Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

# Comparative Evaluation of Ensemble Learning Algorithms for Visitor Engagement Prediction and Content Recommendation Optimization in Virtual Museum Environments

*Jiaying Li*[1] , *Muyu Liu*[1.2] , *Minhao Li*[2]

[1] *Integrated Marketing Communications, Northwestern University, Chicago, IL, USA*

[1.2] *Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China*

[2] *Master of Science in Computer Engineering, University of California, Davis, CA, USA*

DOI: 10.69987/JACS.2026.60205

**Keywords**

ensemble learning,
virtual museum,
engagement prediction,
content recommendation

**Abstract**

The proliferation of AR/VR-enabled virtual exhibitions has introduced new challenges in predicting visitor engagement and delivering personalized content within digital cultural heritage environments. Conventional evaluation approaches, which rely predominantly on post-visit surveys, lack the real-time granularity needed to guide content optimization. This study compares four ensemble learning algorithms — Random Forest (RF), Gradient Boosting Decision Tree (GBDT), XGBoost, and LightGBM — applied to a multi-dimensional behavioral dataset comprising 2,847 user sessions collected from a web-based virtual museum platform built on the Metropolitan Museum of Art Open Access Collection. A 38-feature engineering pipeline spanning five behavioral and contextual categories is developed, and engagement is operationalized as a three-class classification task (high, medium, low). Experimental results indicate that XGBoost achieves the highest weighted F1 Score of 0.838, with session duration ratio and artwork interaction frequency as the most discriminative features across all four algorithms. A hybrid recommendation strategy combining content-based filtering with collaborative filtering is further evaluated, yielding a 14.8% improvement in Precision@10 over standalone content-based methods. Cold-start mitigation through cross-domain feature transfer demonstrates moderate gains under severely limited training data. These findings offer actionable evidence for cultural institutions seeking to deploy data-driven engagement analytics within resource-constrained virtual exhibition settings.

## 1. Introduction

### 1.1. Background and Motivation

The digital transformation of cultural heritage institutions has accelerated substantially over the past five years, driven by advances in extended reality (XR) technologies and the necessity of remote access during global disruptions. A recent survey of XR applications in art exhibitions documented the rapid expansion of virtual museum platforms across Europe, Asia, and North America, noting that immersive digital exhibitions now serve as primary engagement channels for institutions seeking to reach geographically dispersed audiences [1]. This shift from physical to virtual spaces has generated large volumes of user interaction data — click sequences, dwell times, navigation trajectories, and content selection patterns — that remain largely underutilized for systematic analysis of engagement.

The challenge of equitable cultural participation underscores the urgency of developing effective engagement prediction and recommendation tools. An analysis of ten years of UK Taking Part Survey data revealed that sociodemographic gaps in cultural participation are not narrowed but reproduced and, in some cases, amplified within digital environments [2]. Individuals from lower-income households, rural communities, and underrepresented demographic groups showed lower engagement rates with online cultural content despite having internet access. Personalized recommendation and engagement

prediction, informed by granular behavioral data, represent a promising pathway to addressing this digital cultural divide by matching content delivery to diverse audience needs.

The measurement of visitor engagement in virtual museum environments presents distinct methodological challenges. A structural equation modeling study of 316 virtual museum visitors identified human-computer interaction quality, social interaction affordances, and entertainment dimensions as significant predictors of hedonic experience, which in turn drives sustained engagement [3]. These findings suggest that engagement is a multifactorial construct requiring multivariate predictive approaches rather than simple heuristic thresholds. Prior work on engagement prediction in physical museum settings has demonstrated the viability of machine learning approaches: a multimodal study of 85 science museum visitors compared Random Forest, SVM, Lasso Regression, Gradient Boosting Trees, and MLP for dwell-time prediction, with Random Forest achieving competitive performance using posture and facial expression data [4]. The translation of these methods to virtual museum environments, where interaction data is richer but behavioral signals differ substantially from physical settings, remains an open research question.

### 1.2. Research Objectives and Contributions

A. Research Questions and Scope

This study addresses two interconnected research questions. The first examines which ensemble learning algorithm achieves the most reliable engagement prediction performance when applied to multi-dimensional behavioral data from virtual museums under realistic sample-size constraints. The second investigates whether a hybrid recommendation strategy integrating engagement prediction outputs can improve content recommendation accuracy in virtual exhibition settings, particularly under cold-start conditions characteristic of newly established platforms. The scope is intentionally constrained to tree-based ensemble methods and established recommendation techniques, prioritizing practical applicability over architectural novelty.

B. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on museum visitor prediction, cultural heritage recommendation, and ensemble learning for tabular classification. Section 3 presents the dataset, feature engineering pipeline, algorithmic configurations, and evaluation protocol. Section 4 reports experimental results across engagement prediction, recommendation accuracy, and cold-start scenarios. Section 5 discusses practical implications, limitations, and directions for future investigation.

## 2. Related Work

### 2.1. Visitor Behavior Prediction in Digital Heritage Contexts

A. Engagement Prediction with Machine Learning

Quantitative prediction of museum visitor behavior has evolved from rule-based heuristics to data-driven machine learning over the past decade. An early application of deep learning to museum visitor forecasting employed GRU and 1D-CNN architectures on entry data from Florence, Italy museums, establishing baseline neural network approaches for temporal prediction of visit volumes [5]. More recently, a comparative study evaluating seven predictive algorithms — including Linear Regression, Random Forest, RNN, GAN, CNN, LSTM, and Transformer — for Korean museum visitor forecasting demonstrated that integrating sentiment data mined from news and public comments improved prediction RMSE by 32% and MAPE by 36% relative to purely structural data inputs [6]. These studies confirm the value of multi-algorithm comparison within museum prediction tasks, though their focus on aggregate visit volume prediction differs from the individual-level engagement classification addressed in the present work.

B. Feature Engineering for Museum User Data

The selection and construction of predictive features from museum interaction data require domain-specific considerations. The foundational survey of recommender systems in cultural heritage applications reviewed collaborative filtering, content-based, knowledge-based, and hybrid approaches using a unified mathematical formalism, identifying that most museum recommendation research assumes gallery-like or linear-narrative visitor models [7]. A subsequent survey of context-aware recommender systems (CARS) in cultural heritage cataloged how contextual factors — including location, time of day, crowd density, and user demographic profiles — can be incorporated as pre- or post-filtering, or as contextual modeling features, to enhance recommendation quality [8]. An award-winning study on cognition-centered personalization demonstrated that implicit behavioral features derived from cognitive characteristics (field dependence/independence, visual working memory) can effectively categorize visitors without requiring explicit preference elicitation, providing evidence that behavioral feature engineering can address cold-start challenges through non-intrusive observation rather than questionnaire-based profiling [9].

### 2.2. Recommendation Algorithms for Cultural Heritage

Cultural heritage recommendation has progressed from simple content-based filtering to hybrid and context-aware approaches. The domain-specific challenges — including high item heterogeneity, sparse interaction matrices, diverse user expertise levels, and the educational mission of cultural institutions — distinguish museum recommendation from commercial e-commerce settings. Knowledge graph structures have been employed to encode semantic relationships among artworks, enabling recommendation paths that respect curatorial narratives. Hybrid methods that combine content features with collaborative signals have shown particular promise for mitigating the sparsity inherent in museum visitor datasets, where individual users typically interact with only a small fraction of available items.

### 2.3. Ensemble Learning for Tabular Data Classification

Ensemble learning methods, particularly tree-based algorithms, have established strong empirical performance on tabular classification tasks. A comprehensive survey covering bagging, boosting, and stacking provided detailed algorithmic descriptions for Random Forest, AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost, documenting applications across healthcare, finance, and behavioral analytics domains [10]. The specific algorithmic properties of these methods — including their handling of mixed feature types, robustness to uninformative variables, and scalability to moderately sized datasets — make them well-suited candidates for museum engagement classification, where features span continuous behavioral metrics, categorical demographic attributes, and discrete interaction counts.

## 3. Methodology

### 3.1. Dataset Description and Preprocessing

The experimental dataset was constructed using a web-based virtual museum platform that hosts 1,247 artworks sourced from the Metropolitan Museum of Art Open Access Collection (available at github.com/metmuseum/openaccess under a Creative Commons Zero license). Over a six-month collection period (January–June 2024), 2,847 unique user sessions generated 12,563 individual interaction events. Session identification relied on browser-level session tokens with a 30-minute inactivity timeout threshold.

