https://doi.org/10.5817/StS2024-2-12

Forecasting Stadium Attendance Using Machine Learning Models: A Case of the National Football League

Pang Yu[1], Wang Fengchen[1, 2]

[1]School of Economics and Management, National Research University Higher School of Economics, 3A, Kantemirovskaya Street, St Petersburg, Russia

[2]Université Côte d'Azur, 43, Avenue Stephen Liégeard, Nice, France.

ABSTRACT

The study examines the use of machine learning models to forecast attendance at sports stadiums, specifically analyzing National Football League (NFL) games from 2000 to 2019, with over 5,055 regular-season games. The models, including Linear Regression, Classification and Regression Trees (CART), Random Forest, CatBoost, and XGBoost, integrate a diverse set of variables such as team performance, economic indicators, stadium characteristics, and weather conditions. Each model's accuracy and effectiveness are assessed using five statistical metrics. With a Mean Absolute Error (MAE) of 0.02 and a Root Mean Squared Error (RMSE) of 0.04, the models display high precision in predicting stadium attendance. The coefficient of determination (R^2) reaches 77.27% after optimization. These figures suggest that the models, particularly Random Forest and CatBoost, are highly effective in forecasting attendance rates for NFL games. Key influences on game attendance include factors like 'stadium_name,' 'personal_income,' 'stadium_age,' and 'home_club_age', which emerge as significant predictors. This study fills a theoretical gap in the limited research on the NFL and provides valuable insights for strategic planning and decision-making in professional sports management.

Keywords: machine learning, stadium attendance forecast, Random Forest, CatBoost, XGBoost.

INTRODUCTION

The National Football League (NFL) is a cornerstone of the sports industry in the United States, not only for its cultural significance but also for its substantial economic impact. Each season, NFL games draw millions of fans, playing a crucial role in the league's financial success. Especially, NFL attendance directly influences various revenue streams, including ticket sales, in-stadium purchases, and sponsorships, all of which are closely tied to game attendance. Furthermore, attendance figures significantly impact broadcast rights and advertising revenues, as they reflect the sport's popularity (Buraimo, 2008). However, these attendance numbers can fluctuate due to many factors, such as team performance, economic conditions, and even weather patterns (Ge et al., 2020; Paul et al., 2021). Consequently, understanding and accurately forecasting these factors is vital for financial planning and strategic decision-making within the NFL (Şahin & Uçar, 2020).

Traditional forecasting methods often do not fully capture the complex, multifaceted nature of the factors influencing audience turnout. This inadequacy highlights the need for more sophisticated approaches, particularly at a time when data-driven decision-making is becoming increasingly prevalent in sports management. In this context, machine learning emerges as an effective tool, offering a more advanced method for forecasting NFL attendance. By using large amounts of data and detecting complex trends, machine learning models have the potential to provide more accurate and reliable forecasts than traditional statistical methods (Rein & Memmert, 2016).

The application of machine learning methods to forecast attendance in various sports has been increasingly recognized and adopted in the existing literature. For instance, significant research has been conducted across Major League Soccer (MLS) (King & Rice, 2018) , Major League Baseball (MLB) (Gupta, 2019; Mueller, 2020), the National Basketball Association (NBA) (King, 2017). However, studies focusing on the NFL are limited and typically cover only short-term data (Şahin & Uçar, 2020). This lack of comprehensive research presents a critical opportunity to conduct a more in-depth analysis that spans a longer timeframe and incorporates a broader set of variables in the context of NFL attendance forecasting.

Our study aims to bridge this gap by offering a comprehensive analytical insight by comparing five proven-effective machine learning-based models for NFL game attendance forecasts, including Linear Regression, Classification and Regression Trees, Random Forest, XGBoost, and CatBoost. Also, we compiled a comprehensive dataset that includes over 5,055 NFL regular-season games from 2000 to 2019, incorporating a wide range of predictors such as team performance, economic indicators, stadium conditions, and weather patterns. The accuracy and effectiveness of these models were examined using statistical metrics such as MAE, MAPE, MSE, RMSE, and R^2. Among these models, Random Forest, CatBoost, and XGBoost demonstrated the best performance. This study not only fills the existing theoretical research gap but also offers valuable practical insights for strategic planning and decision-making in professional sports management.

DATA AND METHOD

Data collection and classification

In this study, data were gathered on all National Football League (NFL) games played from 2000 to 2019, including a total of 5,324 games. These games included various stages of competition such as regular season, Wild Card playoffs, Divisional playoffs, Conference Championships, and the Super Bowl. To focus on forecasting stadium attendance with greater precision, the analysis was limited to regular season games. These games typically exhibit higher team engagement and more consistent outcomes. Consequently, all playoff games were removed from the dataset, leaving 5,104 regular season games for analysis. Furthermore, certain regular season games held internationally—at venues including Wembley, Twickenham, and Tottenham in England, as well as Toronto, Canada, and Mexico City, Mexico—were excluded. A total of 38 games were omitted due to their distinctive characteristics and the potential variability in their attendance patterns. Additionally, 11 games that were relocated to alternative stadiums due to extreme weather damage during this period were also excluded to maintain consistency within the dataset. Following these adjustments, the dataset was refined to encompass 5,055 regular-season games. Table 1 outlines the details and descriptive statistics of these games.

Table 1. The descriptive statistics of NFL regular-season matches and attendance

Season	Count	Mean	SD	Season	Count	Mean	SD
2000/2001	248	65934.44	9345.13	2010/2011	253	67036.61	9229.69
2001/2002	248	65753.58	9514.52	2011/2012	254	67418.85	7982.57
2002/2003	256	66325.67	8901.60	2012/2013	254	67632.18	8347.35
2003/2004	255	66649.50	9572.38	2013/2014	253	68397.32	8249.17
2004/2005	256	67462.60	9032.13	2014/2015	252	68650.90	8081.27
2005/2006	247	67947.58	8348.47	2015/2016	253	68212.84	8336.46
2006/2007	256	68773.64	6483.32	2016/2017	252	69324.70	8229.04
2007/2008	255	68692.19	6405.20	2017/2018	251	67167.65	10971.11
2008/2009	254	68206.94	6844.63	2018/2019	253	66887.30	10877.42
2009/2010	254	67505.53	9684.24	2019/2020	251	66433.25	11075.18

Source: authors' elaboration according to Pro Football Reference. URL: https://www.pro-football-reference.com/ (accessed on 25.12.2023).

Building on the classification methodology developed by Borland and Macdonald (2003), we categorized variables into five groups: consumer preferences, economic factors, quality of viewing, attraction of the sporting contest, and supply capacity. This categorization, grounded in sports economics and fan behavior research, ensures a thorough analysis of factors affecting stadium attendance. As presented in Table 2, the data collection included various open sources, including:

- Pro Football Reference (https://www.pro-football-reference.com/) provided detailed match-specific data, including the year, round, weekday, starting hour, home and away team names, teams' winning percentages, previous season performance, stadium names, and game attendance numbers. Acknowledging that home teams are typically the same as their home stadiums, the variable "Stadium_Name" was expressly incorporated into our predictive model. This addition caters to the instances within our dataset where teams have transitioned to different stadiums over the years, ensuring that our analysis captures any influence that a change in stadium might have on game attendance.

