	0.102	0.355	0.017	0.092	0.127	0.100	0.306	0.016	0.123	0.088
GMSC	6.649	0.359	1.902	0.301	1.360	6.064	0.338	1.997	1.383	0.295
FP	10.096	0.414	162.922	0.413	12.603	10.445	0.391	171.094	12.835	0.392
NETRTG	20.329	0.231	1.614	0.221	1.251	19.027	0.235	1.764	1.304	0.230
PIE	0.060	0.158	0.850	0.158	0.910	0.052	0.150	0.804	0.877	0.151
PM	12.154	0.264	2.247	0.251	1.485	11.814	0.277	2.597	1.586	0.269
PTS	6.415	0.543	67.106	0.530	8.028	6.915	0.498	76.950	8.540	0.511
REB	2.691	0.590	11.945	0.519	3.371	2.648	0.584	11.638	3.304	0.501
STL	0.960	0.633	1.718	0.585	1.267	0.929	0.621	1.543	1.197	0.569
TENDEX	0.856	1.697	1.148	0.250	1.061	0.866	0.282	1.196	1.070	0.242
TOV	1.354	0.607	3.069	0.596	1.702	1.371	0.637	3.158	1.709	0.594
USG PCT	0.047	0.249	0.004	0.048	0.059	0.050	0.249	0.004	0.062	0.050

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 18 DT Forecasting Performance

DT	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	1.667	0.505	5.196	0.502	2.151	1.825	0.488	6.333	2.321	0.503
AST RATIO	7.461	0.429	96.716	0.818	9.506	7.756	0.458	104.700	9.688	0.753
AST TO	1.177	0.717	3.259	0.571	1.699	1.284	0.651	3.999	1.819	0.582
BLK	0.612	0.724	1.012	0.491	0.939	0.587	0.740	0.894	0.857	0.471
EFF	7.072	0.358	1.525	0.278	1.219	7.554	0.326	1.644	1.258	0.268
EFG PCT	0.145	0.323	0.035	0.121	0.182	0.135	0.298	0.031	0.170	0.111
FOUR FACTORS	0.085	0.300	0.012	0.077	0.107	0.082	0.263	0.011	0.102	0.073
GMSC	5.790	0.315	1.444	0.262	1.182	6.003	0.298	1.522	1.211	0.256
FP	8.736	0.377	121.309	0.364	10.896	8.961	0.339	127.209	11.039	0.338
NETRTG	17.454	0.203	1.213	0.195	1.087	19.139	0.207	1.318	1.127	0.201
PIE	0.051	0.136	0.628	0.136	0.780	0.054	0.125	0.579	0.740	0.127
PM	10.768	0.239	1.771	0.224	1.315	12.577	0.247	1.975	1.384	0.231
PTS	5.624	0.481	50.583	0.470	6.999	6.135	0.441	62.008	7.601	0.441
REB	2.312	0.504	8.990	0.446	2.919	2.243	0.494	8.464	2.817	0.429
STL	0.777	0.443	1.184	0.534	1.062	0.753	0.446	1.157	1.038	0.516
TENDEX	0.714	2.149	0.817	0.213	0.895	0.725	0.235	0.856	0.905	0.202
TOV	1.127	0.450	2.248	0.521	1.454	1.088	0.442	2.158	1.402	0.499
USG PCT	0.040	0.217	0.003	0.041	0.050	0.042	0.215	0.003	0.053	0.042

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Table 19 PA Forecasting Performance