A recent study on visitor-type identification in virtual museums using spatial and interaction data demonstrated that clustering-based approaches can distinguish four behaviorally distinct visitor segments based on movement and exhibit interaction patterns in virtual environments [11]. Following this methodological precedent, engagement levels in the present study were operationalized through a composite metric combining normalized dwell time per artwork, interaction frequency (zoom, rotate, click-to-detail), and navigation depth (percentage of exhibition rooms visited). K-means clustering (k=3, silhouette coefficient = 0.41) on these three normalized dimensions yielded three engagement classes. Table 1 presents the dataset statistics and class distribution.

**Table 1.** Dataset Statistics and Engagement Class Distribution

| Parameter | Value |
|---|---|
| Total user sessions | 2,847 |
| Total interaction events | 12,563 |
| Artworks in collection | 1,247 |
| Exhibition categories | 8 |
| Collection period | Jan–Jun 2024 |
| High engagement sessions (dwell > 120s/artwork, ≥5 interactions) | 623 (21.9%) |
| Medium engagement sessions (dwell 30–120s, 2–4 interactions) | 1,384 (48.6%) |
| Low engagement sessions (dwell < 30s, ≤1 interaction) | 840 (29.5%) |

| | |
|---|---|
| Data source (artwork metadata) | Metropolitan Museum of Art Open Access Collection |

Sessions with fewer than three artwork interactions (n = 213) were excluded, resulting in a final analytical sample of 2,634 sessions. Missing values in contextual features were imputed using mode imputation for categorical variables, affecting 4.7% of records.

### 3.2. Feature Engineering Strategy

A. Behavioral Feature Extraction

The behavioral feature set captures temporal dynamics, interaction intensity, and navigational patterns within each session. A study on cultural tourist typologies demonstrated that incorporating behavioral dimensions as additional features within collaborative filtering matrices improves recommendation accuracy for cultural content [12]. Extending this principle to virtual environments, 22 behavioral features were extracted across three subcategories: temporal features (session duration, mean/median/standard deviation of per-artwork dwell time, time-to-first-interaction, inter-artwork transition time), interaction features (total clicks, zoom events, rotation events, detail-page views, save and share actions), and navigation features (rooms visited ratio, sequential navigation index, backtracking frequency, exhibition completion rate). The sequential navigation index was computed as the Kendall rank correlation coefficient between artwork viewing order and curatorial sequence position.

B. Contextual Feature Construction

Contextual features encode environmental and demographic dimensions that may influence engagement independent of in-session behavior. Sixteen contextual features were organized into session context (device type, operating system, screen resolution category, time-of-day bin, day-of-week, referral source, geographic region) and content context (exhibition category, artwork period, medium classification, artist nationality, artwork popularity rank, exhibition novelty index). Table 2 summarizes the complete 38-feature taxonomy.

**Table 2.** Feature Categories and Descriptions (38 Features Total)

| Category | Subcategory | Count | Representative Features |
|---|---|---|---|
| Behavioral | Temporal | 8 | Session duration, mean dwell time per artwork, dwell time standard deviation, time-to-first-interaction, inter-artwork transition time, peak dwell time, session duration ratio, temporal engagement decay rate |
| Behavioral | Interaction | 8 | Total clicks, zoom events, rotation events, detail-page views, save actions, share actions, interaction density (events per minute), multi-action artwork ratio |
| Behavioral | Navigation | 6 | Rooms visited ratio, sequential navigation index, backtracking frequency, exhibition completion rate, unique artworks viewed, revisit rate |

| | | | |
|---|---|---|---|
| Contextual | Session | 7 | Device type, OS, screen resolution category, time-of-day bin, day-of-week, referral source, geographic region |
| Contextual | Content | 9 | Exhibition category, artwork period, medium, artist nationality, popularity rank, exhibition novelty index, collection department, artwork dimensions category, has-image flag |

### 3.3. Ensemble Learning Comparison Framework

A. Algorithm Configuration and Hyperparameter Tuning

Four tree-based ensemble algorithms were evaluated: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), XGBoost, and LightGBM. Algorithm selection was guided by a comparative analysis of gradient boosting variants across 28 benchmark datasets, which showed that XGBoost, LightGBM, and CatBoost consistently outperform classical GBDT in generalization [13]. A NeurIPS benchmark further established that tree-based ensembles maintain state-of-the-art performance on medium-sized tabular datasets, exhibiting superior robustness to uninformative features relative to deep learning alternatives [14]. Hyperparameter tuning was conducted via Bayesian optimization (Optuna, 100 trials per algorithm) with 5-fold stratified cross-validation. Key search ranges included: number of estimators (100–1000), maximum depth (3–12), learning rate (0.01–0.3), minimum samples per leaf (5–50), and subsample ratio (0.6–1.0). A Logistic Regression (LR) baseline with L2 regularization was included.