- Weather data: Our primary weather data was sourced from Meteostat via their API (https://dev.meteostat.net/python/api/timeseries/), which allowed us to link stadium locations with match day weather conditions, including temperature, precipitation, and atmospheric pressure. For instances where Meteostat lacked historical records, supplementary data were obtained from Weather Underground (https://www.wunderground.com/). This information reflects conditions at the time decisions to attend games are typically made.
- Economic indicators: Data were gathered based on the Metropolitan Statistical Areas (MSAs) corresponding to the home territories of each team, as delineated by Welki & Zlatoper (1999). This data was acquired from the Federal Reserve Economic Data (FRED) (https://fred.stlouisfed.org/). Economic measures, particularly Per Capita Personal Income, were standardized to 2017 dollar values using the Consumer Price Index (CPI) to facilitate consistent economic comparisons across years.
- Team official websites: Data such as the age of clubs, construction year of stadiums, and roof types were sourced from the teams' official websites. Notable stadium features, such as roof type adaptability at So-Fi Stadium and post-2016 improvements at Hard Rock Stadium, were categorized accordingly to reflect their weather protection capabilities. Stadium capacity data, which may fluctuate due to operational changes, were calibrated for consistent comparison.
- Geographic Data: Position coordinates were obtained from Google Maps, and distances from away city centers to home team stadiums were calculated using the Haversine equation.

In addition, the attendance percentage was chosen as the dependent variable and calculated by dividing the actual attendance by the stadium's capacity, as suggested by Falls and Natke (2014), and Bowley and Berger (2014).

Table 2. Description of independent variables

Classification	Variable	Description	Coding/Value	Reference

Consumer preferences	Age of home team clubs
	Age of away team clubs
		Year of team establishment	Min = 0
			Max = 121
			Mean = 54.54
				(Coates & Humphreys, 2007; Depken, 2001)

Economic factors	Real Per Capita Personal Income	Average income in the team's MSA area (U.S. Dollar, 2017 = 100)	Min = 39081.05
			Max = 96225.64
			Mean = 52717.83
				(Depken, 2001; Spenner et al., 2004; Welki & Zlatoper, 1999)

	Population	Resident population of the metropolitan area (in thousands)	Min = 283.34
			Max = 13266.52
			Mean = 3659.76
				(Depken, 2001; Hart et al., 1975; Jennett, 1984; King & Rice, 2018)

	Unemployment Rate	Economic indicator of the team's city (%)	Min = 2.0
			Max = 15.0
			Mean = 5.5
				(Jennett, 1984; Lenten, 2011)

Classification	Variable	Description	Coding/Value	Reference

Quality of viewing	Year	The starting year of NFL season	Min = 2000
			Max = 2019
				(Depken, 2001; Jennett, 1984; Spenner et al., 2004)

	Starting hour	The starting hour of the game	Min = 12
			Max = 23
			Mean = 15.02
				(Alonso & O'Shea, 2013; King & Rice, 2018)

	Weekday	The weekday of the game	Min = 1
			Max = 7
			Mean = 6.46
				(Coates & Humphreys, 2010; King & Rice, 2018; Paul et al., 2021)

	Distance	Kilometers from away team city center to home stadium	Min = 5.91
			Max = 4395.72
			Mean = 1568.45
				(Hart et al., 1975; Şahin & Erol, 2018)

	Temperature	Refers to the temperature several hours prior to the game (Fahrenheit)	Min = -13.0
			Max = 109.04
			Mean = 58.72
				(Gropper & Anderson, 2018; Paul et al., 2021)

	Precipitation	Refers to the rainfall several hours prior to the game (Inch)	Min = 0
			Max = 0.72
				(Ge et al., 2020; Paul et al., 2021)

	Pressure	Refers to atmospheric pressure of mercury (Inch)	Min = 24.07
			Max = 30.85
			Mean = 30.06
				(Paul et al., 2021)

	Roof type	Indoor/ Outdoor/ retractable	Min = 0
			Max = 2
				(Ge et al., 2020; Paul et al., 2021)

	Age of stadium	Years since built	Min = 0
			Max = 96
			Mean = 21.62
				(Depken, 2001; Gropper & Anderson, 2018; Paul et al., 2021; Spenner et al., 2004)

Attraction of sporting contest	Round	Regular season weeks	Min = 1
			Max = 17
				(Falls & Natke, 2014; Mueller, 2020)

	Stadium Name	The code of stadium where the match was held	Min = 0
			Max = 48
				(King & Rice, 2018)

	Home/Away team name	The code of home and away team names	Min = 0
			Max = 33
				(Falls & Natke, 2014; Hansen & Gauthier, 1989; King & Rice, 2018; Mueller, 2020)

	Home winning percentage	The home team winning percentage in the season before the game	Min = 0
			Max = 1
			Mean = 0.462
				(Depken, 2001; Gropper & Anderson, 2018; Paul et al., 2021; Welki & Zlatoper, 1999)

	Away winning percentage	The away team winning percentage in the season prior to the game	Min = 0
			Max = 1
			Mean = 0.477
				(Coates & Humphreys, 2010; Gropper & Anderson, 2018; Mueller, 2020)

	Home/Away Last season play-off teams	Whether home/away team last season is play-off team	0 = No 1 = Yes	(Coates & Humphreys, 2007; Falls & Natke, 2016; Gropper & Anderson, 2018; Nesbit & King, 2010)
	Home/Away Last season Super Bowl teams	Whether home/away team last season is Super Bowl team	0 = No 1 = Yes	(Jennett, 1984; Nesbit & King, 2010)

Supply capacity	Capacity	Maximum attendance capacity	Min = 27000
			Max = 100000
			Mean = 70901.09
				(Depken, 2001; Gropper & Anderson, 2018; Hansen & Gauthier, 1989; King & Rice, 2018)

Source: authors' elaboration.

Variables' selection and validation

We employed the Spearman correlation coefficient to evaluate the association between diverse variable types within our dataset. Prior to assessment, categorical variables such as home team, away team, stadium name, and roof type were numerically encoded to enable their inclusion. This non-parametric measure was chosen for its capacity to handle mixed data types effectively, without the requirement of a normal distribution (Hauke & Kossowski, 2011).