PA	Forecasting in Test Environment					Forecasting with Unseen data				
	MAE	MAPE	MSE	RMSE	RMSLE	MAE	MAPE	MSE	RMSE	RMSLE
AST	2.000	0.599	7.259	0.573	2.509	2.104	0.632	7.721	2.563	0.557
AST RATIO	8.448	0.507	125.705	0.875	10.667	8.933	0.565	135.142	10.994	0.839
AST TO	1.387	0.816	4.080	0.632	1.934	1.636	0.804	6.159	2.186	0.677
BLK	0.787	0.692	1.390	0.507	1.060	0.733	0.674	1.191	0.966	0.479
EFF	7.889	0.380	1.887	0.304	1.347	7.789	0.364	2.144	1.423	0.308
EFG PCT	0.170	0.361	0.049	0.139	0.209	0.171	0.362	0.049	0.207	0.135
FOUR FACTORS	0.099	0.345	0.016	0.088	0.122	0.097	0.297	0.015	0.118	0.084
GMSC	6.382	0.349	1.796	0.287	1.305	6.387	0.345	2.115	1.404	0.301
FP	9.791	0.423	150.432	0.405	12.096	10.421	0.382	173.556	12.648	0.386
NETRTG	20.214	0.228	1.587	0.219	1.239	19.134	0.244	1.925	1.347	0.238
PIE	0.057	0.151	0.769	0.149	0.864	0.053	0.152	0.878	0.895	0.152
PM	12.050	0.261	2.214	0.246	1.457	11.454	0.293	2.921	1.650	0.285
PTS	6.457	0.556	67.798	0.523	7.978	6.885	0.520	75.716	8.415	0.515
REB	2.609	0.562	11.668	0.510	3.276	2.733	0.567	12.593	3.332	0.503
STL	0.909	0.620	1.466	0.557	1.156	0.902	0.632	1.404	1.131	0.546
TENDEX	0.856	2.444	1.169	0.246	1.056	0.892	0.290	1.270	1.088	0.244
TOV	1.270	0.565	2.696	0.567	1.588	1.346	0.624	3.118	1.673	0.583
USG PCT	0.045	0.238	0.003	0.046	0.057	0.050	0.247	0.004	0.061	0.050

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

References

1. Bai Z, Bai X (2021) Sports big data: management, analysis, applications, and challenges. Complexity 2021:1–11. https://doi.org/10.1155/2021/6676297
2. Li B, Xu X (2021) Application of artificial intelligence in basketball sport. J f Educ, Health Sport 11(7):54–67. https://doi.org/10.12775/JEHS.2021.11.07.005
3. Watanabe NM, Shapiro S, Drayer J (2021) Big Data and Analytics in Sport Management. J Sport Manag 35(3):197–202. https://doi.org/10.1123/jsm.2021-0067
4. Sarlis V, Tjortjis C (2020) Sports analytics — Evaluation of basketball players and team performance. Inf Syst. https://doi.org/10.1016/j.is.2020.101562
5. Aoki, R. Y. S., Assuncao, R. M., & Vaz de Melo, P. O. S. (2017). Luck is Hard to Beat. In": *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 1367–1376. https://doi.org/10.1145/3097983.3098045
6. Nguyen NH, Nguyen DTA, Ma B, Hu J (2022) The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity. J Inf Telecommun 6(2):217–235. https://doi.org/10.1080/24751839.2021.1977066
7. Morgulev E, Azar OH, Lidor R (2018) Sports analytics and the big-data era. Int J Data Sci Anal 5(4):213–222. https://doi.org/10.1007/s41060-017-0093-7
8. Terner Z, Franks A (2021) Modeling player and team performance in basketball. Annu Rev Stat Appl 8(1):1–23. https://doi.org/10.1146/annurev-statistics-040720-015536
9. Vinué G, Epifanio I (2019) Forecasting basketball players' performance using sparse functional data*. Stat Anal Data Min 12(6):534–547. https://doi.org/10.1002/sam.11436