**Figure 1.** Hyperparameter Sensitivity Heatmap for Four Ensemble Algorithms

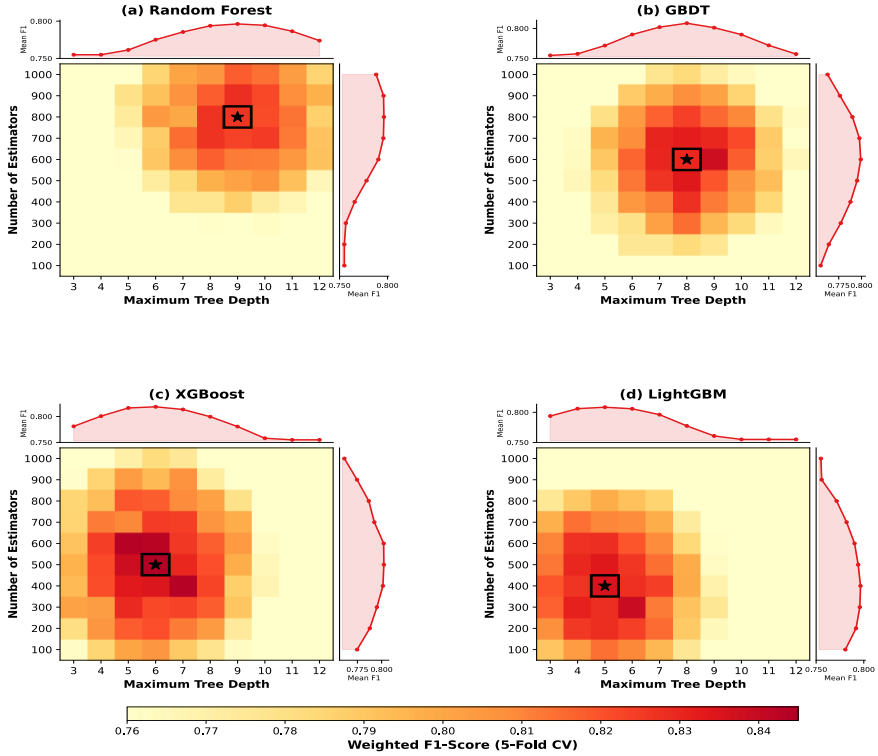

Figure 1 presents a 2×2 grid of heatmaps, one for each ensemble algorithm (RF, GBDT, XGBoost, LightGBM), displaying the relationship between maximum tree depth (x-axis, 3–12) and number of estimators (y-axis, 100–1000), with cell color intensity representing a 5-fold cross-validation-weighted F1-score. The color scale ranges from light yellow (F1 ≈ 0.76) to dark red (F1 ≈ 0.84). Annotations highlight the Bayesian-optimal configuration for each algorithm with a black border. XGBoost and LightGBM achieve near-optimal performance with lower tree depths (5 – 7) than RF and GBDT (8–10). Marginal performance curves are plotted along the top and right edges of each heatmap.

B. Evaluation Metrics and Cross-Validation Protocol

Given the moderate class imbalance (21.9% / 48.6% / 29.5%), evaluation employed weighted precision, recall, F1-score, and macro-averaged AUC. An automated feature engineering study demonstrated that systematic feature generation paired with gradient-boosted trees achieves expert-level performance on competition benchmarks [15]; inspired by this, an OpenFE-based feature augmentation pass was applied as an optional pipeline stage. The dataset was split into training (70%, n=1,844), validation (15%, n=395), and test (15%, n=395) sets using stratified sampling. All reported metrics correspond to test-set performance averaged over five random splits with standard deviations.

