INTERNATIONAL JOURNAL OF BUSINESS AND MANAGEMENT (IJBM)
ISSN: 2815-9330 (Online)

VOL. 5, NO. 3 [Special Issue] 2026 https://iessociety.org/

Digital engagement, competitive balance, and stadium attendance in Indian Women's Football: An explainable machine learning and PLS-SEM framework

Prithu Sarkar*[1], Dr. Shatabdi Som[2] and Dr. Romica Bhat[3]

[1]Research Scholar, Amity School of Communication, Amity University Kolkata India

[2]Associate Professor, Amity School of Communication, Amity University Kolkata India; Email: shatabdizmail@gmail.com; ORCID ID: 0000-0002-9775-4815

[3]Professor, Lovely Professional University, India; Email: romica.32881@lpu.co.in; ORCID ID: 0000-0003-0870-0610

*Correspondence: prithusarkar90@gmail.com; ORCID ID: 0000-0002-9775-4815

Abstract

Attendance underpins the commercial sustainability and visibility of women's football, yet research on attendance demand in emerging markets remains scarce. This study proposes an integrated predictive and explanatory model for Indian women's football, combining conventional inferential statistics with seven machine learning algorithms, explainable artificial intelligence, and composite PLS SEM. The analytical dataset comprised 1,000 match level observations spanning economic, venue, team quality, competitive balance, match context, digital communication, digital engagement, fan base, and environmental indicators. Prediction was evaluated through an 80:20 holdout split with baseline and tuned models, alongside nested 5x5 cross validation. Model performance was assessed using R squared, RMSE, and MAE, while SHAP, LIME, and partial dependence and individual conditional expectation analyses supported interpretation. Theory guided structural relationships were tested through a Mode B composite PLS SEM model with 5,000 bootstrap resamples. Random Forest achieved the strongest tuned holdout performance (R squared = 0.155, RMSE = 4,946, MAE = 4,128), while SGD Regressor produced the lowest mean nested cross validation RMSE (4,855). Match importance and weekend scheduling emerged as the strongest predictive signals. The structural model explained 18.0 percent of the variance in attendance and showed positive predictive relevance (Q squared = 0.167). Match Context was the only construct with a strong, statistically significant direct structural effect on attendance (beta = 0.418, p < 0.001, f squared = 0.211), while digital communication and engagement showed weaker and structurally uncertain effects. These findings indicate that, within the joint predictive and structural framework, match context is the most actionable lever for scheduling, marketing, and fan development decisions in Indian women's football.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

***Keywords*:** Women's Football, Stadium Attendance, Sports Analytics, Digital Fan Engagement, Explainable Artificial Intelligence, Machine Learning, PLS SEM, Sports Marketing

JEL Classification: Z22, C45, C55

1. Introduction

One of the most obvious signs of the evolution of a market in professional sport is its stadium attendance. It can impact gate receipts, concession and merchandising income, sponsor exposure, perceived product attractiveness and the atmosphere of a competition. These functions are particularly significant in women's football, as the development of the sport relies not just on the level of play, but also on translating the increased awareness of football onto steady match day demand. Traditional sport-demand literature has conceptualized attendance as a result of a combination of economic, sporting and venue, and other contextual factors (Borland & Macdonald, 2003; Schreyer & Ansari, 2022). This rationale has since been extended to women's football, where scheduling, local and venue characteristics, quality of the teams, stage of the competition and uncertainty of the outcome have all proven to be relevant in various settings (Valenti et al., 2020; Valenti et al., 2024; Buraimo et al., 2026). The burgeoning literature also warns against taking results from well-known men's leagues for granted when they are applied to women's leagues. Valenti et al. (2020) identified four factors that were related to attendance (554 UEFA Women's Champions League matches): competition stage, uncertainty, competitive intensity, and away-club reputation, and weather. Valenti et al. (2024) found that in the FA Women's Super League, weekends, weather, distance, local economies, and capacity dictated demand. In a more recent study, Buraimo et al. (2026) demonstrated that the impact of the uncertainty of outcome is not universal, but depends on the country, with outcome uncertainty remaining important in England, France and Germany. These differences suggest that evidence for attendance research needs to be local and not assumed.

The second spectator demand change is the fast growth of digital communication. Clubs are now in constant communication with fans via Instagram, Facebook, X/Twitter, video platforms and other social media platforms. It can provide fans with information, deepen brand connections, drive engagement, and foster community connections between fixtures. Football club research indicates that there are motivations related to information, social, identity and brands that shape online engagement (Vale & Fernandes, 2018), and that the content produced by clubs can influence both the brand associations and how fan communities interpret the team brand (Parganas et al., 2017; Anagnostopoulos et al., 2018). The importance of fan involvement and league–club brand relationships is also stressed in the concept of sport-brand scholarship (Kunkel et al., 2014).

But a key conceptual difference is important: digital engagement does not equal physical attendance. A fan may like, share, comment on or even view a post without being at a match and

DOI: https://doi.org/10.56879/ijbm.v5i3.134

matchday attendance may be limited by scheduling, location, ticket price or even the importance of the game. The Chinese Women's Super League is especially illustrative. Liu et al. (2025) determined that different factors drove stadium attendance and online streaming demand: weekend scheduling and stadium structure were linked to in-person demand, while competitive balance and team quality were more related to online streaming demand. Thus, the online-to-offline connection should not be assumed but should be tested.

Machine learning offers a complementary point of view to explanatory models, as it is intended to learn predictive patterns that may be nonlinear, interactive, and/or distributed over many variables. AI has seen a significant surge in its applications in football analytics, ranging from player performance analysis and event detection in the game to predicting match outcomes and injury analysis (Elstak et al., 2024). Recently, Beermann et al. (2026) tested how linear regression, Random Forest and XGBoost performed in the prediction of attendance demand in 2,069 international football matches and revealed that the nonlinear models can significantly outperform linear regression in a comprehensive international-football dataset. This study also demonstrated the importance of the features to better understand the predictive power of an economic, venue and team-quality information. Prediction alone, nevertheless, does not establish a structural relationship. A variable may be highly useful to a Random Forest because it partitions observations effectively, but its corresponding theoretical path may be weak after controlling for other constructs. Conversely, a theoretically important variable can show a stable structural path while contributing relatively little incremental predictive information. This distinction motivates the combination of explainable machine learning with PLS-SEM in the present research.

Four connected gaps motivate the study. First, stadium-attendance research remains concentrated in North American and European markets, with comparatively little evidence from women's football and emerging football markets, a limitation previously highlighted by Schreyer and Ansari (2022). Second, existing women's-football studies have largely used econometric or regression frameworks (Valenti et al., 2020; Valenti et al., 2024; Liu et al., 2025; Buraimo et al., 2026), leaving room for broader model comparison under leakage-aware machine learning. Third, digital communication and engagement are often studied as marketing outcomes in their own right, while the transition from pre-match communication to online engagement and then to stadium attendance is less frequently tested at the match level. Fourth, predictive importance and structural evidence are normally reported separately. A framework that combines the two can provide decision-makers with a clearer distinction between variables that are useful for forecasting and variables that receive support within a theory-guided structural model.

The study therefore addresses four research questions. RQ1 asks how accurately stadium attendance in Indian women's football can be predicted using economic, venue, team, match-context, digital, fan-base, and environmental information. RQ2 asks which match-level and digital variables contribute most strongly to predictive performance. RQ3 asks which structural

DOI: https://doi.org/10.56879/ijbm.v5i3.134

relationships among digital communication, digital engagement, competitive conditions, match context, and attendance receive empirical support. RQ4 asks how predictive importance from explainable machine learning can be combined with PLS-SEM evidence to generate an actionable prioritization framework.

The study makes four contributions. Contextually, it extends attendance-demand research to Indian women's football, an under-researched setting. Methodologically, it compares seven regression learners using baseline models, tuning, a nested 5×5 cross-validation procedure, and multiple explainability tools rather than relying on a single predictive algorithm. Theoretically, it tests a multi-construct model that links digital communication, digital engagement, team quality, match context, economic/venue conditions, fan base, and the physical attendance outcome. Managerially, it introduces an ML×PLS prioritization matrix in which SHAP-based predictive importance is crossed with PLS-SEM structural evidence. The framework is intended to identify not merely what predicts attendance, but what deserves managerial priority under the joint predictive and structural evidence.

2. Literature Review and Theoretical Framework

2.1 Economic Theory of Spectator Demand

The most general theoretical base of the study is the economics of demand for of sport spectacles. Based on the work of Borland and Macdonald (2003), attendance is regarded as a choice of consumption that can be influenced by monetary cost, quality of the sporting contest, uncertainty of the result, the viewing experience and alternatives the consumer has. It means that demand can be affected by ticket price, venue conditions, accessibility, team quality and context of the match. Previous literature, summarized by Schreyer and Ansari (2022), found that studies have explored the impact of price, income, stadium quality, scheduling, weather, team performance, star players, rivalry, uncertainty, and substitution effects on attendance with significant variation across sports and markets.

Different cost and accessibility mechanism can apply for women's football than in the highly commercialized men's football. The convenience, match time, attraction of the venue and travel time are all part of the total cost of attendance which a relatively low ticket price may minimize. Valenti et al. (2024) reported that weekend scheduling and shorter travel distance were correlated with higher levels of attendance to FA WSLs, which represents the additional costs beyond the ticket price associated with attending a match. For the current research, ticket price is thus used as an economic/access measure, and capacity of the stadium is used as a venue/structural measure.

2.2 Outcome Uncertainty, Team Quality and Competitive Balance

In sports economics, one of the most controversial concepts is that of outcome uncertainty. The classical argument is that contests are more attractive the less predictable the outcome is, which is questionable from the empirical evidence. Pawlowski and Anders (2012) demonstrated that in

DOI: https://doi.org/10.56879/ijbm.v5i3.134

German football the link between uncertainty and stadium attendance does not have to be monotonic. A more general meta-literature also shows that fans decision-making is based on relative as well as absolute quality, not simply on uncertainty. In the case of women's football, Valenti et al. (2020) revealed a positive correlation between outcome uncertainty and attendance at UEFA Women's Champions League matches, while Buraimo et al. (2026) identified some evidence of the impact of uncertainty at the game level in Germany but not in England and France. The results recommend context-sensitive approach of competitive balance.

Home and away team rankings, rank differences and a competitive-balance index are used to represent competitive conditions in the present study. The lower the numbers, the stronger the teams, and the greater the difference between the numbers, the greater the disparity between the opponents, and the closer the match-ups, the larger the competitive-balance index. They are all conceptually grouped together under Team Quality and Competitive Balance (C3). Since there is some uncertainty as to whether fans prefer uncertainty or absolute quality, H3 is stated as a significant relationship, with no restriction on the sign of the attendance response.