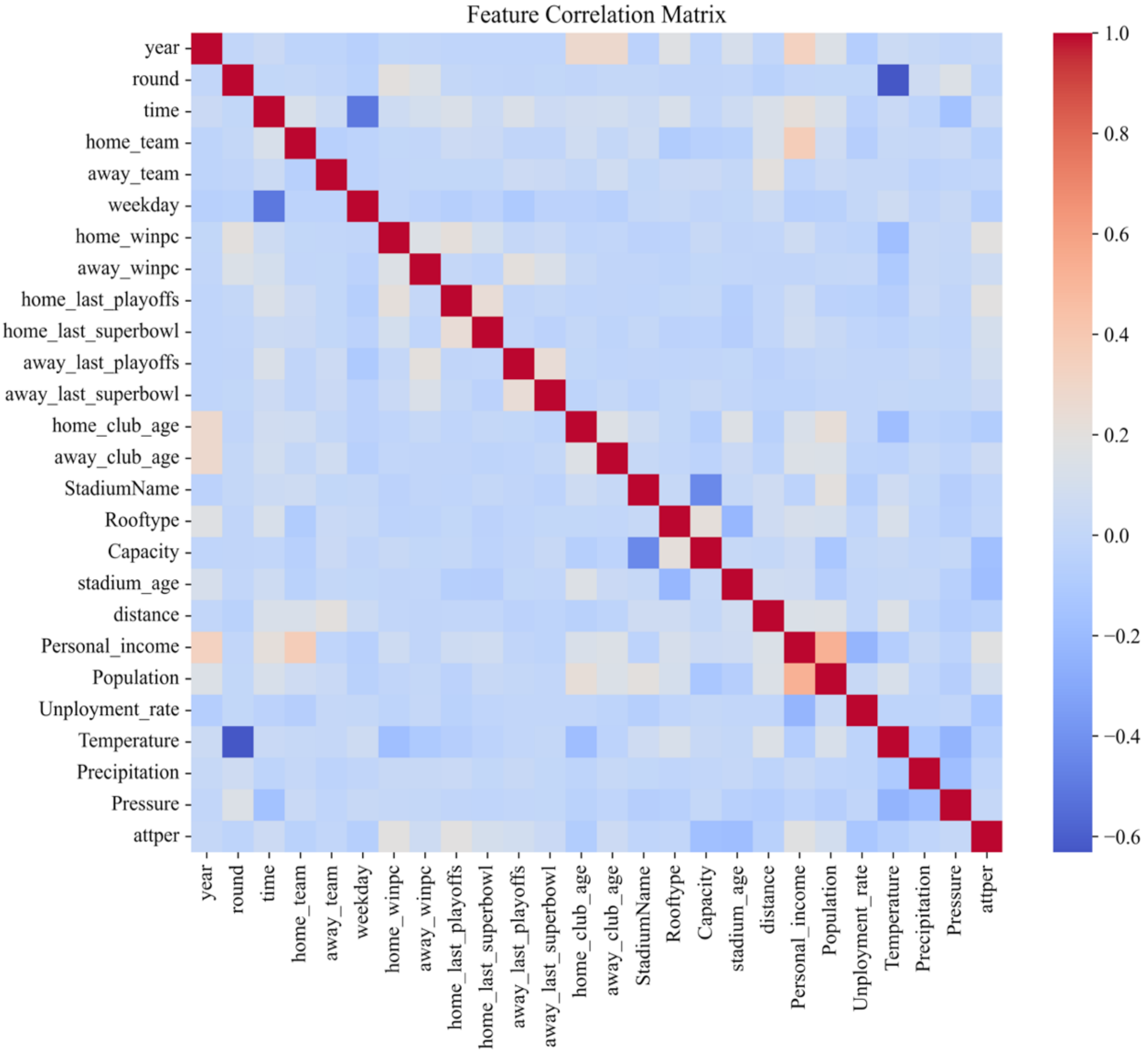

Fig. 1. Feature correlation matrix

Figure 1 presents a heatmap that illustrates the Spearman correlation matrix, offering a visual exploration of the relationships between the variables. It is important to note that Spearman coefficients close to or exceeding 0.7 may signal the presence of multicollinearity, which can distort model estimation (Rodionova et al., 2022). Such multicollinearity necessitates careful variable selection to avert the redundancy that highly correlated variable pairs can cause in the modeling process. In our analysis, the most notable correlation is observed between the *"round"* and *"weather"*, presenting a negative coefficient of -0.63. This notable inverse relationship likely arises from the

NFL season‘s span from September to January, where later rounds are typically associated with colder weather conditions. The other variables demonstrate negligible correlations, indicating a substantial level of independence and diminishing concerns for multicollinearity within our dataset.

Building upon the insights from the Spearman correlation matrix, we employed the forward selection method to refine the variables for subsequent analysis. This technique involves gradually adding variables to an initial base model, ranking them according to the magnitude of their relationship with the dependent variable. The inclusion criterion for each variable is the statistical significance of its p-value. Adhering to conventional significance benchmarks, we incorporated variables exhibiting p-values below the 0.05 threshold. This rigorous selection process yielded a subset of variables strongly correlated with the attendance percentage, which are anticipated to enhance the predictive capacity of our machine learning model. After this selection process, 17 variables remained, with the previous season‘s performance ranking as the most significant predictor. Table 3 displayed the detailed p-value for each variable considered.

Table 3. Selected variables from forward selection process

Variable	P-Value
home_last_playoffs	0.0000
stadium_age	0.0000
Personal_income	0.0000
StadiumName	0.0000
Capacity	0.0000
home_winpc	0.0000
distance	0.0000
away_last_playoffs	0.0000
Temperature	0.0000
home_club_age	0.0000
Unemployment_rate	0.0000
round	0.0000
Population	0.0000
away_club_age	0.0000
weekday	0.0012
Rooftype	0.0042
time	0.0090

Models' selection and validation

Python 3.11.0 within the PyCharm Integrated Development Environment was deployed in this study. A suite of machine learning algorithms was chosen due to their distinct advantages and proven track records in domains similar to sports attendance forecasting:

Linear Regression: This method works well when there's a direct, straight-line relationship between factors like how well a team is doing and the number of fans attending the games. It serves as an initial benchmark to ascertain the efficacy of alternative, more intricate methodologies under investigation, in order to assess their comparative performance in predicting attendance (Du et al., 2022; King, 2017; King et al., 2018; King & Rice, 2018).

CART (Classification and Regression Trees): CART employs a binary tree structure; it extends the simplicity of linear regression through a hierarchical decision-making process (Lewis, n.d.). It systematically segments data based on singular variable criteria, which is instrumental in isolating pivotal determinants of attendance such as game-specific attributes and team participation. Its efficacy in delineating influential predictors has garnered recognition and application across related scholarly investigations (Du et al., 2022; Mueller, 2020).

Random Forest: This ensemble method, synthesizing a multitude of decision trees, stands out for its powerful defense against overfitting—common pitfall in forecast analytics. Random Forest effectively navigates our dataset's multidimensional complexity, accommodating the abundance of variables without demanding feature reduction. Such an attribute is essential in the context of sports attendance, where a myriad of factors converge to influence outcomes (Du et al., 2022; King, 2017; King et al., 2018; King & Rice, 2018; Mueller, 2020).

XGBoost: XGBoost is highly regarded for its rapid processing capabilities and precision in model outcomes (Chen & Guestrin, 2016). This algorithm's flexibility renders it particularly compatible with the diverse nature of our dataset. The proven efficacy of XGBoost in similar applications supports its selection, offering a robust tool for accurately forecasting NFL game attendance (Du et al., 2022; King et al., 2018; King & Rice, 2018).

CatBoost: Research validated CatBoost's superior performance, often in comparison with XGBoost and Random Forest, attesting to its effectiveness in forecast modeling (Hong, 2020; Huang et al., 2019; Jabeur et al., 2021). Its adeptness in gradient boosting is key to its rapid and accurate forecasting abilities (Prokhorenkova et al., 2018), excelling in handling datasets rich in categorical features, which makes it particularly well suited for our dataset of NFL game attendance forecasting.

In our analysis model, we partitioned the dataset into a training set, constituting 90% of the data, and a testing set, making up the remaining 10%. A total of 506 matches were randomly selected to form the testing subset. To thoroughly assess the efficacy of our forecasting model, we used a suite of five metrics designed to capture its performance dimensions. These metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and R-squared. The initial four metrics quantify the deviation between the predicted and actual outcomes, with lower values indicative of a more accurate model. Conversely, the R-squared metric gauges the model's capacity to explain the variability of the data, where a higher value denotes a stronger explanatory power. While MAE provides a direct measure of average error, it does not differentiate between underestimation and overestimation; hence, RMSE is utilized to offer an alternative measure by representing the square root of the average squared errors, thus highlighting larger errors more prominently.