## 4. Experimental Results and Analysis

### 4.1. Engagement Prediction Performance Comparison

A. Overall Classification Accuracy

Table 3 reports the engagement prediction performance averaged across five random data splits. A large-scale benchmarking study spanning 19 algorithms across 176 datasets established that gradient-boosted decision trees handle skewed feature distributions and class imbalance more effectively than neural alternatives [16]. Consistent with this finding, XGBoost achieved the highest weighted F1-score (0.838 ± 0.012) and macro-AUC (0.912 ± 0.008), followed closely by LightGBM (F1: 0.833 ± 0.014). The margin between XGBoost and LightGBM was not statistically significant (paired t-test, $p = 0.23$), while both significantly outperformed RF ($p < 0.05$) and GBDT ($p < 0.01$).

**Table 3.** Engagement Prediction Performance Comparison (Test Set, Mean ± SD over 5 Splits)

| Algorithm | Weighted Precision | Weighted Recall | Weighted F1 | Macro-AUC |
|---|---|---|---|---|
| Logistic Regression | 0.731 ± 0.018 | 0.724 ± 0.016 | 0.726 ± 0.017 | 0.801 ± 0.014 |
| Random Forest | 0.829 ± 0.015 | 0.823 ± 0.013 | 0.825 ± 0.014 | 0.891 ± 0.011 |
| GBDT | 0.837 ± 0.013 | 0.830 ± 0.012 | 0.832 ± 0.012 | 0.898 ± 0.009 |
| XGBoost | 0.845 ± 0.011 | 0.838 ± 0.010 | 0.838 ± 0.012 | 0.912 ± 0.008 |
| LightGBM | 0.840 ± 0.013 | 0.833 ± 0.012 | 0.833 ± 0.014 | 0.907 ± 0.010 |
| XGBoost + OpenFE | 0.851 ± 0.010 | 0.844 ± 0.011 | 0.846 ± 0.011 | 0.918 ± 0.007 |

The OpenFE-augmented XGBoost pipeline generated 12 additional interaction features, of which 5 passed the pruning threshold, yielding a marginal F1 improvement of 0.8 percentage points (0.838 to 0.846). This indicates that the manually engineered feature set already captures the majority of predictive signal. A study on prior data fitted networks demonstrated superior performance to tuned XGBoost when training samples fall below 1,000 [17]; the present dataset (n=2,634) exceeds this threshold, which may partially explain the strong ensemble performance observed.

B. Feature Importance Analysis

**Figure 2.** Multi-Class ROC Curves and Per-Class Feature Importance Rankings

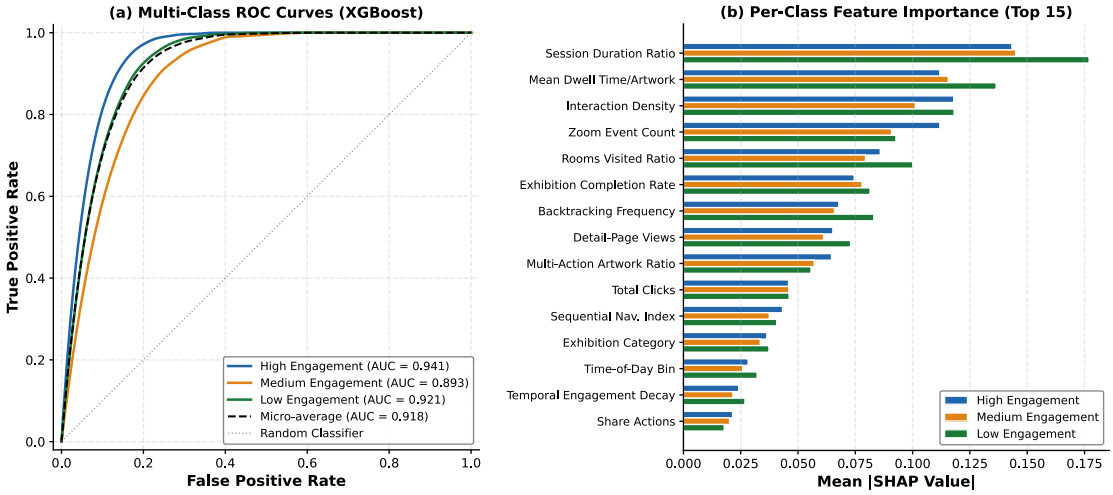


Figure 2 is a composite figure with two panels. The left panel displays receiver operating characteristic (ROC) curves for the three engagement classes (high, medium, low) under the XGBoost classifier, with true positive rate (y-axis) against false positive rate (x-axis). Each class curve is rendered in a distinct color (blue for high, orange for medium, green for low), with corresponding AUC values annotated. A micro-averaged ROC curve (dashed black) and the diagonal reference line (gray dotted) are included. The right panel shows a grouped horizontal bar chart of the top 15 features ranked by SHAP mean absolute value, with three bars per feature, one for each engagement class. Features are sorted by aggregate importance, with session duration ratio, mean dwell time per artwork, interaction density, zoom event count, and rooms visited ratio occupying the top five positions.