2.3 Match Context and Scheduling

Match context is the fit that captures features that make one fixture more consequential and/or more convenient than another. The competition level always plays a pivotal role in football for women. Valenti et al. (2020) identified higher demand in the later stages of the UEFA Women's Champions League and Buraimo et al. (2026) demonstrated that match importance can have an impact on attendance even if the uncertainty at game level is not very high. Another thing that comes up many times is weekend scheduling. A number of studies have linked weekend games and bigger stadium attendance such as Valenti et al. (2024) and Liu et al. (2025) in the Chinese Women's Super League. The results offer solid evidence for the use of the composite Match Context (C4) with match importance and weekend scheduling.

There is a theoretical separation between the match context and the quality of the teams. A final can be very appealing even if it lacks competitive balance because of the scarcity and significance of the event, while a balanced league can lack attendance when the season is off-peak. The difference matters in making management decisions because it can be more readily acted upon in the important areas of scheduling and event packaging than it can in the makeup of playing squads.

2.4 Digital Communication as Stimulus

For digital sports marketing, we can take a look at the Stimulus–Organism–Response perspective (Mehrabian & Russell, 1974). For this study, stimulus is defined as pre-match social-media communication. Instagram, X/Twitter, and Facebook posts provide multiple touchpoints for information, reminder, expression, socialization and promotion. As shown in the sport-marketing literature, these are not just media that are broadcast but are interactive brand worlds where the followers are consuming, contributing, and creating content (Vale & Fernandes, 2018). Other

DOI: https://doi.org/10.56879/ijbm.v5i3.134

studies also reveal that Facebook and Twitter interactions can lead to various brand association effects (Parganas et al., 2017), and how professional football clubs exploit Instagram in a strategic manner for visual branding (Anagnostopoulos et al., 2018).

Therefore, Digital Communication Effort (C5) is defined as a formative behavioural composite that consists of the number of pre-match posts on the three platforms. The formative treatment is apt since each platform offers a unique communication medium and a high level of activity on one does not necessarily indicate a high level of activity on another.

2.5 Digital Engagement as a Behavioral Response

Digital Engagement Response (C6) records matches' platform interactions. Engagement is calculated using Instagram, X/Twitter and Facebook engagement counts as well as a follower normalized engagement rate. Conceptually, engagement can be seen as an intermediate behavior which may indicate awareness and interest. But the connection from digital engagement to in-person attendance is not automatic. Attending the stadium means having additional resources and availability, fans may participate online as a result of geographic distance, habit or convenience. Liu et al. (2025) clearly show that the drivers of online and physical attendance may differ in women's football. So, in the present study we do not take the digital-to-physical pathway for granted.

2.6 Fan Base and Brand Strength

Another attendance source is a team's base of fans. Sport-brand research suggests that the relationship between supporter, team and league is important and that it has to be strong if consumers are to be involved (Kunkel et al., 2014). Social platforms also enable the size of the reachable audience to be observed by number of followers. Home/away-team follower counts are used as proxies for Fan Base/Brand Strength (C7). These indicators are not actually measuring psychological identification; they are measuring the size of the digitally addressable audience associated with the fixture. The construct is added as a structural predictor and control as a large fan base may lead to a larger pool from which matchday spectators are drawn, even if short-term digital engagement is low.

2.7 Explainable Machine Learning in Spectator-Demand Research

The motivating force behind the predictive aspect of the study is that machine-learning models are able to handle the nonlinearities and interactions. Random Forest is a combination of many decorrelated decision trees (Breiman, 2001), while Gradient Boost builds models by incrementally addressing the weaknesses of weak learners (Friedman, 2001), with XGBoost giving an efficient regularized implementation (Chen & Guestrin, 2016); CatBoost is specifically designed to handle nonlinear tabular structure by ordered boosting (Prokhorenkova et al., 2018). A regularized linear benchmark is ridge regression (Hoerl & Kennard, 1970), and an alternative flexible margin-based

DOI: https://doi.org/10.56879/ijbm.v5i3.134

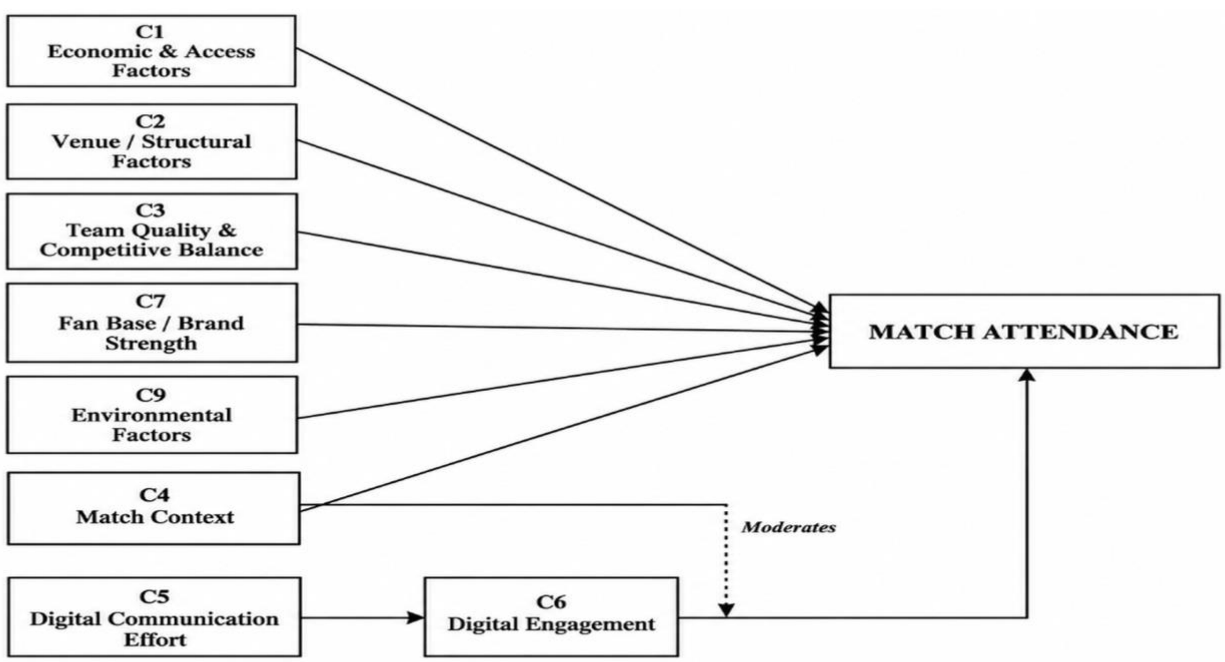
method is support vector regression (Cortes & Vapnik, 1995). Since multiple families are used, conclusions are less likely to be substantive conclusions that rely on a single learning algorithm.

Interpretability is crucial, as a great model is not very useful to a manager if they do not know why the model is giving them the prediction. SHAP calculates additive feature contributions, using cooperative game theory (Lundberg & Lee, 2017), while LIME estimates individual predictions by interpretable models in their vicinity (Ribeiro et al., 2016). These tools are complemented by partial-dependence and ICE analyses, which illustrate the modelled response for the range of observed data for selected predictors. These methods are able to distinguish globally-dominant predictors from case-specific signals.

2.8 Conceptual Model and Hypotheses

The economic, competitive, contextual, digital, fan-base and environmental aspects are brought together in Figure 1 to create a theory-driven model. Seven worthy hypotheses are tested. The hypothesis of H1 is that the Digital Communication Effort is positively related to Digital Engagement. H2 suggests that Digital Engagement has a positive effect on Match Attendance. H3 proposes a significant relationship between Team Quality/Competitive Balance and Match Attendance. H4 is proposing a positive relationship between Match Context and Match Attendance. It is proposed by H5 that Digital Engagement's impact on Attendance is moderated by Match Context. H6: Economic and Access Barriers have a negative effect on Attendance. H7 suggests Venue/Structural Advantage has a positive impact on Attendance. Focal paths are not interpreted in isolation by retaining Fan Base and Environmental Conditions as structural controls.

Figure 1. Conceptual Research Model and Hypothesized Relationships



DOI: https://doi.org/10.56879/ijbm.v5i3.134

Table 1. Constructs, Theoretical Foundations, Indicators, and Hypothesized Roles

Construct	Theoretical basis	Indicators	Model role	Expectation
C1 Economic & Access	Consumer demand theory	Ticket price	Direct predictor	H6: negative expected
C2 Venue/Structural	Sport-demand / viewing quality	Stadium capacity	Direct predictor	H7: positive expected
C3 Team Quality & Competitive Balance	Outcome uncertainty / absolute quality	Home rank, away rank, competitive-balance index	Direct predictor	H3: significant
C4 Match Context	Event significance and scheduling	Match importance, weekend match	Direct predictor and moderator	H4 positive; H5 moderation
C5 Digital Communication Effort	S-O-R stimulus	Instagram, X/Twitter, Facebook pre-match posts	Antecedent of C6	H1 positive
C6 Digital Engagement Response	S-O-R behavioral response	Platform engagement counts, engagement rate	Direct predictor	H2 positive
C7 Fan Base / Brand Strength	Sport brand / fan community	Home and away followers	Structural control	Control relationship
C9 Environmental Conditions	Sport-demand context	Weather score	Structural control	Control relationship
Attendance	Spectator demand	Match attendance	Outcome	Primary dependent variable

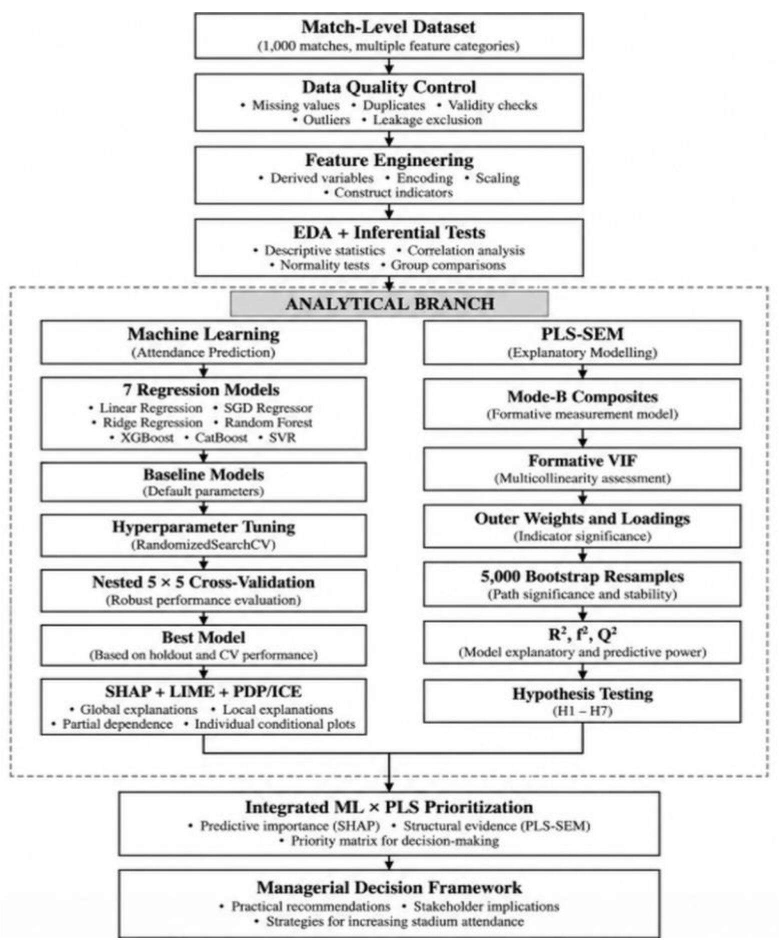
DOI: https://doi.org/10.56879/ijbm.v5i3.134

3. Methodology

3.1 Research Design and Analytical Architecture

The study adopted a quantitative match-level predictive–explanatory design. The analytical architecture consisted of three connected layers. The first layer assessed the dataset through quality-control checks, exploratory analysis, correlations, outlier screening, and classical inferential tests. The second layer evaluated attendance prediction through seven machine-learning algorithms under baseline and tuned configurations, an 80:20 holdout split, and nested 5×5 cross-validationThe third layer was built around theory-based composite relationships, estimated by theory-guided Mode-B PLS-SEM and combined the structural evidence with SHAP-derived predictive importance of the relations. The overall workflow is summarized in figure 2. The design deliberately dissociates prediction and structural explanation: accuracy of the model refers to whether attendance can be predicted or not, and PLS-SEM refers to whether or not the hypothesized relationships between constructs are supported.