RESULTS

Performance comparison of all models

Table 4 presents an assessment of the performance metrics for various models, evaluated using their default settings with no parameter optimization. For consistency across models, a random seed of 42 was employed during the analysis.

Table 4. Models' performance comparison

Model	MAE	MSE	MAPE	RMSE	R^2
Linear Regression	0.04904	0.00555	0.05846	0.07449	0.21452
CART	0.02718	0.00336	0.03199	0.05794	0.52465
Random Forest	0.02062	0.00174	0.02452	0.04171	0.75373
CatBoost	0.0236	0.00171	0.02763	0.04132	0.75823
XGBoost	0.02483	0.00195	0.02896	0.04412	0.72444

The Linear Regression model served as the foundational benchmark of this study, presenting a modest R^2 of 0.21452. While less sophisticated than the ensemble algorithms, its transparent structure provided a valuable preliminary insight and a baseline for comparison.

Upon examining the ensemble methods, a marked improvement in prediction accuracy was observed. The CART model demonstrated enhanced explanatory power with an R^2 of 0.52465, indicating a considerable improvement over the baseline Linear Regression model. The Random Forest algorithm emerged as the superior performer within this comparison, achieving the R^2 of 0.75373. This reflects an optimal balance of bias and variance, as further evidenced by the lowest observed scores in MAE (0.02062), RMSE (0.04171), and MAPE (0.02452). The Random Forest model showcased its effectiveness in managing the complexities of the dataset while avoiding overfitting. Both XGBoost and CatBoost models yielded impressive performances, with R^2 scores of 0.72444 and 0.75823, respectively. They adapted robustly to the dataset, with XGBoost displaying particular accuracy in predicting attendance figures, as indicated by its MAPE score of 0.02896.

Detailed model analysis of the performance

Upon obtaining preliminary outcomes, our investigation pivoted to examining the predictive efficacy of the Random Forest, CatBoost, and XGBoost models. The refinement of these models was conducted through Grid Search, a paramount hyperparameter optimization method in machine learning. This procedure is essential for enhancing the accuracy of algorithmic predictions. Grid Search operates on the principle of exhaustively probing various hyperparameter configurations to discern the combination that optimizes model performance (Bergstra & Bengio, 2012). It methodically tunes hyperparameters, with each cycle meticulously calibrating the model's learning process based on the given data. Utilizing cross-validation within this approach permits a well-substantiated selection of hyperparameters, bolstering the model's capacity for generalization and its predictive power. We applied a 3-fold cross-validation technique as part of the Grid Search to ascertain the dependability of our results. This tripartite validation scheme is instrumental in ensuring the reliability and robustness of the tuning process. Table 5 presents the optimized hyperparameters for each model.

Table 5. Optimized hyperparameters of all models

Random Forest		**CatBoost**		**XGBoost**	
Hyperparameters	Optimized	Hyperparameters	Optimized	Hyperparameters	Optimized
bootstrap	False	depth	5	n_estimators	400
max_depth	20	iterations	3000	max_depth	7
max_features	sqrt	l2_leaf_reg	1	learning_rate	0.1
n_estimators	400	border_count	128	subsample	0.7
min_samples_leaf	1	subsample	0.9	colsample_bytree	0.8
min_samples_split	5	learning_rate	0.1	lambda	2
-	-	-	-	alpha	0

Table 6 provides a detailed look at the improved performance of the models following hyperparameter tuning. This process has clearly led to enhanced precision and predictive strength across the models, as demonstrated by reductions in MAE, RMSE, MSE, and MAPE values. Such improvements suggest that the models‘ forecasts are more closely aligned with actual observations, thereby increasing the reliability of their predictions.

Table 6. Detailed performance of optimized models

Model	**MAE**	**MSE**	**MAPE**	**RMSE**	**R²**
Random Forest	0.02062	0.00174	0.02452	0.04171	0.75373
Random Forest Optimization	**0.02035**	**0.00161**	**0.02404**	**0.04007**	**0.77268**
CatBoost	0.02360	0.00171	0.02763	0.04132	0.75823
CatBoost Optimization	0.02286	0.00162	0.02650	0.04022	0.77098
XGBoost	0.02483	0.00195	0.02896	0.04412	0.72444
XGBoost Optimization	0.02244	0.00175	0.02628	0.04185	0.75201

The optimized Random Forest model shows marginal advancements, most notably in MAPE, with an improved figure of 0.02404, and an MAE that is close to the smallest observed value at 0.02035. The CatBoost model, despite not surpassing the Random Forest in RMSE or MAE, exhibited a notable rise in the coefficient of determination, with its R^2 value reaching 0.77098, indicative of a strong fit to the dataset. The XGBoost model, on the other hand, displayed the most pronounced gains from the optimization, achieving lower RMSE at 0.04185 and MSE at 0.00175 post-optimization. It also showed a significant improvement in the R^2 value, increasing it to 0.75201, which underscores its enhanced capacity to explain the variation in NFL game attendance, cementing its position as a robust model in our analysis.