10. Ahmadalinezhad M, Makrehchi M (2020) Basketball lineup performance prediction using edge-centric multi-view network analysis. Social Netw Anal Min. https://doi.org/10.1007/s13278-020-00677-0
11. Migliorati M (2020) Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. Electr J Appl Stat Anal 13(2):454–473. https://doi.org/10.1285/i20705948v13n2p454
12. Zhang F, Huang Y, Ren W (2021) basketball sports injury prediction model based on the grey theory neural network. J Healthcare Eng. https://doi.org/10.1155/2021/1653093
13. Rangel W, Ugrinowitsch C, Lamas L (2019) Basketball players' versatility: Assessing the diversity of tactical roles. Int J Sports Sci Coach 14(4):552–561. https://doi.org/10.1177/1747954119859683
14. Siemon, D., Ahmad, R., Huttner, J.-P., & Robra-Bissantz, S. (2019). *Predicting the Performance of Basketball Players Using Automated Personality Mining BeDien-Begleitforschung Personennahe Dienstleistungen View project Collaboration with AI View project*. https://www.researchgate.net/publication/327344755
15. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, Lemaçon A, Soucy P, Glubb D, Rostamianfar A, Bolla MK, Wang Q, Tyrer J, Dicks E, Lee A, Wang Z, Allen J, Keeman R, Eilber U, Easton DF (2017) Association analysis identifies 65 new breast cancer risk loci. Nature. https://doi.org/10.1038/nature24284
16. Kiliç Depren S (2019) FARKLI MAKİNE ÖĞRENMESİ ALGORİTMALARININ BASKETBOL OYUNCULARININ ATIŞ PERFORMANSI ÜZERİNDEKİ ETKİNLİĞİ. Spor ve Performans Araştırmaları Dergisi. https://doi.org/10.17155/omuspd.507797
17. Oughali MS, Bahloul M, El Rahman SA (2019) Analysis of nba players and shot prediction using random forest and XGBoost models. Int Conf Comput Inf Sci (ICCIS) 2019:1–5. https://doi.org/10.1109/ICCISci.2019.8716412
18. Cene E, Parim C, Ozkan B (2018) Comparing the performance of basketball players with decision trees and TOPSIS. Data Sci Appl 1(1):21–28
19. Soliman G, El-Nabawy A, Misbah A, Eldawlatly S (2017) Predicting all star player in the national basketball association using random forest. Intell Syst Conf (IntelliSys) 2017:706–713. https://doi.org/10.1109/IntelliSys.2017.8324371
20. Zimmermann A (2016) Basketball predictions in the NCAAB and NBA: Similarities and differences. Stat Anal Data Min 9(5):350–364. https://doi.org/10.1002/sam.11319
21. Evans BA, Roush J, Pitts JD, Hornby A (2018) Evidence of skill and strategy in daily fantasy basketball. J Gambl Stud 34(3):757–771. https://doi.org/10.1007/s10899-018-9766-y
22. South C, Elmore R, Clarage A, Sickorez R, Cao J (2019) A starting point for navigating the world of daily fantasy basketball. Am Stat 73(2):179–185. https://doi.org/10.1080/00031305.2017.1401559
23. Chen WJ, Jhou MJ, Lee TS, Lu CJ (2021) Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. Entropy. https://doi.org/10.3390/e23040477