Figure 2. End-to-End Analytical Architecture



DOI: https://doi.org/10.56879/ijbm.v5i3.134

3.2 Dataset, Unit of Analysis, and Data Quality Control

The third layer was built around theory-based composite relationships, estimated by theory-guided Mode-B PLS-SEM and combined the structural evidence with SHAP-derived predictive importance of the relations. The overall workflow is summarized in figure 2. The design deliberately dissociates prediction and structural explanation: accuracy of the model refers to whether attendance can be predicted or not, and PLS-SEM refers to whether or not the hypothesized relationships between constructs are supported. The dependent variable for the primary analysis was match_attendance. Internal quality-control procedures checked row count, duplicate match identifiers, same-team pairings, non-positive values, missing core variables, and attendance values relative to the harmonized stadium-capacity field. The final file contained no duplicate match identifiers, no same home/away team records, no core missing values, and no attendance values above the analysis capacity field. Audit/QC columns and target-derived fields were excluded from predictive feature sets.

Outliers were treated as diagnostic observations rather than automatically deleted. IQR rules and Isolation Forest were used to flag unusual cases, but model training retained the complete analytical sample because extreme attendance or engagement values can represent meaningful event conditions. A robustness analysis subsequently compared the full sample with the 892 records that did not carry a repair flag in the audit layer. The similar cross-validated RMSE values in the two samples were used to assess sensitivity to data preparation.

Table 2. Dataset Quality-Control and Sample Profile

Check	Result	Status
Rows	1,000	Pass
Duplicate match_id	0	Pass
Same home-away team	0	Pass
Attendance above capacity	0	Pass
Missing core values	0	Pass
Team labels	8	Information
Stadiums	6	Information

DOI: https://doi.org/10.56879/ijbm.v5i3.134

3.3 Construct Operationalization and Feature Engineering

Construct operationalization followed the theoretical model while avoiding deterministic redundancy. C1 Economic & Access was represented by ticket price. C2 Venue/Structural was represented by stadium capacity. C3 Team Quality and Competitive Balance combined reverse-oriented home and away team rankings with a competitive-balance indicator. C4 Match Context combined match importance and weekend scheduling. C5 Digital Communication used the three pre-match platform post counts. C6 Digital Engagement used Instagram, X/Twitter, and Facebook engagement together with engagement rate. C7 Fan Base used home and away follower counts, and C9 Environmental Conditions used weather score. All construct indicators entered the PLS model after standardization.

Several derived features were created for prediction and diagnostics. The absolute rank difference was defined as:

$$RD_i = |Rank\_home,_i - Rank\_away,_i| \quad (1)$$

Competitive balance was represented with an inverse-gap index in which closer rankings receive larger values:

$$CBI_i = 1 / (1 + RD_i) \quad (2)$$

Total pre-match social communication and total digital engagement were calculated as:

$$SP_i = IGP_i + TWP_i + FBP_i \quad (3)$$

$$DE_i = IGE_i + TWE_i + FBE_i \quad (4)$$

A follower-normalized engagement rate was also computed:

$$ER_i = [DE_i / (Followers\_home,_i + Followers\_away,_i)] \times 100 \quad (5)$$

Others predictive features were platform concentration measures, follower imbalance, engagement per post, active social platforms, and ticket price per 1,000 units of capacity. Occupancy rate and the high-attendance label were not included in attendance regression since they were target derived. The key observed and engineered variables that were kept for interpretation are summarized in Table 3.

Table 3. Principal Variables and Operational Definitions

Variable	Operational meaning	Analytical role	Type
match_attendance	Number of spectators	Primary outcome	Observed

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Variable	Operational meaning	Analytical role	Type
ticket_price	Match ticket price	Economic/access predictor	Observed
stadium_capacity	Analysis-level venue capacity	Venue predictor	Observed/harmonized
team_rank_home	Home-team rank	Team-quality predictor	Observed
team_rank_away	Away-team rank	Team-quality predictor	Observed
rank_gap_abs	Absolute difference between team ranks	Competitive balance	Engineered
competitive_balance_index	1/(1+rank gap)	Competitive balance	Engineered
match_importance	League / semi-final / final	Match context	Observed
weekend_match	1=weekend, 0=weekday	Match context	Observed
instagram_posts_pre_match	Instagram pre-match post count	Digital communication	Observed
twitter_posts_pre_match	X/Twitter pre-match post count	Digital communication	Observed
facebook_posts_pre_match	Facebook pre-match post count	Digital communication	Observed

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Variable	Operational meaning	Analytical role	Type
instagram_engagement	Instagram engagement	Digital engagement	Observed
twitter_engagement	X/Twitter engagement	Digital engagement	Observed
facebook_engagement	Facebook engagement	Digital engagement	Observed
engagement_rate	Total engagement / combined followers ×100	Digital engagement	Engineered
team_followers_home	Home-team followers	Fan base	Observed
team_followers_away	Away-team followers	Fan base	Observed
weather_score	Normalized match-weather score	Environmental	Observed
high_attendance_70pct	Attendance ≥70% of capacity	Auxiliary classification target	Outcome-derived; classification only

3.4 Exploratory and Inferential Statistical Analysis

Distribution, central tendency, dispersion, pairwise relationships, and group differences were explored in the distribution of the data. Descriptive statistics of the primary outcome and core predictors were created. The Pearson and Spearman correlations were used to differentiate between linear and monotonic relationships. The normality of attendance was assessed using Shapiro–Wilk and Q–Q diagnostics, since the formal tests for normality can be very sensitive at N=1,000. Variance inflation factors were used to check the multicollinearity. Outliers were detected using IQR rules as well as Isolation Forest, however they were not automatically removed. Weekend versus weekday attendance was assessed with Welch's independent-samples t-test and Mann–

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Whitney U, with Cohen's d as an effect-size measure. Match-importance groups were compared with classical one-way ANOVA, Welch ANOVA, and Kruskal–Wallis tests, with η^2 and epsilon-squared reported where appropriate. Attendance across stadiums was assessed with Welch ANOVA and Kruskal–Wallis. False-discovery-rate correction was applied in multi-test contexts. The preliminary tests were used to describe empirical regularities; they were not substituted for the predictive or structural analyses.

3.5 Machine-Learning Models and Preprocessing

Seven regression algorithms represented different predictive families. Linear Regression provided an unregularized benchmark. Ridge Regression represented regularized linear estimation. SGD Regressor provided a stochastic linear learner suitable for high-dimensional encoded features. Random Forest represented bagged decision trees and nonlinear interactions (Breiman, 2001). XGBoost and CatBoost represented modern boosting methods (Chen & Guestrin, 2016; Prokhorenkova et al., 2018). Support Vector Regression represented a nonlinear kernel-based alternative (Cortes & Vapnik, 1995). Numeric variables were median-imputed when required; categorical variables were most-frequent imputed and one-hot encoded. Scaling was applied to algorithms that depend on feature magnitude, while tree-based models used unscaled numeric predictors. All preprocessing was included inside model pipelines to reduce leakage during cross-validation.

The final primary regression set contained 27 raw predictors after excluding identifiers, audit fields, target variables, and direct target-derived variables. Categorical inputs included home team, away team, stadium, and match importance. Numeric inputs included economic, venue, team-rank, competitive-balance, communication, engagement, follower, weekend, weather, and selected non-redundant engineered features. Deterministically related variables were not interpreted as independent causal effects; their role was predictive.

3.6 Training, Hyperparameter Tuning, Nested Cross-Validation, and Performance Metrics

The data were separated into 80% training and 20% holdout test observations, with quantile-based stratification on attendance to preserve the distribution of the outcome. Baseline versions of all seven models were first trained for continuity with conventional attendance-prediction studies. Tuned versions were then selected using RandomizedSearchCV. To prevent hyperparameter optimization from being evaluated on the same folds used to choose parameters, a nested cross-validation design was implemented: five outer folds estimated generalization performance and five inner folds optimized model hyperparameters. This design is stricter than a single train/test split and supports fairer comparison among model families.

Predictive performance was summarized with the coefficient of determination, root mean squared error, and mean absolute error:

$$R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2] \quad (6)$$

DOI: https://doi.org/10.56879/ijbm.v5i3.134

$RMSE = \sqrt{[(1/n)\ \Sigma(y_i - \hat{y}_i)^2]}$ (7)

$MAE = (1/n)\ \Sigma|y_i - \hat{y}_i|$ (8)

For nested CV, mean, standard deviation, and confidence intervals were reported across outer folds. Fold-wise RMSE values were also compared with paired t-tests and Wilcoxon signed-rank tests. Because only five outer folds were available, these significance tests were interpreted cautiously and in conjunction with the magnitude and consistency of performance differences.

3.7 Auxiliary Classification and Explainable Artificial Intelligence

The definition of "label" is attendance at or above 70% of the analysis capacity. This outcome was only used for classification and not in the primary attendance regression model as a predictor. Using Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC, Logistic Regression, SGD Classifier, Random Forest, XGBoost, CatBoost, and support vector classification were compared. As only 13.1% of observations were in the high-attendance class, ROC-AUC was viewed in combination with precision–recall performance.