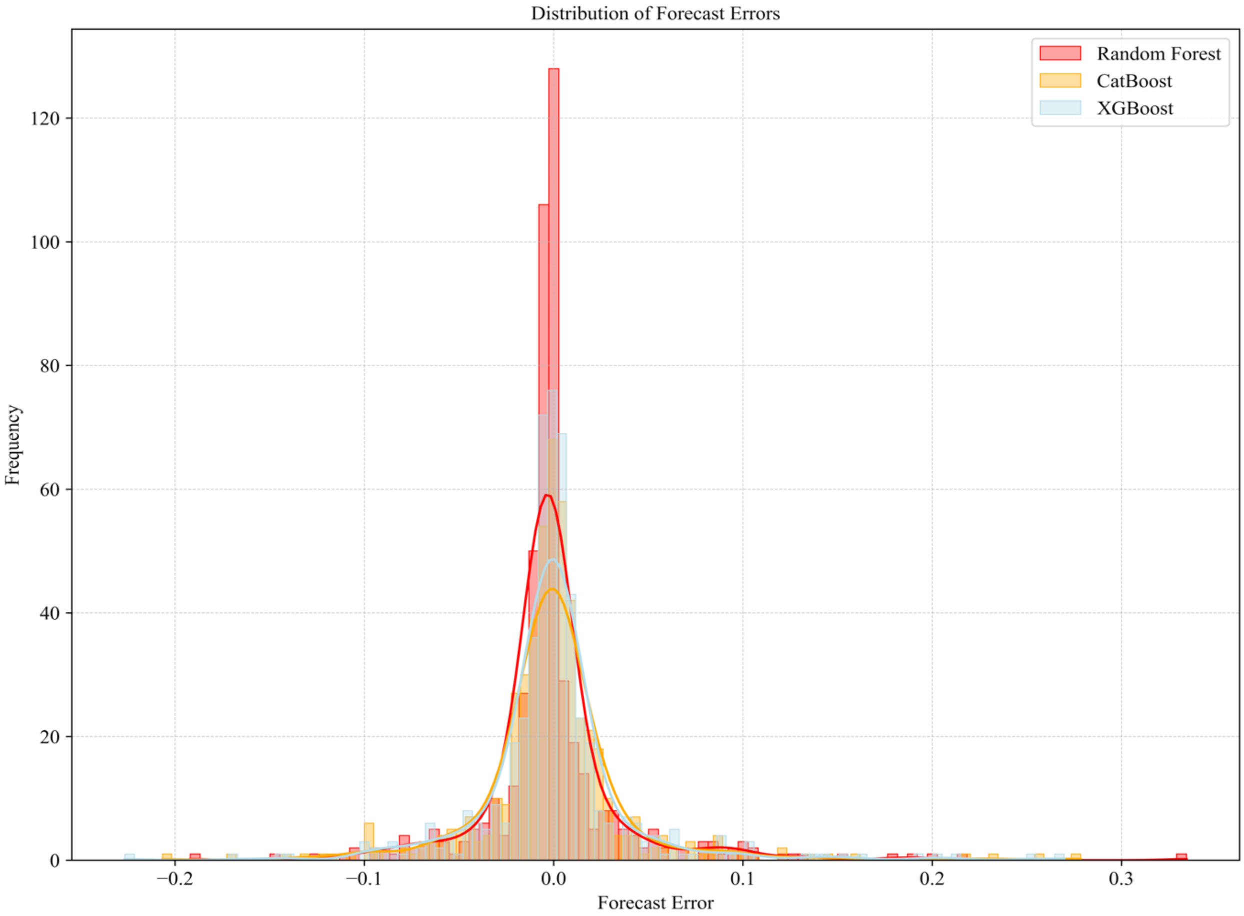

Fig. 2. Histograms of the distribution of forecast errors

Figure 2 illustrates histograms and kernel density estimates (KDE) for forecast errors generated by the Random Forest, CatBoost, and XGBoost models. The histograms reveal a clustering of errors around the zero mark for all three models, with the KDE peaks aligning closely with this central point. This aggregation near zero indicates that, on average, the models' predictions are quite accurate.The width of the KDE peaks and the spread of the histograms reflect the variability of the forecasts. A slimmer peak suggests a tight grouping of errors around the mean error, signaling more consistent performance from the model. Among the three, the Random Forest model's histogram is the narrowest, suggesting that its predictions are less varied and potentially more reliable than the others. The CatBoost model exhibits a slightly wider distribution, indicating a small increase in the variability of its predictive accuracy. XGBoost demonstrates a level of forecast error variability similar to CatBoost, as seen in the spread of its histogram.

Overall, the Random Forest model shows the highest stability in its predictions. CatBoost and XGBoost, while presenting somewhat more variability in their forecasts, still maintain a commendable level of accuracy, with all models demonstrating a generally reliable predictive performance.

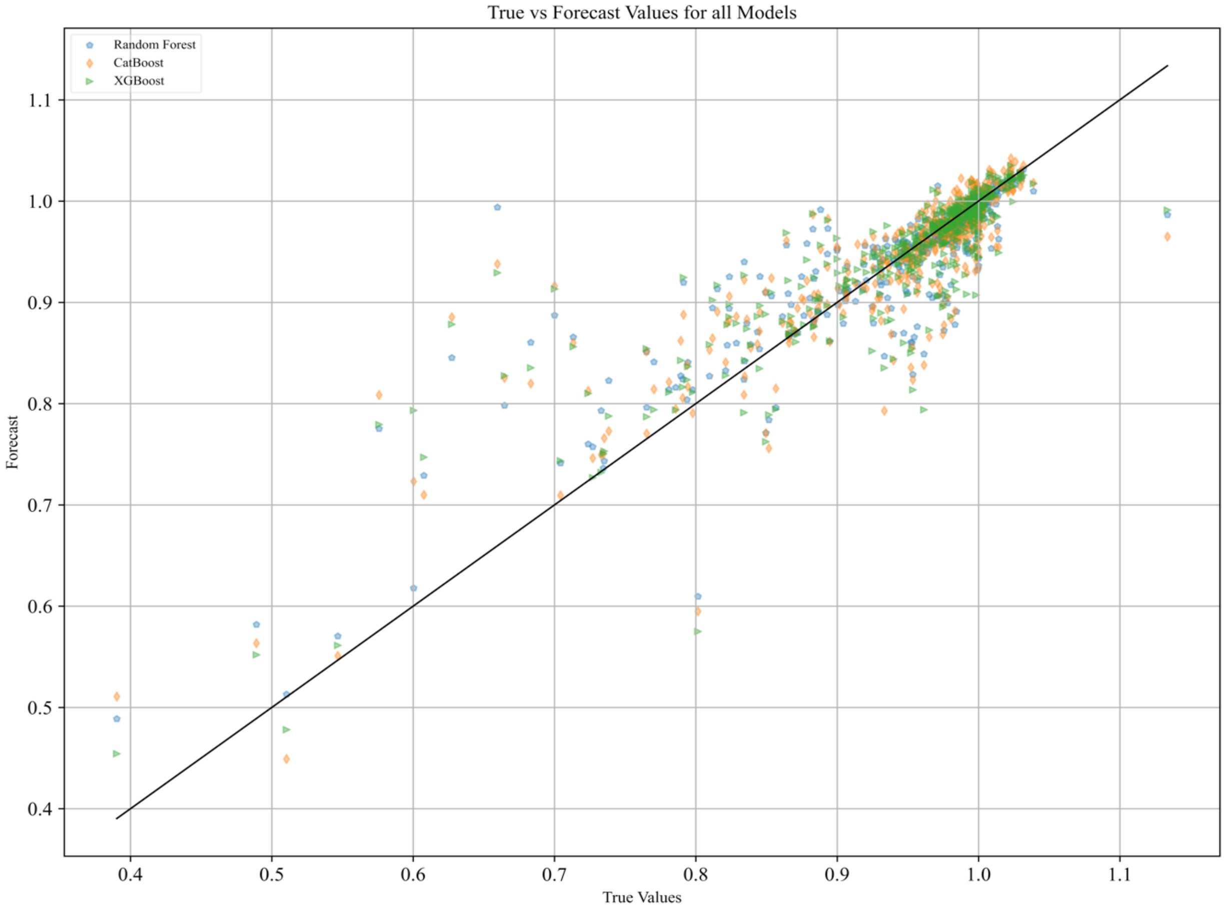

Fig. 3. Comparison of true and forecast values.

Figure 3 compares the actual versus predicted attendance values post-optimization for the three models. The diagram reveals a trend of underprediction within the lower attendance brackets, particularly within the 0.6 to 0.7 range on the x-axis. This recurrent underestimation points to a potential systematic bias or a deficiency in relevant predictive factors for this segment of attendance. In contrast, at higher attendance values—approaching the full capacity mark of 1.0 on the x-axis—the models' predictions align more closely with reality, suggesting increased accuracy when forecasting games with typical or expected attendance levels. While there are a handful of significant deviations from the line of ideal prediction, these outliers are relatively sparse and may reflect exceptional circumstances or anomalies not accounted for in the model's data inputs.

The spread of the data points across the chart highlights the models' variable performance across the spectrum of attendance figures. Predictions for games with lower attendance figures are less accurate, possibly due to a lack of comprehensive data or unconsidered factors, whereas forecasts for games with standard attendance figures seem to effectively capture the salient factors, yielding more precise predictions.