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

24. Thabtah F, Zhang L, Abdelhamid N (2019) NBA game result prediction using feature analysis and machine learning. Annals Data Sci 6(1):103–116. https://doi.org/10.1007/s40745-018-00189-x
25. Huang M-L, Lin Y-J (2020) Regression tree model for predicting game scores for the golden state warriors in the national basketball association. Symmetry 12(5):835. https://doi.org/10.3390/sym12050835
26. Cheng G, Zhang Z, Kyebambe M, Kimbugwe N (2016) Predicting the outcome of NBA playoffs based on the maximum entropy principle. Entropy 18(12):450. https://doi.org/10.3390/e18120450
27. The National Basketball Association. (2022). *nba.com*.
28. Zhang S, Gomez MÁ, Yi Q, Dong R, Leicht A, Lorenzo A (2020) Modelling the relationship between match outcome and match performances during the 2019 FIBA basketball world cup: a quantile regression analysis. Int J Environ Res Public Health 17(16):5722. https://doi.org/10.3390/ijerph17165722
29. Dehesa R, Vaquera A, Gonçalves B, Mateus N, Gomez-Ruano MÁ, Sampaio J (2019) Key game indicators in NBA players' performance profiles. Kinesiology 51(1):92–101. https://doi.org/10.26582/k.51.1.9
30. Khanmohammadi, R., Saba-Sadiya, S., Esfandiarpour, S., Alhanai, T., & Ghassemi, M. M. (2022). *MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs*.
31. Atkinson AC, Riani M, Corbellini A (2021) The box-cox transformation: review and extensions. Stat Sci. https://doi.org/10.1214/20-STS778
32. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519. https://doi.org/10.1007/s10115-012-0487-8
33. Katrutsa A, Strijov V (2017) Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Syst Appl 76:1–11. https://doi.org/10.1016/j.eswa.2017.01.048
34. Zien, A., Krämer, N., Sonnenburg, S., & Rätsch, G. (2009). *The Feature Importance Ranking Measure* (pp. 694–709). https://doi.org/10.1007/978-3-642-04174-7_45
35. Imaam F, Subasinghe A, Kasthuriarachchi H, Fernando S, Haddela P, Pemadasa N (2021) Moderate automobile accident claim process automation using machine learning. Int Conf Comput Commun Inf (ICCCI) 2021:1–6. https://doi.org/10.1109/ICCCI50826.2021.9457017
36. Ali, M. (2020). *PyCaret: An open source, low-code machine learning library in Python*. https://www.pycaret.org
37. Wang, Z., Sun, D., Jiang, S., & Huang, W. (2022). AChEI-EL:Prediction of acetylcholinesterase inhibitors based on ensemble learning model. In: *2022 7th international conference on big data analytics (ICBDA)*, 96–103. https://doi.org/10.1109/ICBDA55095.2022.9760329
38. Triguero I, García S, Herrera F (2015) Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inf Syst 42(2):245–284. https://doi.org/10.1007/s10115-013-0706-y
39. Kumar, A., Naughton, J., & Patel, J. M. (2015). Learning generalized linear models over normalized data. In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pp 1969–1984. https://doi.org/10.1145/2723372.2723713
40. Ampomah EK, Qin Z, Nyame G (2020) Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information 11(6):332. https://doi.org/10.3390/info11060332
41. Mondal AR, Bhuiyan MAE, Yang F (2020) Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. SN Appl Sci 2(8):1372. https://doi.org/10.1007/s42452-020-03196-x
42. Lobo JL, Del Ser J, Bifet A, Kasabov N (2020) Spiking neural networks and online learning: an overview and perspectives. Neural Netw 121:88–100. https://doi.org/10.1016/j.neunet.2019.09.004
43. Metzler D, Bruce Croft W (2007) Linear feature-based models for information retrieval. Inf Retrieval 10(3):257–274. https://doi.org/10.1007/s10791-006-9019-z
44. James G, Witten D, Hastie T, Tibshirani R (2021) An Introduction to Statistical Learning. Springer, US
45. Ranstam J, Cook JA (2018) LASSO regression. Br J Surg 105(10):1348–1348. https://doi.org/10.1002/bjs.10895
46. Zhang Z, Lai Z, Xu Y, Shao L, Wu J, Xie G-S (2017) Discriminative elastic-net regularized linear regression. IEEE Trans Image Process 26(3):1466–1481. https://doi.org/10.1109/TIP.2017.2651396
47. Efendi, A., & Effrihan. (2017). *A simulation study on Bayesian Ridge regression models for several collinearity levels*. pp 020031. https://doi.org/10.1063/1.5016665
48. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Annals Stat. https://doi.org/10.1214/009053604000000067
49. Sun Q, Zhou W-X, Fan J (2020) Adaptive huber regression. J Am Stat Assoc 115(529):254–265. https://doi.org/10.1080/01621459.2018.1543124