Explainability was based on the most powerful holdout model, which was tree-based. Feature contributions were quantified using SHAP TreeExplainer at both the global and observation level (Lundberg & Lee, 2017). The direction and magnitude of the top predictors were summarized with a SHAP beeswarm plot. To evaluate local explanations (Ribeiro et al., 2016), LIME was applied to representative high, low, typical and high-error observations. Selected numeric predictors had average and case-specific shapes of the response displayed in partial-dependence and ICE curves. These methods are useful for substantive interpretation but do not make claims about causal relationships.

3.8 Composite Measurement Model and PLS-SEM Estimation

A PLS-SEM/PLS path modelling approach was used to estimate the construct model. The behavioral indicators were included as Mode-B composites because the indicators are parts of a whole, and constitute the behavior and not interchangeable signs of a latent psychometric property (Henseler et al., 2009; Hair et al., 2021). This treatment was particularly crucial for Digital Engagement: the platform engagement indicators were not subjected to a reflective measurement model. The generic composite score is represented as:

$C_j = \Sigma_k\ w_{jk}\ z_{jk}$ (9)

where z_{jk} denotes a standardized indicator and w_{jk} is the Mode-B outer weight. Multicollinearity of the indicators was evaluated by VIF, with a more conservative cut-off of 3.3. PLS algorithm converged after 16 iterations. Outer weights and loadings were then examined using bootstrap distributions. Fixed unit weights by construction were used for single-indicator composites, like ticket price, stadium capacity, weather and attendance.

3.9 Structural model, moderation, effect sizes, and predictive relevance

DOI: https://doi.org/10.56879/ijbm.v5i3.134

There were two connecting equations in the structural portion. First, Digital Engagement was predicted by Digital Communication. Second, Digital Engagement, Team Quality/Competitive Balance, Match Context, Economic/Access conditions, Venue/Structural conditions, Fan Base, Environmental conditions and the Digital Engagement × Match Context interaction were all modeled as predictors of standardized Attendance:

$$C6_i = \alpha_0 + \alpha_1 C5_i + \varepsilon_i \quad (10)$$

$$ATT_i = \beta_0 + \beta_1 C6_i + \beta_2 C3_i + \beta_3 C4_i + \beta_4 C1_i + \beta_5 C2_i + \beta_6 C7_i + \beta_7 C9_i + \beta_8 (C6_i \times C4_i) + \varepsilon_i \quad (11)$$

Statistical inference used 5,000 requested bootstrap resamples, with successful resamples retained for coefficient distributions, confidence intervals, and two-sided bootstrap p-values. The model reported R^2 and adjusted R^2 for endogenous constructs. Local effect size was calculated as:

$$f^2 = (R^2\_included - R^2\_excluded) / (1 - R^2\_included) \quad (12)$$

Predictive relevance was assessed through cross-validated Q^2:

$$Q^2 = 1 - (PRESS / SSO) \quad (13)$$

Values of Q2 greater than zero mean that predictions made with the cross validation are better than the naive prediction based on the mean of the outcome. Structural coefficients were deemed associations that were or were not compatible with the hypotheses, and the nature of the observational design does not prove the experimental causality.

3.10 Integrated ML–PLS Prioritization and Robustness Analysis

Predictive and structural evidence were combined in the last analytical step. Raw SHAP values were summed from the model features to the model's theory constructs. Predictive importance was compared to the bootstrap p-values of the respective PLS paths for each construct. The median construct-level SHAP importance discriminated between higher and lower predictive relevance and $p<0.05$ discriminated structurally supported versus unsupported paths. This resulted in four managerial categories: Priority Lever (high SHAP + significant PLS), Predictive but Structurally Uncertain (high SHAP + non-significant PLS), Long-term Structural Lever (lower SHAP + significant PLS), and Lower Priority (lower SHAP + non-significant PLS). Robustness checks involved predicting with and without repair flags (a subset of 892 records) and using the 18-predictor core feature set versus the 27-predictor extended model. The integration was intended to avoid black-box feature importance or relying solely on structural coefficients, which have a weak predictive capability to make managerial recommendations.

4. Results and Discussion

4.1 Sample Integrity and Descriptive Profile

The analytical sample contained 1,000 match observations and passed the hard quality-control checks summarized in Table 2. Mean attendance was 12,838.8 spectators (SD=5,298.6), with a

DOI: https://doi.org/10.56879/ijbm.v5i3.134

median of 12,259.5 and a maximum of 29,037. The average ticket price was 176.2, while the mean analysis-level stadium capacity was 27,953.5. The mean home-team and away-team rankings were 5.02 and 5.04, respectively. The average match contained 13.72 pre-match posts across the three platforms and 22,184 digital engagement actions. Weekend fixtures represented 48.7% of the sample. Table 4 presents descriptive statistics for the core observed and engineered variables.

Table 4. Descriptive statistics of core variables (N=1,000)

Variable	Mean	SD	Min	Median	Max
Match attendance	12,838.817	5,298.551	1,058	12,259.500	29,037
Ticket price	176.211	71.157	50	179	299
Stadium capacity	27,953.493	1,364.764	25,954	28,782	29,472
Home-team rank	5.024	2.573	1	5	9
Away-team rank	5.039	2.578	1	5	9
Instagram posts	4.596	2.883	0	5	9
X/Twitter posts	4.526	2.904	0	5	9
Facebook posts	4.599	2.966	0	5	9
Instagram engagement	9,729.272	5,842.155	0	9,714	19,967
X/Twitter engagement	5,058.182	2,904.686	0	5,265	9,997
Facebook engagement	7,396.969	4,276.893	0	7,271.5	14,985

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Variable	Mean	SD	Min	Median	Max
Home-team followers	251,772.630	145,912.665	2,146	253,299.5	499,517
Away-team followers	253,544.511	138,469.917	4,668	261,038	499,864
Weather score	0.647	0.206	0.300	0.642	0.999
Rank gap	2.925	2.128	0	3	8
Competitive-balance index	0.368	0.257	0.111	0.250	1.000
Total social posts	13.721	5.170	0	14	27
Total digital engagement	22,184.423	7,746.449	1,178	22,042	42,262

4.2 Preliminary Statistical Results

Preliminary inference showed that match context was already visible before the machine-learning and PLS analyses. Weekend fixtures averaged 13,774.4 spectators compared with 11,950.6 for weekday fixtures, a difference of approximately 1,824 spectators. Welch's t-test was significant ($t=5.53$, $p<0.001$), with Cohen's $d=0.349$, and the Mann–Whitney robustness test reached the same substantive conclusion ($p<0.001$). Match importance produced an even stronger pattern. League matches averaged 10,080.4 spectators, semi-finals 13,038.5, and finals 15,241.4. One-way ANOVA yielded $F=90.62$ ($p<0.001$) with $\eta^2=0.154$; Welch ANOVA and Kruskal–Wallis tests were also significant. By contrast, attendance did not differ significantly across the six stadium categories under Welch ANOVA ($p=0.441$) or Kruskal–Wallis ($p=0.458$). Table 5 summarizes these tests.

The correlation pattern reinforced the contextual result. Match-importance score correlated $r=0.391$ with attendance and weekend status $r=0.172$. In comparison, total digital engagement correlated $r=0.039$, total social posts $r=-0.020$, competitive balance $r=-0.013$, ticket price $r=0.040$, and weather score $r=-0.018$. While this does not mean that digital variables are not useful for prediction -- nonlinear and interaction effects may still be used by machine-learning models -- it does suggest that the most significant univariate signal in this data is the match context.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Table 5. Preliminary Inferential Tests for Attendance

Comparison	Test	Statistic	p	Effect/interpretation
Weekend vs weekday	Welch t-test	t=5.530	<0.001	Cohen's d=0.349
Weekend vs weekday	Mann–Whitney U	U=150,631	<0.001	Robustness
Match importance	One-way ANOVA	F=90.618	<0.001	η^2=0.154
Match importance	Welch ANOVA	F=92.571	<0.001	Robustness
Match importance	Kruskal–Wallis	H=156.067	<0.001	ε^2=0.155
Stadium	Welch ANOVA	F=0.962	0.441	Not significant
Stadium	Kruskal–Wallis	H=4.666	0.458	Not significant

4.3 Baseline and Tuned Machine-Learning Performance

Table 6 shows the performance of the holdout tasks before and after tuning. Random Forest had the smallest RMSE (5,066.4) and the largest R^2 (0.114) at baseline, just outperforming Linear Regression and Ridge. Tuning boosted the performance of all nonlinear learners, particularly XGBoost (ΔR^2=0.152) and CatBoost (ΔR^2=0.067). Random Forest was the best tuned holdout case with an R^2=0.155, RMSE=4,946.4, and MAE=4,127.6. CatBoost and XGBoost were close behind, with $R^2 \approx 0.141$. Overall these results show a small but consistent level of predictive power: the

Table 6. Baseline Versus Tuned Holdout Regression Performance

Model	Base R^2	Base RMSE	Base MAE	Tuned R^2	Tuned RMSE	Tuned MAE	ΔR^2
Random Forest	0.114	5,066.4	4,189.0	0.155	4,946.4	4,127.6	+0.041

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Model	Base R^2	Base RMSE	Base MAE	Tuned R^2	Tuned RMSE	Tuned MAE	ΔR^2
CatBoost	0.074	5,178.9	4,275.3	0.141	4,988.5	4,184.5	+0.067
XGBoost	-0.012	5,413.9	4,383.6	0.141	4,989.5	4,174.9	+0.152
Ridge Regression	0.103	5,099.0	4,240.2	0.105	5,091.0	4,245.5	+0.003
Linear Regression	0.104	5,093.7	4,230.5	0.104	5,093.7	4,230.5	0.000
SGD Regressor	0.095	5,120.6	4,256.3	0.102	5,101.0	4,245.2	+0.007
SVR	-0.020	5,435.8	4,423.9	0.055	5,231.0	4,304.1	+0.075

models accounted for a significant proportion of the variation in attendance, but with much of the variation still unaccounted for by the match-level variables available.