Feature importance analysis

In the final phase of our analysis, we evaluated the feature importance in the Random Forest, CatBoost, and XGBoost models to determine which variables hold the most influence in the forecasting process. Feature importance offers a quantitative measure of the variables that most significantly influence predictions of NFL game attendance. The corresponding visualizations display these importances, with the bar lengths representing the relative impact of each feature on predictive accuracy.

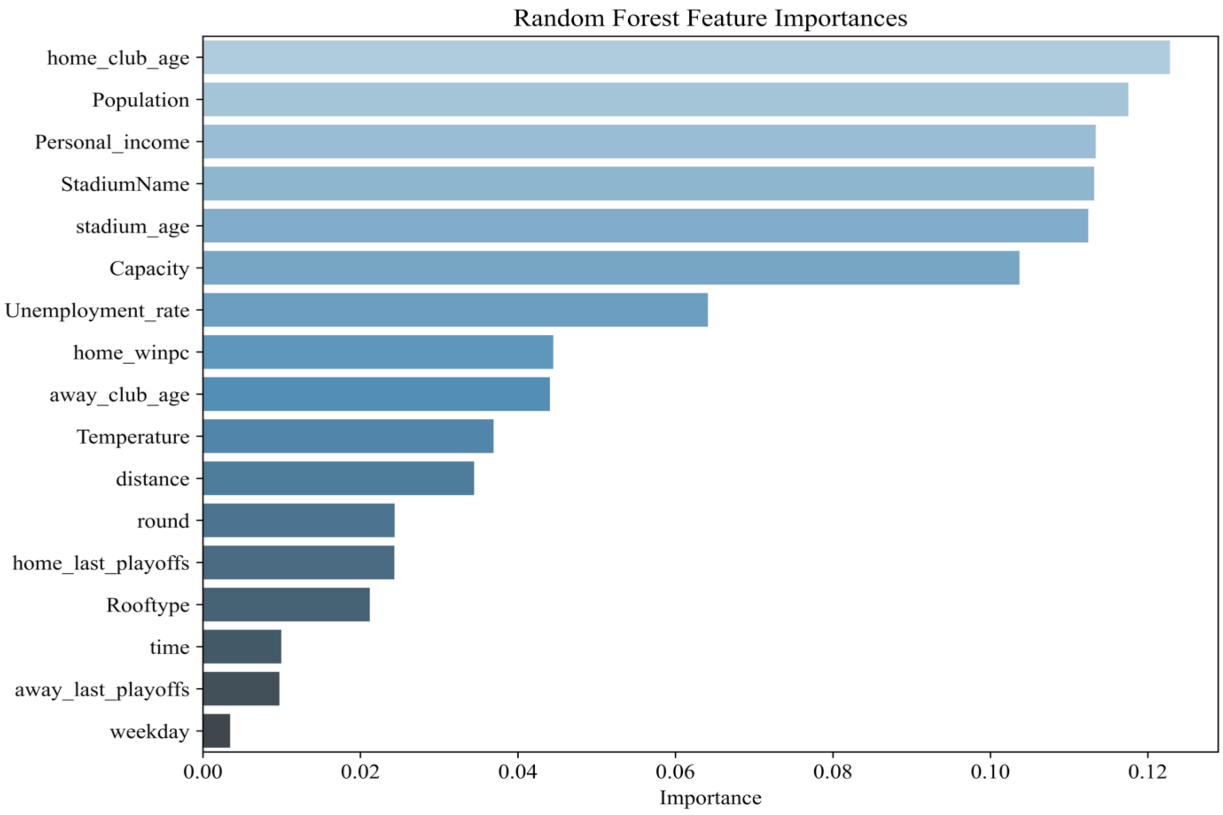

Fig. 4. Feature importance of the Random Forest model

Random Forest and XGBoost exhibited a consistent ranking in the order of feature importance. Despite the alignment in feature prioritization, Random Forest demonstrated a superior performance across several metrics, particularly in the optimized state, with a notable R^2 improvement from 0.75373 to 0.77268. This indicates that Random Forest not only identifies key features similarly to XGBoost but also utilizes them more effectively to capture the underlying patterns in NFL game attendance. CatBoost presented a slightly varied feature importance profile compared to Random Forest, but it showed comparable performance, with an R^2 closely matching that of the optimized Random Forest at 0.77098. This similarity in performance, despite the differences in feature prioritization, suggests that CatBoost may be handling the interactions between variables differently, yet still effectively.

