First-Half Basketball Data Analysis

Buğra Akkuş
Middle East Technical University
Ankara, Turkey
e258974@metu.edu.tr

Damla Başarmış
Middle East Technical University
Ankara, Turkey
e251004@metu.edu.tr

Başak Kabaloğlu
Middle East Technical University
Ankara, Turkey
e242725@metu.edu.tr

Abstract: Predicting the outcome of basketball games based on partial information from a game is a difficult task because of the dynamic and non-linear nature of performance during games. This study examines the feasibility of using first-half team statistics to predict game outcomes and how much the performance in early parts of games represents team performance. Using play-by-play and box score data, a fixed data preprocessing and exploratory analysis pipeline were used to derive first half game features involving scoring, rebounding, turnovers, and defensive actions. Several supervised classification models were trained to predict the game winner given first-half information only, and model performance was checked with the help of accuracy and some classification metrics. The results show that it is still difficult to predict the winner just based on first half statistics, the best performing model predictability on winner by first half stats is about 58%. However, even further analysis shows that while first half data is limited in predicting the outcome, it is very informative in capturing team behavioral patterns across halves. Offensive rebounds and steals are strongly consistent across the first and second halves, and teams that lead in these categories early in the game have a chance of maintaining competitive advantages throughout the course of the game. These results indicate that first-half measures are more useful in characterizing team performance tendencies than in very accurate outcome predictions. The study helps emphasize on interpretation of predictive limitations along with meaningful behavioral insights while modeling sports data.

Keywords— Sports Analytics, Basketball, First-Half Statistics, Decision Tree, XGBoost, Classification, Sports Data Science.

I. INTRODUCTION

Basketball games are dynamic and the performance in the first half of the game may hold information about the final performance. In practice, first-half statistics are often used by coaches and analysts to try to determine how well their team is doing and what developments might occur in the games. However, the degree to which information from the early parts of the game can reliably predict results is less clear, especially given the complexity of the adjustments made in-game and changing team strategies. The main aim of this study is to investigate if it is possible to use statistics of first-half team-level to predict the following outcomes:

(1) the final game winner,

(2) the number of offensive rebounds in the game as each team finishes and

(3) which team wins the game with more stealing.

In addition to being able to predict outcomes, the study hopes to assess whether action-based performance metrics, such as rebounds and steals, are more stable and predictable than outcomes based on final scores. To address these objectives, two supervised learning models i.e. Decision Tree and XGBoost are used. The Decision Tree model gives interpretability and understanding about key decision rules and the XGBoost is used to capture the potential non-linear relationship and give better predictions. By comparing predictive accuracy for various outcome types, this study aims to separate between outcomes that are inherently hard to predict early in a game and those that represent more consistent team behaviors.

II. LITERATURE REVIEW

One of the most recognized questions in sports analytics is the prediction of the basketball game results. Past research has been directed towards the use of in-game statistics like shooting efficiency, rebounds, turnovers, and steals to explain or predict the match results [1]. Of these variables, score differential has always been an effective measure of the eventual outcome of the game, particularly at later points in the game [2]. Several studies have stressed that early-game statistics, including first-half performance, offer helpful, though imprecise, information to predict end results [3]. Since basketball games can be competitive at halfway, the difference in first-half scores can be very overlapping in the end-winners and losers, thus making them less predictive [4]. This complicates the prediction of winners using data of initial stages of the game. Conversely, statistics that are associated with actions like offensive rebounds and steals have been identified to be relatively more consistent indicators of team dominance [1]. Forces generate more scoring chances and so offensive rebounds are used to assess the capacity of a given team, and defensive pressure and forced turnovers are reflected in steals [3,5]. Such statistics are more likely to display more distinct division between teams and can thus be simpler to forecast with the help of machine learning models. In terms of the modeling techniques, decision trees are widely adopted because they are easy to interpret and model non-linear associations [6]. Nevertheless, ensemble approaches like XGBoost have been demonstrated to be better than single-tree models because they combine scores of a great number of weak learners and decrease their variation [7]. Consequently, XGBoost is often used on structured sports data where there are complicated interactions between variables. Based on these results, this paper makes a comparative analysis of the results of Decision Tree and XGBoost to predict three outcomes using first half data: the final game winner, the offensive rebound winner, and the steals winner.