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

50. McDonald GC (2009) Ridge regression. Wiley Interdisciplinary Rev 1(1):93–100. https://doi.org/10.1002/wics.14
51. Chen, H., Zhang, H., Si, S., Li, Y., Boning, D., & Hsieh, C.-J. (2019). Robustness Verification of Tree-based Models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/cd9508fdaa5c1390e9cc329001cf1459-Paper.pdf
52. Sagi O, Rokach L (2018) Ensemble learning: a survey. WIREs Data Min Knowl Discov. https://doi.org/10.1002/widm.1249
53. Papadopoulos S, Azar E, Woon W-L, Kontokosta CE (2018) Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. J Build Perform Simul 11(3):322–332. https://doi.org/10.1080/19401493.2017.1354919
54. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21. https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX
55. Liu J, Huang J, Zhou Y, Li X, Ji S, Xiong H, Dou D (2022) From distributed machine learning to federated learning: a survey. Knowl Inf Syst 64(4):885–917. https://doi.org/10.1007/s10115-022-01664-x
56. Liu, Y., Wang, Y., & Zhang, J. (2012). *New Machine Learning Algorithm: Random Forest* (pp. 246–252). https://doi.org/10.1007/978-3-642-34062-8_32
57. Chakraborty D, Elhegazy H, Elzarka H, Gutierrez L (2020) A novel construction cost prediction model using hybrid natural and light gradient boosting. Adv Eng Inf. https://doi.org/10.1016/J.AEI.2020.101201
58. Alsariera YA, Adeyemo VE, Balogun AO, Alazzawi AK (2020) AI meta-learners and extra-trees algorithm for the detection of phishing websites. IEEE Access 8:142532–142542. https://doi.org/10.1109/ACCESS.2020.3013699
59. Rathore SS, Kumar S (2016) A decision tree regression based approach for the number of software faults prediction. ACM SIGSOFT Softw Eng Notes 41(1):1–6. https://doi.org/10.1145/2853073.2853083
60. Schapire RE (2013) Explaining AdaBoost. Empirical Inference. Springer, Berlin Heidelberg, pp 37–52
61. Son Y, Byun H, Lee J (2016) Nonparametric machine learning models for predicting the credit default swaps: an empirical study. Expert Syst Appl 58:210–220. https://doi.org/10.1016/j.eswa.2016.03.049
62. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46(3):175–185. https://doi.org/10.1080/00031305.1992.10475879
63. Fontenla-Romero, Ó., Guijarro-Berdiñas, B., Martinez-Rego, D., Pérez-Sánchez, B., & Peteiro-Barral, D. (2013). Online Machine Learning. In *Efficiency and Scalability Methods for Computational Intellect* (pp. 27–54). IGI Global. https://doi.org/10.4018/978-1-4666-3942-3.ch002
64. Wu Q, Zhou X, Yan Y, Wu H, Min H (2017) Online transfer learning by leveraging multiple source domains. Knowl Inf Syst 52(3):687–707. https://doi.org/10.1007/s10115-016-1021-1
65. Yin G, Alazzawi FJI, Mironov S, Reegu F, El-Shafay AS, Rahman ML, Nguyen HC (2022) Machine learning method for simulation of adsorption separation: comparisons of model's performance in predicting equilibrium concentrations. Arabian J Chem 15(3):103612
66. Georgievski B, Vrtagic S (2021) Machine learning and the NBA Game. J Phys Educ Sport 21(06):3339–3343
67. Richter C, O'Reilly M, Delahunt E (2021) Machine learning in sports science: challenges and opportunities. Sports Biomech. https://doi.org/10.1080/14763141.2021.1910334
68. de Myttenaere A, Golden B, Le Grand B, Rossi F (2016) Mean Absolute Percentage Error for regression models. Neurocomputing 192:38–48. https://doi.org/10.1016/j.neucom.2015.12.114
69. Schubert A-L, Hagemann D, Voss A, Bergmann K (2017) Evaluating the model fit of diffusion models with the root mean square error of approximation. J Math Psychol 77:29–45. https://doi.org/10.1016/j.jmp.2016.08.004
70. Botchkarev A (2019) A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdiscip J Inf, Knowl Manag 14:045–076. https://doi.org/10.28945/4184
71. Teramoto M, Cross CL, Rieger RH, Maak TG, Willick SE (2018) Predictive validity of national basketball association draft combine on future performance. J Strength Cond Res 32(2):396–408. https://doi.org/10.1519/JSC.0000000000001798
72. Wu TT, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. Annals Appl Stat. https://doi.org/10.1214/07-AOAS147
73. Mamdouh Farghaly H, Shams MY, Abd El-Hafeez T (2023) Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. Knowl Inf Syst 65(6):2595–2617. https://doi.org/10.1007/s10115-023-01851-4
74. Saqib M (2021) Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. Appl Intell 51(5):2703–2713. https://doi.org/10.1007/s10489-020-01942-7