4.4 Nested Cross-Validation and Model Robustness

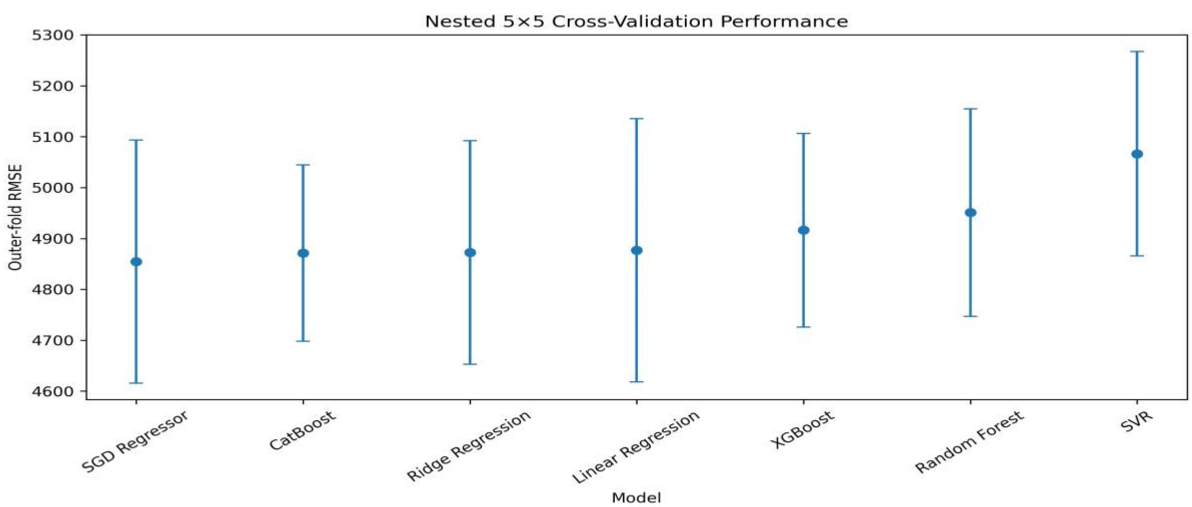
The single holdout split and the nested 5×5 design resulted in a slightly different ranking, which demonstrates the importance of repeated resampling. SGD Regressor achieved the lowest mean outer-fold RMSE (4,854.6) and the highest mean R^2 (0.150), followed closely by CatBoost (RMSE=4,871.4; R^2=0.144), Ridge Regression (4,872.7; 0.143), and Linear Regression (4,876.8; 0.142). The top Random Forest model for the holdout sample ranked lower for the nested CV with mean RMSE=4,950.8 and R^2=0.116. Figure 3 shows the outer-fold RMSE distributions and the complete summary is given in Table 7. Fold-wise significance testing revealed that SGD outperformed Random Forest with a significant RMSE at paired t-test level 0.027, and CatBoost, Ridge, Linear Regression, XGBoost did not show any significant difference ($p>0.05$) with respect to RMSE. This behaviour indicates that there is no one algorithm that was the most successful in all evaluation scenarios. The most defensible conclusion thus is that there is a cluster of similarly performing models and that Random Forest had the best holdout result while SGD had the lowest nested-CV mean error.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Table 7. Nested 5×5 Cross-Validation Performance

Model	Mean R²	SD R²	Mean RMSE	SD RMSE	Mean MAE	95% CI for R²
SGD Regressor	0.150	0.052	4,854.6	238.7	3,985.6	0.085–0.214
CatBoost	0.144	0.033	4,871.4	173.3	4,024.5	0.103–0.185
Ridge Regression	0.143	0.041	4,872.7	219.7	4,013.4	0.092–0.195
Linear Regression	0.142	0.053	4,876.8	258.7	3,987.4	0.076–0.208
XGBoost	0.128	0.035	4,916.3	190.3	4,054.0	0.085–0.171
Random Forest	0.116	0.044	4,950.8	204.1	4,074.1	0.060–0.171
SVR	0.074	0.035	5,066.5	201.1	4,118.2	0.031–0.117

Figure 3. Outer-fold RMSE under nested 5×5 cross-validation

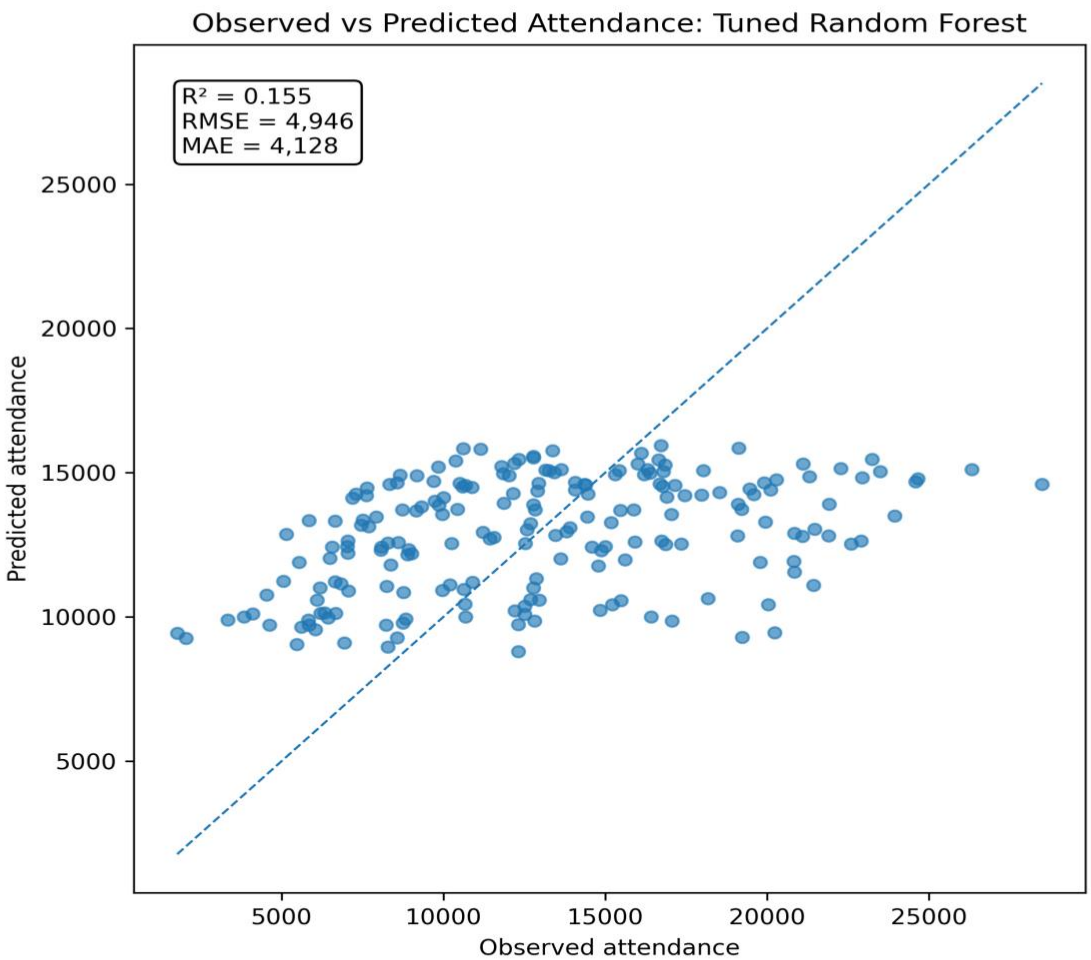


DOI: https://doi.org/10.56879/ijbm.v5i3.134

4.5 Prediction Diagnostics

The comparison of observed and predicted attendance for the tuned Random Forest on the held-out 20% test set is shown in Figure 4. Given that this model was forced into a line fit, it was found that the model was able to capture the overall ordering of low and high behavioural demand matches and to bring the extreme values towards the center, a pattern that is common when trying to predict noisy behavioural demand from a small number of structured match variables. The holdout R^2 of 0.155 is therefore to be interpreted as modest explanatory prediction, not near-deterministic forecasting. Residual and learning curve diagnostics indicated a difference between the training and validation errors that did not diminish as sample size increased, and validation RMSE has plateaued, suggesting that more observations and more temporal/behavioral features might help with generalization.

Figure 4. Observed Versus Predicted Attendance for the Tuned Random Forest Holdout Model



DOI: https://doi.org/10.56879/ijbm.v5i3.134

4.6 Explainable Machine-Learning Results

By far the largest predictor from the SHAP analysis was match importance. The weekend match (446.7) and one-hot encoding of match importance categories (mean absolute SHAP importance of ~2,008.5) were the next important features. The next group consisted of engagement concentration (95.9), home-team identity (86.6), weather score (83.5), X/Twitter engagement (71.4), Facebook engagement (68.7), away followers (68.4), home followers (68.0), and engagement per post (65.2). The ranking is given in Table 8. The SHAP beeswarm of the most important transformed features is displayed in Figure 5.

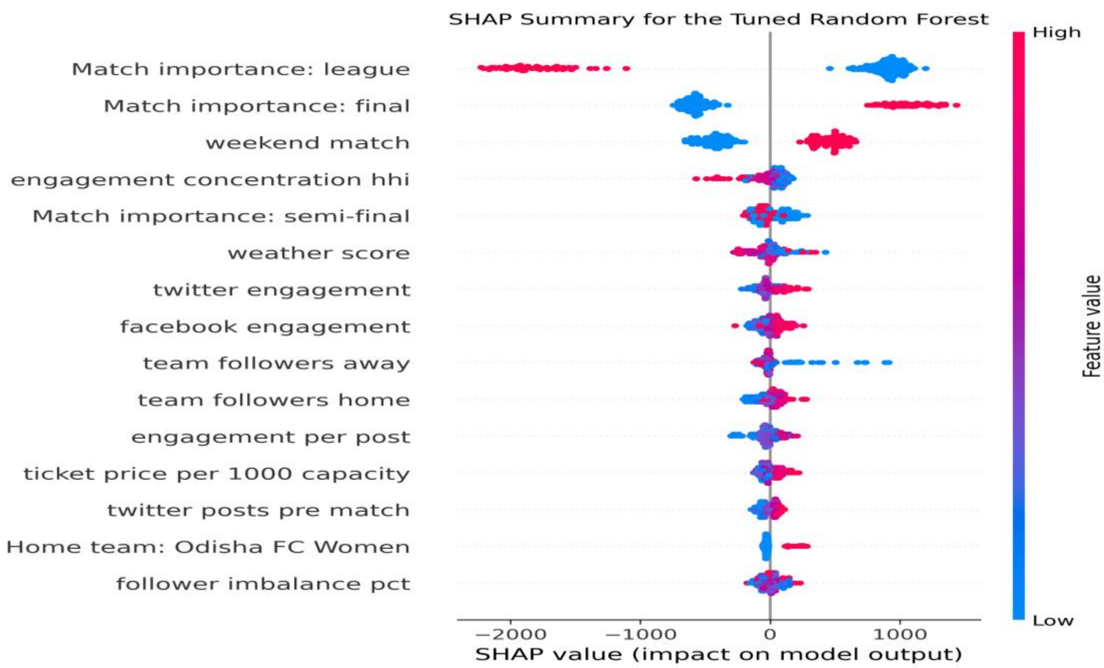
The SHAP distributions are informative in terms of direction. Large negative contributions relative to the prediction baseline were found for league-match membership and large positive contributions were found for final status. The general trend on weekends was to lift predictions. In comparison, the effects of digital engagement variables were smaller and varied, and they tended to improve predictions within a match-context category, but generally were not as prominent as the effects of match importance. This difference is directly predictive of the next ML×PLS result, where the digital constructs have a predictive function but are not structurally determining.