III. METHODOLOGY

A) Dataset

The data in this research is the professional basketball match data of the 2020-2021 season. The initial information is presented in two tables: Actions and Players. The Actions table has 352,541 event-level observations comprising of 46 variables, including shots, rebounds and steals, as well as other detailed in-game actions. Players table has 13, 591 player-level observations and 123 variables which are more contextual details about individual players. After preprocessing, the aggregation and filtering stages, the analysis is performed at the match level, and eventually, a final dataset of 259 matches is obtained. Teams level observations of the same match are combined to get team level statistics which can be compared directly between teams. The event-level data feature the identifiers of matches, teams, seasons, and leagues, and the detailed descriptions of in-game actions. Specifically, period variable is applied to differentiate game halves with Periods 1 and 2 representing the first half of the game. Variables which define the type, subtype, success and value of actions are used to derive meaningful performance measure which could be the points scored, offensive rebound, and steals. Based on event-level data in the first half of every match, first-half team-level performance variables are calculated and served as input features in the predictive models. The use of the outcome variables in defining the model performance and the evaluation of model performance is only using the second half and the full-game statistics.

B) Data Preprocessing and Aggregation

B1. Event Filtering and Half Selection

The original data is the event-level data which reports the elaborated in-game action observations during every basketball match. Since the central aim of this work is to measure the predictive capacity of the first half performance, a preliminary filtering process is used with regards to the period of the game. Events of Periods 1 and 2 only are remembered to create features, because during these periods, the first half of the play is taken up. Periods 3 and 4 are not included in the input feature set as they would have captured all predictive information that is relevant to early-game performance. This division is to avoid the information leakage at the later stages of the game in the predictive models. Besides the half-based filtering, event records are also limited to pertinent in-game activities that make a difference to team performance metrics. Action descriptors like action type, action subtype, success indicator and action value are maintained to enable correct computation of scoring, rebounding, and defensive statistics. Aggregation is done by omitting events that do not add contribution to the team level performance measures. This filtering is necessary because all the inputs of the model are formed based on first half events only and second half and full events are to be applied only to define the outcome variables and to measure the consistency of performance. The analysis has methodological validity due to the strict temporal separation between input features and outcomes, as well as realistic prediction conditions in in-game situations.

B2. Team-Level Aggregation

The event-level data are grouped together to form team-level performance statistics at each match after filtering the events to first-half events. This collective action converts single actions in the in-game to summarized actions that describe the performance of a team in the first half. The events of each match are grouped by match identifier and team identifier, and the statistical information is calculated by adding or counting the occurrences of each event. Variables of scoring are determined by summing the numerical data of successful scoring actions whereas action variables like offensive rebounds and steals are derived as the number of respective events. This would make sure that every feature is based on cumulative performance of the team as opposed to single occurrences. The aggregation process creates a one-observation, team-per-match, which allows comparing a competing team to a competing team in the same game. The analysis can be conducted at the team level and hence is more appropriate to capture collective dynamics of performance, as opposed to that provided by the players, especially when the outcomes of interest are those of the game. All the aggregated features are calculated based on first half events only. No inputs are provided at this stage in terms of second-half and full-game statistics with the latter utilized only to determine the outcome variables and the consistency of results between halves. This design maintains the time integrity and makes predictions based on the information that is available at the halftime.

C) Outcome Variable Construction

Three binary outcome variables are created at the match level to assess the predictive value of first-half performance. These results are characterized with the assistance of full-game statistics and become the target variables of the supervised classification models. Final Game Winner is the first outcome variable that defines a team that is a winner in the match. A team that wins is one whose final game points are higher than that of its opponent; otherwise, such a team is considered the non-winner. This is the most frequent goal in basketball outcome prediction, and it is used as a standard of measuring the predictive difficulty when early-game data is available. Along with the final score, to address the features of team performance that could be more stable throughout the halves of the games, two action-based outcome variables are established. The Offensive Rebound Winner variable refers to the number of offensive rebounds that a particular team had more than its opponent in the entire time that the game was played. In the same manner, the Steals Winner variable is used to refer to the fact that a team had more steals than the opponent by the end of the match. The outcome variables are all coded as binary indicators, thus allowing the use of the same set of classification algorithms across all outcome categories. Critically, outcome labels are based on full-game statistics, but all predictive features are built only based on first-half statistics, which implies that there is a distinct temporal gap between predictors and targets.