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

75. McCann L, Welsch RE (2007) Robust variable selection using least angle regression and elemental set sampling. Comput Stat Data Anal 52(1):249–257. https://doi.org/10.1016/j.csda.2007.01.012
76. Mangasarian OL, Musicant DR (2000) Robust linear and support vector regression. IEEE Trans Pattern Anal Mach Intell 22(9):950–955. https://doi.org/10.1109/34.877518
77. Marquardt DW, Snee RD (1975) Ridge regression in practice. Am Stat 29(1):3–20. https://doi.org/10.1080/00031305.1975.10479105
78. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1):27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x
79. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobotics. https://doi.org/10.3389/fnbot.2013.00021
80. Ruiz-Gazen, A., & Villa, N. (2008). Storms prediction : Logistic regression vs random forest for unbalanced data. *Case Studies in Business, Industry and Government Statistics*, *1*.
81. Sun J, Li J, Fujita H (2022) Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine. Appl Soft Comput 130:109637. https://doi.org/10.1016/j.asoc.2022.109637
82. Ambesange, S., Vijayalaxmi, A., Sridevi, S., Venkateswaran, & Yashoda, B. S. (2020). Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 827–832. https://doi.org/10.1109/WorldS450073.2020.9210404
83. Hayes T, Usami S, Jacobucci R, McArdle JJ (2015) Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. Psychol Aging 30(4):911–929. https://doi.org/10.1037/pag0000046
84. Sun J, Jia M, Li H (2011) AdaBoost ensemble for financial distress prediction: an empirical comparison with data from Chinese listed companies. Expert Syst Appl 38(8):9305–9312. https://doi.org/10.1016/j.eswa.2011.01.042
85. Kumar, T. (2015). Solution of linear and non linear regression problem by k nearest neighbour approach: by using three sigma rule. In: *2015 IEEE international conference on computational intelligence & communication technology*, pp 197–201. https://doi.org/10.1109/CICT.2015.110
86. Inventado, P. S., & Scupelli, P. (2015). Data-driven design pattern production. In: *Proceedings of the 20th European conference on pattern languages of programs*, pp 1–13. https://doi.org/10.1145/2855321.2855336
87. Von Krannichfeldt L, Wang Y, Hug G (2021) Online ensemble learning for load forecasting. IEEE Trans Power Syst 36(1):545–548. https://doi.org/10.1109/TPWRS.2020.3036230
88. Kajy M, Higginbotham DO, Ball G, Vaidya R (2022) "Fantasy Points" associated with professional athlete performance after lumbar discectomy or microdiscectomy. Spartan Med Res J. https://doi.org/10.51894/001c.30766
89. Lu C-J, Lee T-S, Wang C-C, Chen W-J (2021) Improving sports outcome prediction process using integrating adaptive weighted features and machine learning techniques. Processes 9(9):1563. https://doi.org/10.3390/pr9091563
90. Oikonomou, L., & Tjortjis, C. (2018). A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter. In: *2018 South-Eastern European design automation, computer engineering, computer networks and society media conference (SEEDA_CECNSM)*, pp 1–8. https://doi.org/10.23919/SEEDA-CECNSM.2018.8544919
91. Tsiara E, Tjortjis C (2020) Using twitter to predict chart position for songs. IFIP Adv Inf Commun Technol 583:62–72. https://doi.org/10.1007/978-3-030-49161-1_6/TABLES/2
92. Nousi, C., & Tjortjis, C. (2021). 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. https://doi.org/10.1109/SEEDA-CECNSM53056.2021.9566242
93. Koukaras P, Rousidis D, Tjortjis C (2021) An introduction to information network modeling capabilities, utilizing graphs. Commun Comput Inf Sci 1355:134–140. https://doi.org/10.1007/978-3-030-71903-6_14
94. Beleveslis, D., Tjortjis, C., Psaradelis, D., & Nikoglou, D. (2019). A hybrid method for sentiment analysis of election related tweets. In: *2019 4th South-East Europe design automation, computer engineering, computer networks and social media conference (SEEDA-CECNSM)*, pp 1–6. https://doi.org/10.1109/SEEDA-CECNSM.2019.8908289
95. Alberola JM, Garcia-Fornes A (2013) Using a case-based reasoning approach for trading in sports betting markets. Appl Intell 38(3):465–477. https://doi.org/10.1007/s10489-012-0381-9

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

96. Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., & Zemel, R. (2018). Neural relational inference for interacting systems. In J. Dy & A. Krause (Eds.), In: *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2688–2697). PMLR. https://proceedings.mlr.press/v80/kipf18a.html
97. Gómez M-Á, Medina R, Leicht AS, Zhang S, Vaquera A (2020) The performance evolution of match play styles in the spanish professional basketball league. Appl Sci 10(20):7056. https://doi.org/10.3390/app10207056
98. Tjortjis, C., Sinos, L., & Layzell, P. (2003). Facilitating program comprehension by mining association rules from source code. In: *11th ieee international workshop on program comprehension,* pp 125–132. https://doi.org/10.1109/WPC.2003.1199196
99. Hewko, J., Sullivan, R., Reige, S., & El-Hajj, M. (2019). Data Mining in The NBA: An applied approach. In: *2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*, 426–432. https://doi.org/10.1109/UEMCON47517.2019.8993074
100. Ghafari SM, Tjortjis C (2019) A survey on association rules mining using heuristics. WIREs Data Mining Knowl Discov. https://doi.org/10.1002/widm.1307
101. Yu S, Zeng Y, Pan Y, Chen B (2022) Discovering a cohesive football team through players' attributed collaboration networks. Appl Intell. https://doi.org/10.1007/s10489-022-04199-4
102. Raabe D, Nabben R, Memmert D (2023) Graph representations for the analysis of multi-agent spatiotemporal sports data. Appl Intell 53(4):3783–3803. https://doi.org/10.1007/s10489-022-03631-z
103. Jain, S., & Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. In: *2017 3rd international conference on advances in computing,communication & automation (ICACCA) (Fall)*, pp 1–7. https://doi.org/10.1109/ICACCAF.2017.8344688
104. Rodrigues F, Markou I, Pereira FC (2019) Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach. Inf Fusion 49:120–129. https://doi.org/10.1016/j.inffus.2018.07.007

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Content courtesy of Springer Nature, terms of use apply. Rights reserved.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com