Table 8. Top Predictors by Aggregated Mean Absolute SHAP Value

Predictor	Mean |SHAP|
Match importance	2,008.50
Weekend match	446.70
Engagement concentration HHI	95.86
Home team	86.58
Weather score	83.53
X/Twitter engagement	71.44
Facebook engagement	68.72
Away-team followers	68.39
Home-team followers	68.00
Engagement per post	65.23

DOI: https://doi.org/10.56879/ijbm.v5i3.134

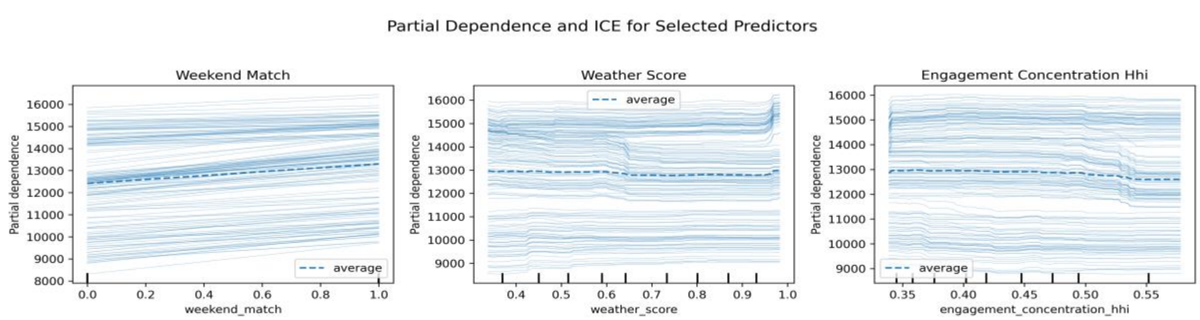
Figure 5. SHAP summary for the tuned Random Forest



4.7 Partial Dependence and Local Explainability

Partial-dependence and ICE analysis helped to gain an additional view of the learned response surface. Weekend status influenced the average predicted attendance to increase as seen in Figure 6, in line with inferential statistics and SHAP ranking. Relationships between weather and engagement concentration were weaker and more varied. ICE trajectories are widely spread, signifying that these variables do not have the same effect on all match records but are influenced by the match context of team, fixture, and digital features. In summary, LIME analyses of high, low, typical, and high-error predictions showed that while a local digital signal can be important in some predictions, in the case of a final, semi-final or league match, the match-context variables seemed to dominate a prediction.

Figure 6. Partial-dependence and ICE profiles for selected predictors



DOI: https://doi.org/10.56879/ijbm.v5i3.134

4.8 Auxiliary High-Attendance Classification

The auxiliary high-attendance task was deliberately made harder with only 131/1000 observations (13.1%) considered to be high attendance (≥70%). Among all the models, the highest ROC-AUC value (0.660) was obtained by SVC and the PR-AUC value of SVC was 0.208 (Table 9). Similar results were obtained with Logistic Regression (ROC-AUC = 0.657; PR-AUC = 0.210). In some cases, the raw accuracy of tree-based classifiers was high due to the dominating majority class but their recall for the high-attendance class was poor. Thus, the results do not support high discrimination, and are indicative of the need to report precision–recall performance in the case of class imbalance.

Table 9. Auxiliary High-Attendance Classification Performance

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
SVC	0.580	0.178	0.615	0.276	0.660	0.208
Logistic Regression	0.585	0.172	0.577	0.265	0.657	0.210
SGD Classifier	0.575	0.176	0.615	0.274	0.644	0.215
XGBoost	0.655	0.188	0.500	0.274	0.631	0.198
Random Forest	0.825	0.091	0.038	0.054	0.628	0.171
CatBoost	0.860	0.000	0.000	0.000	0.618	0.172

4.9 PLS Composite Measurement Results

The Mode-B measurement model converged after 16 iterations. Formative indicator multicollinearity was not problematic; all values were below 1.06, well below the commonly used and conservative threshold of 3.3. Principal outer weights and loadings, and bootstrap results are shown in Table 10. Match Context was the most stable multi-indicator composite, with match-importance score having an outer weight of 0.913 (p<0.001, loading 0.933) and weekend match having an outer weight of 0.360 (p<0.001, loading 0.411). Instagram engagement had a significant bootstrap weight (p=0.044) and X/Twitter engagement, Facebook engagement, and engagement rate had non-significant outer-weight tests within Digital Engagement. These results support the

DOI: https://doi.org/10.56879/ijbm.v5i3.134

formative interpretation, in which indicators represent varying amounts of behavioral information and do not all necessarily have to show high intercorrelations.

The C7 Fan Base block was not so stable. The original away-team follower weight was large and negative, but was broadly distributed over zero in bootstrap, and the original home-team follower weight was smaller but had an important bootstrap result. The construct was kept as a control due to the structural interpretation being approached with caution, however, the number of followers continues to be conceptually relevant and there is not a significant amount of collinearity.

Table 10. Mode-B Outer Weights, Loadings, VIF, and Bootstrap Significance

Construct	Indicator	Weight	Loading	VIF	Bootstrap p
C3	Home-team rank (reversed)	0.779	0.791	1.000	0.024
C3	Away-team rank (reversed)	0.446	0.455	1.000	0.801
C3	Competitive-balance index	-0.420	-0.431	1.000	0.848
C4	Match-importance score	0.913	0.933	1.003	<0.001
C4	Weekend match	0.360	0.411	1.003	<0.001
C5	Instagram posts	0.322	0.382	1.004	0.057
C5	X/Twitter posts	0.554	0.550	1.001	0.248
C5	Facebook posts	0.752	0.761	1.004	0.208
C6	Instagram engagement	0.612	0.669	1.035	0.044

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Construct	Indicator	Weight	Loading	VIF	Bootstrap p
C6	X/Twitter engagement	0.545	0.549	1.008	0.121
C6	Facebook engagement	0.209	0.257	1.014	0.723
C6	Engagement rate	0.409	0.580	1.053	0.146
C7	Home-team followers	0.112	0.087	1.001	0.033
C7	Away-team followers	-0.997	-0.994	1.001	0.945

4.10 Structural Model and Hypothesis Testing

The final PLS structural results are presented in Table 11 and illustrated in Figure 7. 18.0% of the variance in attendance was explained by the model (adjusted $R^2=0.173$). Clear support was found for Match Context, which had a positive coefficient of $\beta=0.418$ (95% bootstrap interval [0.367, 0.466] with $p<0.001$). This coefficient was significantly greater than all other direct paths. The path from Digital Communication to Digital Engagement (H1) was not supported because the PLS bootstrapped point estimate ($\beta=0.129$) was not significant, as the full PLS bootstrap interval crossed zero and the bootstrap p-value was 0.201. Digital Engagement to Attendance (H2) was also positive but small and not significant ($\beta=0.034$, $p=0.357$). H3, H5, H6 and H7 were not supported.

It is important to understand the difference between the positive point estimates and bootstrap decisions. Communication and engagement (digital) is not irrelevant and has positive coefficients in parts of the model, as it helps predict. However, the conclusion is more limited; the dataset is able to provide good structural evidence only for Match Context after the constructs were modeled

DOI: https://doi.org/10.56879/ijbm.v5i3.134

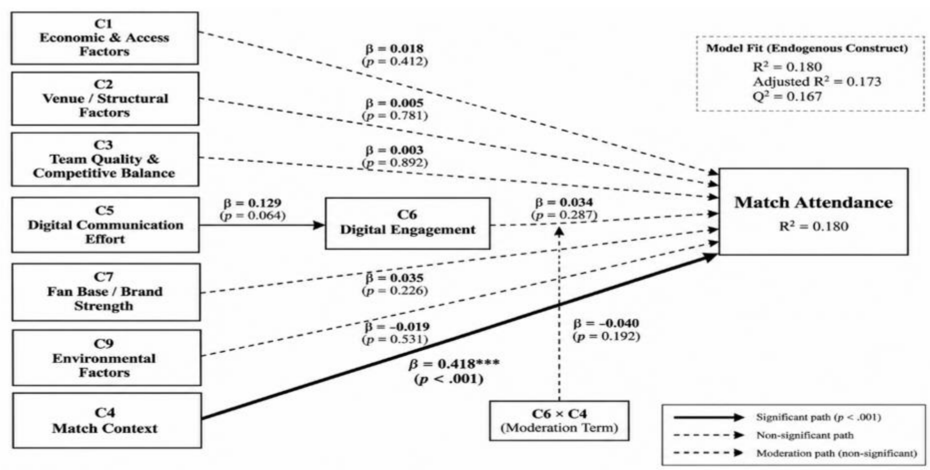
separately and resampling uncertainty was accounted for. This distinction is at the heart of the paper's predictive–explanatory contribution.

Table 11. Final PLS Structural Paths and Hypothesis Decisions

Hyp.	Relationship	β	Bootstrap p	95% CI	Decision
H1	C5 Digital Communication → C6 Digital Engagement	0.129	0.201	[-0.145, 0.201]	Not supported
H2	C6 Digital Engagement → Attendance	0.034	0.357	[-0.063, 0.113]	Not supported
H3	C3 Team Quality/Balance → Attendance	0.003	0.563	[-0.074, 0.080]	Not supported
H4	C4 Match Context → Attendance	0.418	<0.001	[0.367, 0.466]	Supported
H5	C6×C4 → Attendance	-0.040	0.412	[-0.088, 0.044]	Not supported
H6	C1 Economic/Access → Attendance	0.018	0.551	[-0.040, 0.075]	Not supported
H7	C2 Venue/Structural → Attendance	0.005	0.873	[-0.051, 0.062]	Not supported

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Figure 7. Final Composite PLS-SEM Structural Model



4.11 Structural Explanatory and Predictive Power

The value of R^2 for the attendance model was 0.180 and the value of adjusted R^2 was 0.173. The cross-validated predictive relevance was positive (Q^2=0.167), meaning that the set of structural predictors increased the predictive ability compared with a naive mean benchmark. The local effect-size analysis also focused the structural story. Match Context resulted in f2=0.211, while other factors, such as Digital Engagement, Team Quality, Economic/Access, Venue, Fan Base, Environment and the moderation term resulted in f2 values lower than 0.003. Match Context not only had a statistically significant incremental structural contribution in the final specification, but was the only construct with a practically meaningful incremental structural contribution. The 5,000 requested PLS bootstrap resamples were successful, with 4,985 resamples used for inference, which is 99.7% successful resampling. The high ratio of bootstrap completion rate and the positive Q^2 further support the stability of the principal conclusion, but do not mean that the structural model accounts for a great deal of variance in attendance.

4.12 Integrated ML×PLS Results and Robustness

Table 12 integrates construct-level SHAP importance with the corresponding PLS structural evidence. Match Context (C4) was separated from every other construct: it had by far the largest SHAP importance (2,455.2 at the construct-aggregation level) and a significant PLS path (β=0.418, p<0.001), placing it in the Priority Lever quadrant. Digital Communication (C5), Digital

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Engagement (C6), and Fan Base (C7) had above-median predictive importance but non-significant structural paths, placing them in the Predictive but Structurally Uncertain category. C1, C2, C3, and C9 had lower predictive importance and non-significant paths under the present specification.