IV. MODELING

D. Statistical Modeling

D1. Model Specification

In this paper, two algorithm types under supervision are used Decision Tree and XGBoost. The model used is the Decision Tree model because of its interpretability and the capability to display non-linear decision boundaries in the form of hierarchical and rule-based splits. The model can be used to analyze the connection between the performance measures of the first half and the outcomes of prediction, which are clear enough to make decisions by using them in a basketball environment. The XGBoost model is a gradient boosting ensemble approach that is employed to model multifaceted non-linear associations and interactions among predictor variables. XGBoost is highly applicable to structured sporting data and is bias and variance-reducing by sequentially combining many weak learners instead of using a single decision tree. The two models are trained independently on each of the three outcome variables, final game winner, offensive rebound winner, and steals winner with the same set of first-half features. The method allows the comparison of the model performance directly in the outcomes.

D2. Training Procedure

The cumulative data are further subdivided into testing and training data to estimate the out-of-sample predictive performance. Each model is trained on first-half team-level statistics as the input feature, which captures the information during the half-time. The same data splittings and preprocessing pipeline are used across all models and outcome variables, to achieve consistency between experiments. No transformation of outcomes-specific features is presented in addition to the aggregation processes in Section III.

D3. Metrics of Performance Evaluation

Classification accuracy is mainly used as a measure of model performance, which is a measure that is simple to use to intuitively measure the prediction performance and can be easily compared across models and outcome variables. Furthermore, where needed, precision, recall, and F1-score are also provided to address the possible imbalance in different classes and to have more complete information on model behavior. All evaluation measures are calculated on the test data to make certain that the performance reports represent generalization to unseen information.

V. RESULT

E) Model Evaluation & Comparison

E1. Final Game Winner Prediction

The initial result is the capability of the first half team level statistics to forecast the ultimate game winner. Both Decision Tree and XGBoost are trained with the same first half features and tested on the held out test data. In general, there is a weak predictive performance to winner prediction. The most successful model has an accuracy of about 58, it means that the final outcome of a basketball game is always difficult to predict with the help of the information that is known before the half-time. Such degree of accuracy implies that there is a lot of overlap in the level of performance in the first half between the winning and losing teams. The comparisons of models show low differences in their performance. Although XGBoost can predict a non-linear relationship among first half features, it is not significant than the Decision Tree model in this regard. The confusion matrices provided in the analysis show that misclassification is observed in both directions that are dynamic in nature of basketball games, and the impact of the changes in the second half of the game, including the change in tactics and rotating the players. These results illustrate that the performance measures during the earlier part of the game are not enough to predict the eventually final game outcome with a high degree of accuracy, which is a shortcoming of the halftime-based prediction of whether to be the winner or the loser in the game.

E2. Offensive Rebound Winner Prediction

The second group of outcomes aims at predicting the possibility of a team to end the game with a higher number of offensive rebounds than the opponent. Both models have better performance on this action-based outcome, as compared to the final winner prediction. Results of the accuracy test suggest that the first-half offensive rebound data are useful predictors of the overall rebounding dominance in the entire game. Teams that gain more offensive rebounds during a first half have more chances of keeping this lead throughout the game. This trend implies some form of rebounding behavioral persistence between half. The Decision Tree and the XGBoost models are effective to this end with the XGBoost model being slightly higher in most of the assessments. The difference between model performance is, however, small, which suggests that the predictive signal is already represented by rather simple decision rules that are based on first-half rebounding activity. Such findings confirm the hypothesis that offensive rebounding is a consistent element of team performance, and thus it is more predictable using data at the beginning of the game than it is at the end of the game.

Feature	Importance
Offensive Rebound Difference	%59,12
Steals Difference	%15,80
Points Difference	%13,01
Fouls Difference	%12,06

E3. Steals Winner Prediction

The third result analyzed is the team having the most steals than the opponent at the end of the game. Like offensive rebounds, the performance of steal prediction is higher than final winner prediction. The mathematical simulation results indicate that teams with more steal counts during the first half have a higher probability of defensive pressure being maintained during the game. Both models find quite high accuracy on this outcome, which means that defensive activity in the first half is informative of overall in the game steal dominance. Like the offensive rebound outcome, XGBoost shows the increment in results compared to the Decision Tree model, though the difference is not much. This implies that the correlation between early-game steals and final game results is strong enough to model the same using interpretable models. These results indicate that defensive strength, assessed by steals is a stable team trait which can be deduced with certainty based on one-half performance.