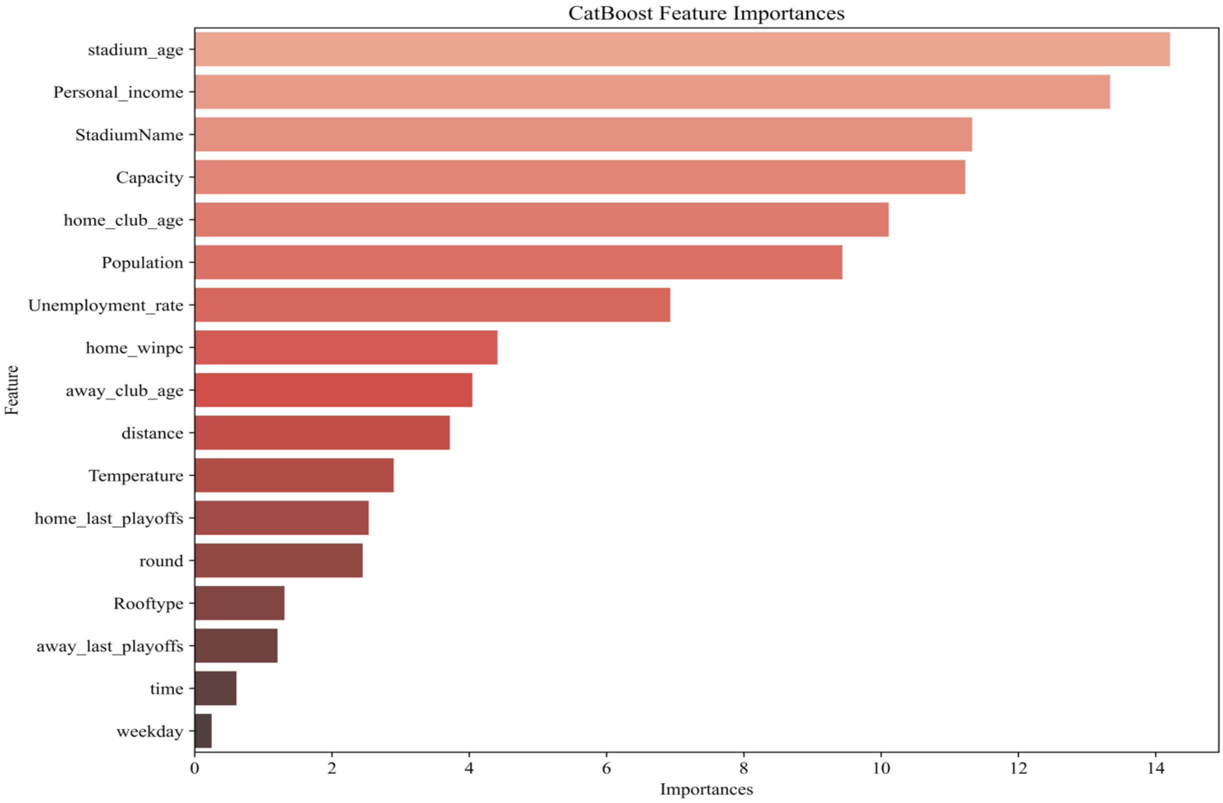

Fig. 5. Feature importance of the CatBoost model

In investigating the feature importance across Random Forest (see Fig. 4), CatBoost (see Fig. 5), and XGBoost models (see Fig. 6), an observable trend emerged, emphasizing the key factors influencing attendance metrics. The evaluation of feature importance across the models reveals that economic factors—Personal Income, Unemployment Rate, and Population—stand out as significant determinants of NFL game attendance, consistently ranking among the top seven features in terms of influence. This underlines the critical role of local economic health in driving attendance figures, confirming the strong connection between economic vibrancy and sports engagement (Du et al., 2022). While there is slight variation in the order of feature importance across different models, four additional key predictors remain constant: the age of the home team, stadium name, the age of stadium and capacity. These elements point to the significance of stadium-related factors in influencing attendance rates, with the stadium‘s brand recognition (captured by its name), seating capacity, and modernity (reflected in its age) being integral to drawing crowds. Furthermore, the age of the home club, indicative of its historical legacy and cultural heritage, is identified as another crucial factor. This suggests that a team‘s longstanding presence and the associated fan base cultivated over time considerably affect game attendance (Mueller, 2020).

Overall, the analysis suggests a multifaceted set of factors that contribute to the forecast of NFL game attendance. The consistency of economic predictors across all models solidifies the importance of these variables in sports analytics. Meanwhile, the identification of club legacy and stadium attributes emphasizes the comprehensive nature of factors that should be considered when predicting attendance, capturing both the tangible and intangible elements that appeal to NFL fans.

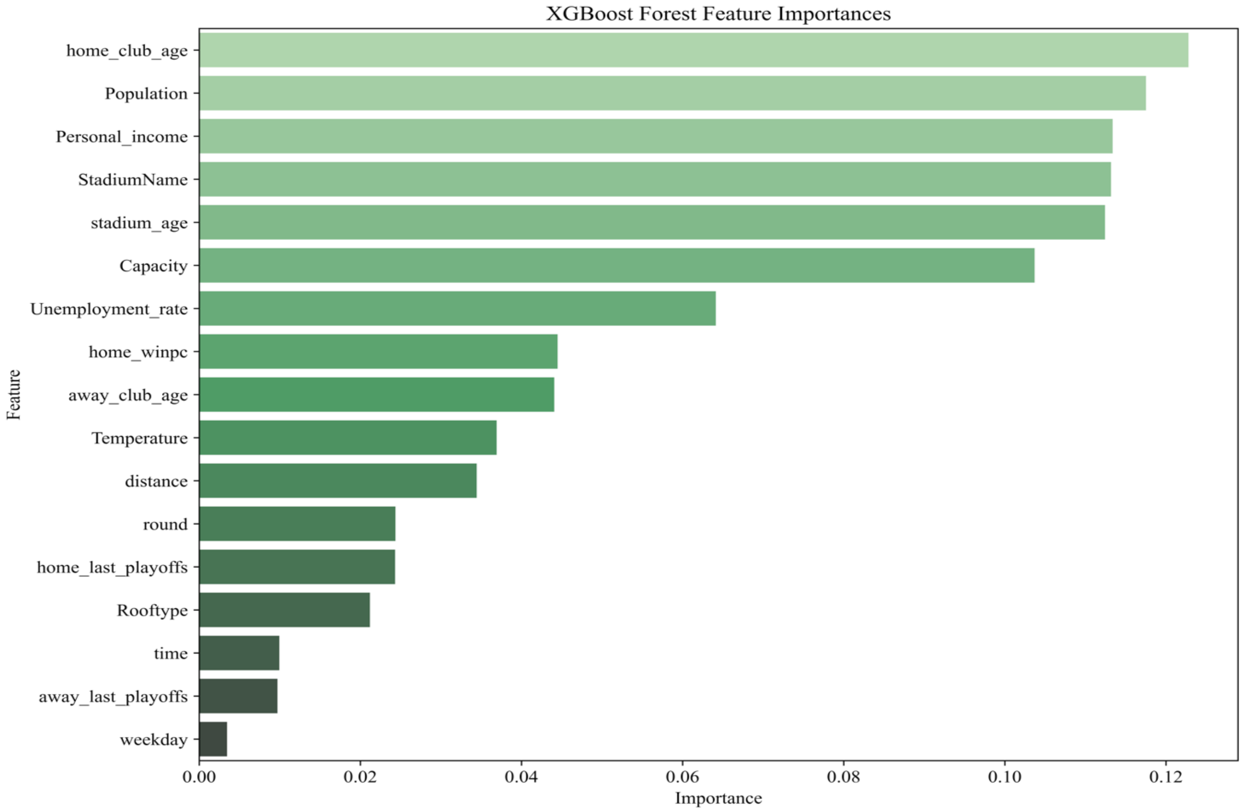

Fig. 6. Feature importance of the XGBoost model.

CONCLUSION

This research is distinctive as it covers the longest period compared to similar studies, providing unique insights into forecasting NFL game attendance by evaluating proven-effective machine learning models such as Linear Regression, CART, Random Forest, CatBoost, and XGBoost. We compiled a comprehensive dataset of over 5,055 regular-season NFL games from 2000 to 2019, incorporating a wide array of variables, including team performance metrics, economic indicators, stadium conditions, and weather patterns.

Our findings reveal that ensemble machine learning models, particularly Random Forest, achieved the best performance among all the models. The R^2 values for all three models — Random Forest, CatBoost, and XGBoost — exceeded 0.75 after tuning, with Random Forest reaching an impressive 0.773 and CatBoost also achieving a strong result at 0.771. Our model demonstrated notably low Mean Absolute Error (MAE) values of approximately 0.02. This contrasts with similar studies on NFL attendance, such as the one by Şahin and Uçar (2020), which reported Mean Absolute Percentage Error (MAPE) values around 0.1, reflecting a forecast error of about 10%. By comparing percentage errors, which provides a fairer assessment against actual attendance numbers, our results show a fivefold reduction in forecast error. This significant improvement marks a considerable advancement in the field. Furthermore, it is noteworthy that Random Forest and CatBoost exhibited unique advantages in prediction accuracy. Visualizations of error distribution highlighted each algorithm's strengths; Random Forest exhibited the most concentrated error distribution. Forecasts for attendance percentages between 0.6 and 0.7 were consistently over predicted. In terms of feature importance, 'Stadium_Name,' 'personal_income,' 'stadium_age,' and 'home_club_age' were consistently identified as key predictors across all models. This finding highlights the considerable impact of factors such as stadium characteristics, local economic conditions, and team history on attendance rates.

Managerial implications

The implications of these findings are significant for stakeholders of NFL, for instance, team managers, and sport investors. Accurate forecasts of game attendance can guide decisions regarding staffing, promotions, ticket pricing, and overall fan engagement strategies. Furthermore, understanding the crucial factors that drive attendance can help in tailoring marketing efforts and enhancing the overall game-day experience for fans. Our research provides a strategic template for similar analyses in other sports contexts.