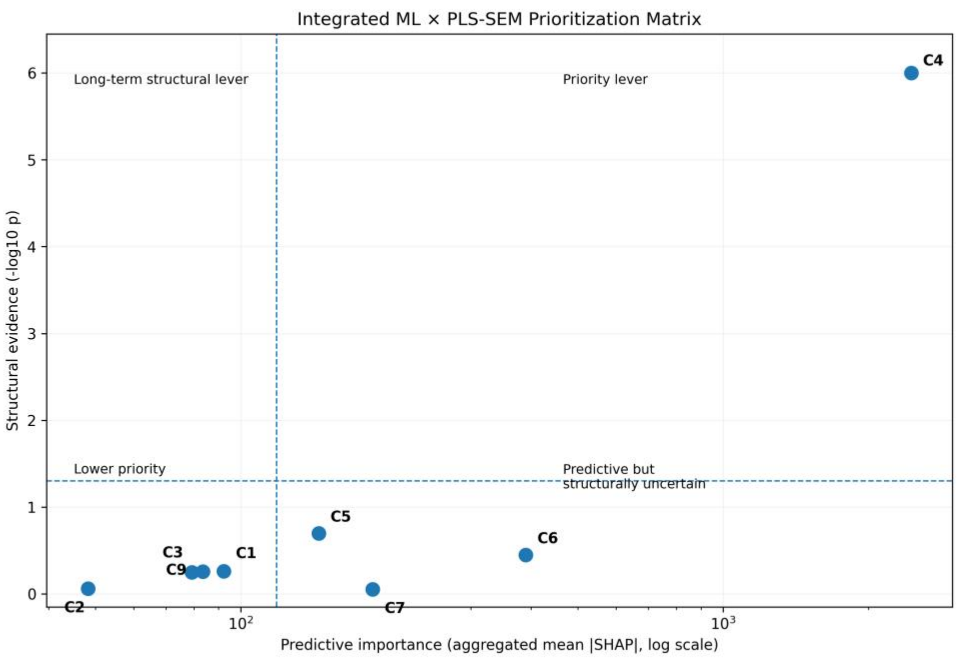
The robustness analyses supported the main model comparison. Mean five-fold CV RMSE was 4,887 for the full 1,000-record dataset and 4,844 for the 892-record subset without repair flags, indicating no deterioration in the stricter subset. The expanded theory-guided 27-predictor feature set also outperformed the core 18-predictor set (RMSE=4,887 versus 5,039). These results suggest that the additional theory-guided features improved predictive usefulness without changing the central finding that match context dominates attendance demand.

Table 12. Integrated Predictive–Structural Priority Framework

Construct	SHAP importance	PLS β	PLS p	Priority category
C1 Economic/Access	92.34	0.018	0.551	Lower priority
C2 Venue/Structural	48.24	0.005	0.873	Lower priority
C3 Team Quality/Balance	79.30	0.003	0.563	Lower priority
C4 Match Context	2,455.20	0.418	<0.001	Priority lever
C5 Digital Communication	145.12	0.129	0.201	Predictive but structurally uncertain
C6 Digital Engagement	389.56	0.034	0.357	Predictive but structurally uncertain
C7 Fan Base	187.83	0.035	0.889	Predictive but structurally uncertain
C9 Environment	83.53	-0.019	0.556	Lower priority

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Figure 8. Integrated ML×PLS-SEM Prioritization Matrix



4.13 Discussion

4.13.1 Match Context as the Dominant Attendance Lever

The most consistent finding in each of the layers of analysis is the centrality of match context. The mean attendance at finals was highest, semi-final match second, and league match third. Weekend games also were more widely attended than weekdays. The descriptive patterns were confirmed by the inferential tests, the Mode-B outer weights of the C4, the PLS structural path and the f^2 effect-size analysis. This finding of a correlation across methods is more persuasive than any one large coefficient in that the same substantive mechanism emerges in the univariate, predictive, and structural analyses.

The discovery is in line with the women's-football literature. The competition stage is one of the important factors in the UEFA Women's Champions League attendance (Valenti et al., 2020). Valenti et al. (2024) found weekend scheduling to be a positive factor in the FA WSL, and Liu et al. (2025) also concluded the same for stadium attendance in the Chinese Women's Super League.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Buraimo et al. (2026) also demonstrate that match significance can have an impact even when there is low short-term outcome uncertainty. The similarities between countries indicate that the convenience and perceived consequence of a fixture may be particularly significant when attempting to engage with women's-football audiences.

The implication is very real for women's football in India. Organizers can offer better attendance increases by offering better fixture timing, weekend concentration and eventization of important games rather than just generic digital activity. Finals and semi-finals are already scarce, and carry enough consequence; league organizers can help sell the perceived importance by themed matchdays, double-headers, rivalry stories, fan events, or even by orchestrating local promotion. While the latter should not be assumed, it is clear that the present model suggests that match context should be evaluated empirically.

4.13.2 Digital Engagement: Predictive Signal Versus Structural Effect

The second big contribution is the separation of the usefulness of the digit and its structural support. Construct Digital Engagement (C6) was highly important and several platform variables were among the most important predictors. However the direct path C6→Attendance was small and not significant in the entire PLS bootstrap. Likewise, the variable Digital Communication (C5) was found to contain valuable predictive value but H1 was not given any boostrap support in the final PLS model. It should not be construed as proof that digital communication isn't important, however. Instead, it points to the fact that for the current match-level model, online activity is a much more stable predictor of audience attention than of actual match attendance as confirmed in itself.

This finding is consistent with Liu et al. (2025) who find that stadium attendance and online streaming of the Chinese Women's Super League are driven by different factors. Digital behaviour can thus add to, rather than replace, or even necessarily become consumption on the match day. This discovery also contributes to the social-media research in football. Vale and Fernandes (2018), Parganas et al. (2017), and Anagnostopoulos et al. (2018) make it clear that the social platforms can be effective environments for engagement and brand management; the present study makes it clear that the transfer from digital engagement to physical attendance depends on conditions and is context specific.

The managerial takeaway is thinking in terms of multiple objectives in digital campaigns. Engagement metrics can still be a helpful tool for understanding where there are audience interest, content resonance and match points that need to be activated. Clubs should not, however, view likes, comments or followers as a direct measure of ticket demand. What makes for a better approach is to mix high digital engagement with site conversion elements – clearly defined matchday calls-to-action, reminders to register and book tickets, links to matchday tickets, localised promotions and event based campaigns around high importance fixtures.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

4.13.3 Competitive Balance and Team Quality

No direct path of significance was found in this dataset to Team Quality and Competitive Balance. This outcome is noteworthy as outcome uncertainty is a central concept in the field of sports economics. The women's-football literature is mixed. Valenti et al. (2020) pointed to an uncertainty effect in the UEFA Women's Champions League, while Buraimo et al. (2026) confirmed no uncertainty effect at a game level in England and France but some in Germany. The result of the present study aligns with a larger trend of uncertainty not being valued the same in all markets.

In an emerging attendance market, one explanation is that subtle differences in competitive balance are not as significant or just as easy to find as the importance and accessibility of the event. If it is a final, fans can pick to go to the final even if the teams are identical. Media interest, digital discussion and perceived prestige may still have a direct effect, but not the current match level model of team quality and competitive balance; the direct effect has not been isolated in the current model. This could be further explored in the future using more detailed team-form, player-quality, betting-odds and historical rivalry metrics.

4.13.4 Comparison with the 2026 Machine-Learning Base Paper

The study contributes to the body of literature on predicting attendance by Beermann et al. (2026). They analysed 2,069 international football matches, compared the three machine learning algorithms, Linear Regression, Random Forest and XGBoost and showed that in a rich UEFA national-team context, machine learning can outperform conventional specifications. This study is different in terms of context and purpose. It provides a platform to look at Indian women's football and includes platform-specific pre-match communication, digital engagement, follower measures, match context and an expanded set of machine-learning families. In addition, nested 5×5 cross-validation, SHAP, LIME, PDP/ICE, a secondary classification task and composite PLS-SEM is used.

The absolute R2 values should not be mechanically compared between the two studies as the data sets, target distributions, competitions, variables and sample structures are significantly different. The novel contribution here is not to suggest a higher predictive accuracy than Beermann et al. (2026), but to expand the question of analysis. Within the integrated framework, the question is whether variables that are useful for predicting attendance are related in the same way to variables that are supported in a structural model. The answer is not always yes and that is exactly what the ML×PLS prioritization framework should show.

5. Conclusion, Implicatons, Limitations and Future Research Directions

This study was aimed at creating an integrated explainable machine-learning and PLS-SEM framework for explaining stadium attendance in Indian women's football. Seven regression algorithms were compared using baseline, tuning, holdout, and nested cross-validation procedures

DOI: https://doi.org/10.56879/ijbm.v5i3.134

and explainability using SHAP, LIME, and partial-dependence/ICE analysis across 1,000 match-level observations. The structural layer had Mode-B composites and a total of 5 000 requested bootstrap resamples to test the relationships between digital communication, engagement, competitive conditions, match context, economic and venue conditions, fan base, environment and attendance guided by the theory.

The empirical main conclusion is obvious. The best and most consistent attendance indicator was match context, especially match importance and weekend match schedules. It was most important in predictive terms, the only significant direct PLS path to attendance was from it, and the largest structural effect size was for it. Digital communication and engagement proved helpful predictive information, but had no direct structural paths that were separately supported after including bootstrap uncertainty. This is meaningful because it provides a way to gauge interest online without ensuring attendance in person.

The resulting ML×PLS prioritization framework shifts the emphasis from prediction to decision support. High feature-importance values for a variable are not the sole reason for prioritizing the variable; nor are the statistically significant structural coefficients the sole reason for prioritizing the variable. Predictive relevance and structural support are two key factors for priority. Match Context is the only factor in the present data that meets both criteria, and thus is the most defensible short-term managerial leverage. A wider point is that the development of women's football's fan base might rely on the synchronicity of online involvement and on the tangible and symbolic requirements that make a game worth watching. For club and league organisers looking to sustainable attendance levels, the most effective approach, therefore, is not a campaign based on digital promotion alone, but digitally informed activation around well-timed, consequential and compelling matchday events.

5.1 Theoretical Contributions

There are three theoretical contributions. Second, the study is in an environment where the development of sport-demand is still underway and where the convenience of a match and the importance of an event can be more important than the marginal effects noted in more established sport leagues. The strong C4 result implies that the decision to attend a game is strongly affected by the context: the match importance and weekend scheduling work together in influencing attendance, whereas price, venue size, competitive balance and weather have a lesser impact in the current model.