Feature	Importance
Steals Difference	%56,78
Points Difference	%15,29
Fouls Difference	%14,44
Offensive Rebound Difference	13,4753

E4. Model Comparison Across Outcomes

When the model performance is compared in the three outcome variables, it is evident that the level of predictability varies. The lowest accuracy is always observed in final game winner prediction whereas action-based outcome, which includes offensive rebound winner and steals winner, is predicted at a significantly greater rate. In all the results, XGBoost outperforms the Decision Tree model by a tiny margin although the differences between them are rather small. This indicates that although the ensemble techniques can be effective to improve performance, the relationship between predictability and model complexity is not the main relationship but the nature of the outcome variable. In general, the findings prove that the first-half statistics are more efficient to reveal the systematic tendencies in the team behavior rather than to forecast the overall game outcomes. Although it is not easy to predict winners based on the halftime, action-based result offers valuable pieces of information about the tendencies in the work of the team which remain constant during the game.

Prediction Target	Decision Tree Accuracy	XGBOOST Accuracy	Best Model
Game Winner	0,585	0,462	Decision Tree
Offensive Rebound Leader	0,738	0,769	XGBoost
Steals Leader	0,708	0,723	XGBoost

- Game Winner: Prediction proved difficult. The best model was the Decision Tree, achieving an accuracy of 58.5%.
- Offensive Rebound Leader: This was highly predictable. The XGBoost model achieved an accuracy of 76.9%.
- Steals Leader: This was also predictable with high accuracy. The XGBoost model achieved 72.3% accuracy.

VI. CONCLUSION

This paper analyzed to what degree, team level statistics on the first half of basketball games might be utilized in forecasting various outcome of the game. The analysis was conducted on three prediction tasks, including the ultimate winner of a game, the dominance of offensive rebound, and the dominance of steal, using supervised models of classification. The findings indicate that it is intrinsically hard to make predictions regarding the ultimate game winner with the knowledge of first-half only. Although interpretable and ensemble-based models are applied, predictive accuracy is still low, and this is because basketball games are dynamic and second-half changes affect the result of the game. These results demonstrate how inaccurate using early-game performance as a predictor of final results is. Conversely, the researcher concludes that performance measures based on action are more accurately forecasted with first-half data and in offensive rebounds and steals. In the case of teams with higher rebounding and defensive action in the first half, the team is likely to retain the benefits in the game. This implies that these metrics measure long-term team behavior and not the temporary changes in performance. Generally, it can be highlighted in the analysis that first-half statistics might not be adequate to achieve very accurate projection of the winner, yet they are useful in gaining insight on the underlying tendencies of the team performance. These results add to a more refined conception of the informational content of early-game basketball data and the significance of predictive findings to be interpreted as analytical outcomes instead of failure to model them.

VII. REFERENCES

[1] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," Journal of Quantitative Analysis in Sports: Vol. 3: Iss. 3, Article 1. DOI: 10.2202/1559-0410.1070

[2] He, Y., Zhou, Z., Qian, C., & Ma, H. (2023). *Predicting score differences in NBA regular season basketball games*. Proceedings of the 5th Asia Pacific Information Technology Conference (APIT 2023), Ho Chi Minh City, Vietnam, 7–11. https://doi.org/10.1145/3588155.3588169

[3] Cervone, D., Bornn, L., & Goldsberry, K. (2014). POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data.

[4] Manner, H. (2016). Modeling and forecasting the outcomes of NBA basketball games. *Journal of Quantitative Analysis in Sports, 12*, 31 - 41.

[5] Matulaitis, K., & Bietkis, T. (2021). Prediction of offensive possession ends in elite basketball teams. International Journal of Environmental Research and Public Health, 18(3), 1083. https://doi.org/10.3390/ijerph18031083

[6] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[7] Ouyang, Y., Li, X., Zhou, W., Hong, W., Zheng, W., Qi, F., & Peng, L. (2024). Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. PLOS ONE, 19(7), e0307478. https://doi.org/10.1371/journal.pone.0307478