Research limitations and future directions

While this study offers valuable insights into NFL game attendance forecasting, it is not without limitations. Firstly, the scope of our analysis was restricted to regular-season games, excluding playoff matches and games in atypical venues, which could exhibit different attendance dynamics. Additionally, our models' forecast power may be limited by the availability and accuracy of external data sources, such as weather conditions. The inherent unpredictability of certain events, like extraordinary performance or unexpected team changes, also poses challenges to the models' accuracy. Future research could expand on our work by including playoff games and exploring the impact of atypical venues on attendance. Investigating the effect of real-time social media sentiment or fan engagement metrics could also enrich the predictive model. Lastly, applying our methodology to other sports businesses or incorporating emerging machine learning techniques, such as deep learning, could offer broader insights into sports attendance dynamics.

ACKNOWLEDGEMENTS

The author gratefully acknowledges financial support from China Scholarship Council.

REFERENCES

Alonso, A. D., & O'Shea, M. (2013). The links between reasons for game attendance of a new professional sports league and revenue management: An exploratory study. *International Journal of Revenue Management*, *7*(1), 56–74. https://doi.org/10.1504/IJRM.2013.053359

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Borland, J., & Macdonald, R. (2003). Demand for Sport. *Oxford Review of Economic Policy*, *19*(4), 478–502. https://doi.org/10.1093/oxrep/19.4.478

Bowley, J. L., & Berger, P. D. (2017). Predicting National Football League (NFL) stadium attendance. *International Journal of Social Science and Business*, *2*(3).

Buraimo, B. (2008). Stadium attendance and television audience demand in English league football. *Managerial and Decision Economics*, *29*(6), 513–523. https://doi.org/10.1002/mde.1421

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Coates, D., & Humphreys, B. R. (2007). Ticket prices, concessions and attendance at professional sporting events. *International Journal of Sports Finance, 2*(3), 161–170.

Coates, D., & Humphreys, B. R. (2010). Week to week attendance and competitive balance in the National Football League. *International Journal of Sport Finance*, *5*(4), 239.

Depken, C. A. (2001). Fan loyalty in professional sports: An extension to the National Football League. *Journal of Sports Economics*, *2*(3), 275–284. https://doi.org/10.1177/152700250100200306

Du, P., Wang, Y., Liao, C., & Xian, T. (2022). Sports games attendance forecast using machine learning. *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, 181–188. https://doi.org/10.1109/ICDSCA56264.2022.9987748

Falls, G. A., & Natke, P. A. (2014). College football attendance: A panel study of the Football Bowl Subdivision. *Applied Economics*, *46*(10), 1093–1107. https://doi.org/10.1080/00036846.2013.866208

Falls, G. A., & Natke, P. A. (2016). College football attendance: A panel study of the Football Championship Subdivision. *Managerial and Decision Economics*, *37*(8), 530–540. https://doi.org/10.1002/mde.2740

Ge, Q., Humphreys, B. R., & Zhou, K. (2020). Are fair weather fans affected by weather? Rainfall, habit formation, and live game attendance. *Journal of Sports Economics*, *21*(3), 304–322. https://doi.org/10.1177/1527002519885427

Gropper, C. C., & Anderson, B. C. (2018). Sellout, blackout, or get out: the impacts of the 2012 policy change on TV blackouts and attendance in the NFL. *Journal of Sports Economics*, *19*(4), 522–561. https://doi.org/10.1177/1527002516661600

Gupta, R. (2019). *Prediction of major factors affecting fans attendance for the teams of major league baseball*. Dublin, National College of Ireland.

Hansen, H., & Gauthier, R. (1989). Factors affecting attendance at professional sport events. *Journal of Sport Management*, *3*(1), 15–32. https://doi.org/10.1123/jsm.3.1.15

Hart, R. A., Hutton, J., & Sharot, T. (1975). A statistical analysis of association football attendances. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *24*(1), 17–27. https://doi.org/10.2307/2346700

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Hong, J. (2020). An application of XGBoost, LightGBM, CatBoost algorithms on house price appraisal system. *Housing Finance Research*, *4*, 33–64. https://doi.org/10.52344/hfr.2020.4.0.33

Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., Zeng, W., & Zhou, H. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology*, *574*, 1029–1041. https://doi.org/10.1016/j.jhydrol.2019.04.085

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Jennett, N. (1984). Attendances, uncertainty of outcome and policy in Scottish League Football. *Scottish Journal of Political Economy*, *31*(2), 176–198. https://doi.org/10.1111/j.1467-9485.1984.tb00472.x

King, B. E. (2017). Predicting National Basketball Association game attendance using random forests. *Journal of Computer Science*, *5*(1), 1–14. https://doi.org/10.15640/jcsit.v5n1a1

King, B. E., & Rice, J. (2018). Predicting attendance at major league soccer matches: A comparison of four techniques. *Journal of Computer Science and Information Technology*, *6*, 15–22. https://doi.org/10.15640/jcsit.v6n2a2

King, B. E., Rice, J. L., & Vaughan, J. (2018). Using machine learning to predict National Hockey League average home game attendance. *The Journal of Prediction Markets*, *12*(2), 85–98. https://doi.org/10.5750/jpm.v12i2.1608

Lenten, L. J. (2011). Long-run trends and factors in attendance patterns in sport: Australian Football League, 1945–2009. *Handbook on the Economics of Leisure, Edward-Elgar, Northampton*, 360–380. https://doi.org/10.4337/9780857930569.00026

Lewis, R. J. (2000.). An introduction to Classification and Regression Tree (CART) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California* (Vol. 14). San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Mueller, S. Q. (2020). Pre- and Within-Season attendance forecasting in major league baseball: A random forest approach. *Applied Economics*, *52*(41), 4512–4528. https://doi.org/10.1080/00036846.2020.1736502

Nesbit, T. M., & King, K. A. (2010). The impact of fantasy football participation on NFL attendance. *Atlantic Economic Journal*, *38*(1), 95–108. https://doi.org/10.1007/s11293-009-9202-x

Paul, R. J., Ehrlich, J. A., & Losak, J. (2021). Expanding upon the weather: Cloud cover and barometric pressure as determinants of attendance for NFL games. *Managerial Finance*, *47*(6), 749–759. https://doi.org/10.1108/MF-06-2020-0295

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS 2018)*. Montréal, Canada.

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1410. https://doi.org/10.1186/s40064-016-3108-2

Rodionova, M., Skhvediani, A., & Kudryavtseva, T. (2022). Prediction of crash severity as a way of road safety improvement: The case of Saint Petersburg, Russia. *Sustainability*, *14*(16), 9840. https://doi.org/10.3390/su14169840

Şahin, M., & Erol, R. (2018). Prediction of attendance demand in European football games: Comparison of ANFIS, Fuzzy Logic, and ANN. *Computational Intelligence and Neuroscience*, *2018*, 1–14. https://doi.org/10.1155/2018/5714872

Şahin, M., & Uçar, M. (2020). Prediction of sports attendance: A comparative analysis. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *236*(2), 106–123. https://doi.org/10.1177/1754337120983135

Spenner, E. L., Fenn, A. J., & Crooker, J. (2004). The demand for NFL attendance: A rational addiction model. *Colorado College Economics and Business Working Paper*, *2004–01*. http://dx.doi.org/10.2139/ssrn.611661

Welki, A. M., & Zlatoper, T. J. (1999). U.S. professional football game-day attendance. *Atlantic Economic Journal*, *27*(3), 285–298. https://doi.org/10.1007/BF02299579

Contact Information:

yupang@hse.ru; https://orcid.org/0000-0002-8330-4581
fwang@hse.ru; https://orcid.org/0000-0003-3103-5049