Second, the study does not assume but rather tests the S-O-R-inspired sequence from digital communication to digital engagement to attendance. The weak bootstrap support for the digital structural paths enhances the theoretical narrative. Digital communication and engagement can be as attention and relationship operations without necessarily resulting in direct one-to-one conversion to physical consumption. This helps to clarify a multi-channel conception of fandom in which “going online” and “going to the stadium” are somewhat separable but related outcomes.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Third, the study provides a predictive–structural distinction. The traditional econometric or SEM studies can prove association under a certain theory, but the research has limited predictability; machine learning can make very good prediction, but it may not be easy to get variables that can be structurally interpreted. This is a practical synthesis of SHAP and PLS-SEM. Only if a construct is predictively important and structurally supported is it given the strongest managerial status. This helps to minimize the risk of erroneous interpretation of the blackbox importance scores or of using effects with a low predictive power but high statistical significance.

5.2 Managerial implications

For league organizers, the results focus on fixture design and fixture scheduling. The most consistent attendance levers are weekend scheduling and perceived importance of a match. Whenever possible, high-potential fixtures should be scheduled in easy-to-reach weekend windows, and a large amount of advance promotion and packaging of the event should be used. Standings narrative, qualification implications, rivalry framing, and coordinated matchday experiences are other ways in which league communications can add salience to regular season matches.

Digital metrics can be used as targeting and diagnostic signals for club marketing teams. While high engagement can help to target audiences and fixtures that are showing interest, conversion calls for a matchday proposition. Reach, engagement, click-through, ticket action and actual attendance should therefore be separated in campaign evaluation. Content can be customized for the various purposes of each platform: Instagram as a visual identity, X/Twitter for real-time content and engagement, Facebook for social reach, and direct ticket links for conversions.

When applied as a probabilistic decision support system, rather than an exact point forecast, even modest predictive models can play a role in venue operations for staffing and stewarding, concessions and seating allocations, and fan-experience planning. The ML×PLS model suggests digital engagement is still important for getting people to pay attention, even if it's not a good direct attendance model. This opens up opportunities to run campaigns that feature both digital and match context activation – particularly for big events.

5.3 Limitations and Future Research

There are several limitations in interpreting the findings. Firstly the study is situated in one national women's football context and one collated match level analytical dataset. The magnitudes of the coefficients should thus be checked in other Indian competitions, seasons and countries before they can be generalized. Second, the data are cross-sectional at the match level and do not lend themselves to causation. Structural coefficients are associations that are consistent with a given model, and causal effects would require experimental, quasi-experimental, or stronger longitudinal identification strategies. Third, the predictive accuracy is modest. A holdout R^2 of approximately 0.155 and structural R^2 of 0.180 indicate that important determinants of attendance remain outside

DOI: https://doi.org/10.56879/ijbm.v5i3.134

the current feature set. Future work should add chronology, season/round dynamics, historical attendance, team form, player availability, ticket sales trajectories, travel distance, local events, broadcast exposure, sentiment from actual social content, and campaign-level spending. Temporal validation would be particularly valuable because random folds can still share team and venue patterns across training and validation samples.

Fourth, the digital constructs use behavioral counts rather than perceptual measures. Engagement counts can indicate attention but cannot directly measure fan identification, perceived content quality, attendance intention, or psychological attachment. A complementary spectator survey would allow genuinely reflective psychological constructs to be modeled alongside the behavioral match-level composites. Fifth, the venue-capacity field was harmonized during dataset preparation and should be interpreted as an analysis-level venue descriptor rather than an independently audited official-capacity measure. Capacity-derived outcomes were therefore not used as predictors of raw attendance.

Future research can build on the present framework in three directions. The first is longitudinal prediction using multiple seasons and time-aware cross-validation. The second is multi-source data integration that joins ticket transactions, media exposure, social content, and local socioeconomic context at the match level. The third is causal evaluation of actionable levers, such as weekend scheduling, promotional intensity, or matchday packages, using natural experiments or staggered interventions. These extensions would help determine whether the strong contextual pattern observed here represents a causal opportunity for attendance growth.

Declarations

Data availability statement

Authors will keep the analytical data set and code that support their reported analyses and will be willing to share it based on the data-sharing policies of the proposed target journal and any restrictions from the sources. The manuscript includes all necessary model specifications and principal numerical results to replicate the analytical logic.

Funding

No external funding is declared for this manuscript unless otherwise specified by the authors at submission.

Conflict of interest

The authors declare no conflict of interest unless otherwise specified at submission.

Ethics statement

DOI: https://doi.org/10.56879/ijbm.v5i3.134

The study analyzes match-level, team-level, and digital behavioral data and does not report personally identifiable participant information. Any additional institutional ethics requirements should be completed according to the target journal and author affiliation before submission.

References

Anagnostopoulos, C., Parganas, P., Chadwick, S., & Fenton, A. (2018). Branding in pictures: Using Instagram as a brand management tool in professional team sport organisations. European Sport Management Quarterly, 18(4), 413–438. https://doi.org/10.1080/16184742.2017.1410202

Beermann, S., Dallmeyer, S., & Breuer, C. (2026). Cheering for your country: The prediction of attendance at international football games using machine learning techniques. Applied Economics. Advance online publication. https://doi.org/10.1080/00036846.2026.2618091

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

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

Buraimo, B., Farnell, A., Hargreaves, J., & Simmons, R. (2026). Outcome uncertainty and the demand for women's football. Sports Economics Review, 15, 100074. https://doi.org/10.1016/j.serev.2026.100074

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297. https://doi.org/10.1007/BF00994018

Elstak, I., Salmon, P., & McLean, S. (2024). Artificial intelligence applications in the football codes: A systematic review. Journal of Sports Sciences, 42(13), 1184–1199. https://doi.org/10.1080/02640414.2024.2383065

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, 61(1), 101–107.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). A primer on partial least squares structural equation modeling (PLS-SEM) (3rd ed.). SAGE.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. European Business Review, 31(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. Advances in International Marketing, 20, 277–319. https://doi.org/10.1108/S1474-7979(2009)0000020014

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

Kunkel, T., Doyle, J. P., & Funk, D. C. (2014). Exploring sport brand development strategies to strengthen consumer involvement with the product: The case of the Australian A-League. Sport Management Review, 17(4), 470–483. https://doi.org/10.1016/j.smr.2014.01.004

Liu, T., Li, S., Han, Z., & Niu, H. (2025). Distinct drivers of stadium attendance and online streaming: Evidence from the Chinese Women's Super League. Frontiers in Sports and Active Living, 7, 1695179. https://doi.org/10.3389/fspor.2025.1695179

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Mehrabian, A., & Russell, J. A. (1974). An approach to environmental psychology. MIT Press.

Parganas, P., Anagnostopoulos, C., & Chadwick, S. (2017). Effects of social media interactions on brand associations: A comparative study of soccer fan clubs. International Journal of Sports Marketing and Sponsorship, 18(2), 149–165. https://doi.org/10.1108/IJSMS-05-2017-087

Pawlowski, T., & Anders, C. (2012). Stadium attendance in German professional football: The (un)importance of uncertainty of outcome reconsidered. Applied Economics Letters, 19(16), 1553–1556. https://doi.org/10.1080/13504851.2011.639725

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?' Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Schreyer, D., & Ansari, P. (2022). Stadium attendance demand research: A scoping review. Journal of Sports Economics, 23(6), 749–788. https://doi.org/10.1177/15270025211000404

Scelles, N. (2021). Policy, political and economic determinants of the evolution of competitive balance in the FIFA women's football World Cups. International Journal of Sport Policy and Politics, 13(2), 281–297. https://doi.org/10.1080/19406940.2021.1898445

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B, 36(2), 111–147.

Valenti, M., Scelles, N., & Morrow, S. (2020). The determinants of stadium attendance in elite women's football: Evidence from the UEFA Women's Champions League. Sport Management Review, 23(3), 509–520. https://doi.org/10.1016/j.smr.2019.04.005

Valenti, M., Scelles, N., & Morrow, S. (2024). The determinants of stadium attendance in elite women's football: Evidence from the FA Women's Super League. European Sport Management Quarterly, 25(2), 322–338. https://doi.org/10.1080/16184742.2024.2343485

Vale, L., & Fernandes, T. (2018). Social media and sports: Driving fan engagement with football clubs on Facebook. Journal of Strategic Marketing, 26(1), 37–55. https://doi.org/10.1080/0965254X.2017.1359655

Glossary

AI: Artificial intelligence.

ANOVA: Analysis of variance.

CatBoost: A gradient-boosting algorithm designed to handle categorical and nonlinear tabular structure.

FDR: False discovery rate; a multiple-testing error-control procedure.

ICE: Individual conditional expectation; a plot showing the modeled response to a predictor for individual observations.

DOI: https://doi.org/10.56879/ijbm.v5i3.134

IQR: Interquartile range; the difference between the third and first quartiles.

LIME: Local Interpretable Model-agnostic Explanations; a method for explaining individual model predictions.

MAE: Mean absolute error; the average absolute difference between observed and predicted values.

ML: Machine learning.

Mode-B composite: A formative composite estimated from weighted indicators rather than interchangeable reflective indicators.

PLS-SEM: Partial least squares structural equation modelling.

PR-AUC: Area under the precision–recall curve.

Q^2: Cross-validated predictive relevance statistic.

R^2: Coefficient of determination; the proportion of outcome variance explained by a model.

RMSE: Root mean squared error.

ROC-AUC: Area under the receiver operating characteristic curve.

SGD: Stochastic gradient descent.

SHAP: Shapley Additive exPlanations; additive feature-contribution values for model interpretation.

S-To-R: Stimulus–Organism–Response framework.

SVR: Support vector regression.

VIF: Variance inflation factor; a diagnostic for multicollinearity.

XGBoost: Extreme Gradient Boosting; a regularized gradient-boosting implementation.

Appendix 1. Additional robustness evidence

The main text reports the primary holdout, nested cross-validation, PLS-SEM, and ML×PLS results. The following supplementary summaries preserve additional evidence used to assess model stability.

Supplementary Table S1. Robustness of cross-validated RMSE

Analysis	N / feature set	Mean CV RMSE	SD RMSE
Full dataset	1,000	4,886.97	166.64

DOI: https://doi.org/10.56879/ijbm.v5i3.134

Analysis	N / feature set	Mean CV RMSE	SD RMSE
Records without repair flags	892	4,844.47	131.37
Core original predictors	18 predictors	5,039.38	193.22
Expanded theory-guided predictors	27 predictors	4,886.97	166.64

Supplementary Table S2. Fold-wise model-comparison tests

Comparison	Mean RMSE difference	Paired t-test p	Wilcoxon p
SGD vs CatBoost	16.80	0.677	1.000
SGD vs Ridge	18.09	0.309	0.313
SGD vs Linear Regression	22.22	0.251	0.313
SGD vs XGBoost	61.65	0.185	0.063
SGD vs Random Forest	96.22	0.027	0.063
SGD vs SVR	211.90	0.059	0.063

DOI: https://doi.org/10.56879/ijbm.v5i3.134