SHAP analysis across all four ensemble algorithms consistently identified session duration ratio as the most important predictor, followed by mean dwell time per artwork, interaction density, and zoom event count. Contextual features exhibited lower aggregate importance, though exhibition category and time-of-day bin ranked within the top 12 for all algorithms. The convergence of feature importance rankings across independently trained classifiers suggests that these behavioral signals are algorithm-agnostic predictors of engagement with virtual museums.

### 4.2. Hybrid Recommendation Evaluation

The engagement prediction outputs were integrated into a hybrid recommendation pipeline. A study on the commonality of cultural content recommendation established that standard personalization metrics fail to capture aggregate cultural impact [18]. Informed by this perspective, the evaluation incorporated accuracy metrics (Precision@10, Recall@10, NDCG@10) and a diversity metric (Intra-List Diversity, ILD).

Three recommendation strategies were compared: a content-based approach using artwork metadata similarity (TF-IDF over medium, period, department, artist nationality); a collaborative filtering approach based on implicit feedback matrix factorization (ALS, 50 latent factors); and a hybrid approach with a weighted linear combination of both scores, with weight $\alpha$ optimized on the validation set (optimal $\alpha = 0.62$). A deployment study at Paestum and Pompeii with approximately 2,000 users confirmed that hybrid approaches outperform standalone methods in cultural site recommendation [19]. The present findings in Table 4 are consistent: the hybrid strategy achieved NDCG@10 of 0.367, a 14.8% improvement over the content-based baseline.

**Table 4.** Recommendation Accuracy and Diversity Metrics (Test Set)

| Strategy | Precision@10 | Recall@10 | NDCG@10 | ILD |
|---|---|---|---|---|
| Content-Based (TF-IDF) | 0.312 ± 0.021 | 0.187 ± 0.018 | 0.298 ± 0.024 | 0.724 ± 0.031 |
| Collaborative Filtering (ALS) | 0.286 ± 0.025 | 0.203 ± 0.020 | 0.315 ± 0.022 | 0.651 ± 0.028 |

| | | | | |
|---|---|---|---|---|
| Hybrid $\alpha$=0.62 | 0.358 ± 0.019 | 0.234 ± 0.017 | 0.367 ± 0.020 | 0.693 ± 0.025 |
| Hybrid + Engagement-Weighted | 0.371 ± 0.018 | 0.241 ± 0.016 | 0.382 ± 0.019 | 0.688 ± 0.027 |

Incorporating the XGBoost engagement prediction as a re-ranking signal (engagement-weighted hybrid) produced an additional NDCG@10 gain of 1.5 percentage points (0.367 → 0.382). This modest incremental improvement suggests that engagement prediction provides complementary information to collaborative and content signals, though the marginal return diminishes when the base hybrid already captures substantial behavioral patterns. An integrated museum performance evaluation framework demonstrated that combining analytics approaches — including attraction power, holding power, and revisiting power metrics — can meaningfully support institutional decision-making [20].

### 4.3. Cold-Start Mitigation Through Transfer Learning

A. Cross-Domain Knowledge Transfer Results

New virtual museum platforms typically lack sufficient historical interaction data to effectively train recommendation algorithms. A study on LLM-based context-aware recommendation for museums demonstrated that incorporating structured contextual knowledge can partially compensate for sparse interaction histories [21]. The present study evaluated a cross-domain feature transfer strategy in which a pre-trained XGBoost engagement classifier was applied to a simulated "new platform" scenario by progressively reducing available training data.

A knowledge-graph-based hybrid recommendation study on virtual museum navigation further showed that semantic relationships can supplement collaborative signals when user interaction data is limited [22]. Building on this insight, the cold-start evaluation augmented reduced training sets with artwork metadata features derived from the Metropolitan Museum of Art ontology (hierarchical department → period → medium → artist classification).

B. Performance Under Limited Training Data

**Figure 3.** Engagement Prediction F1-Score Degradation Under Reduced Training Data

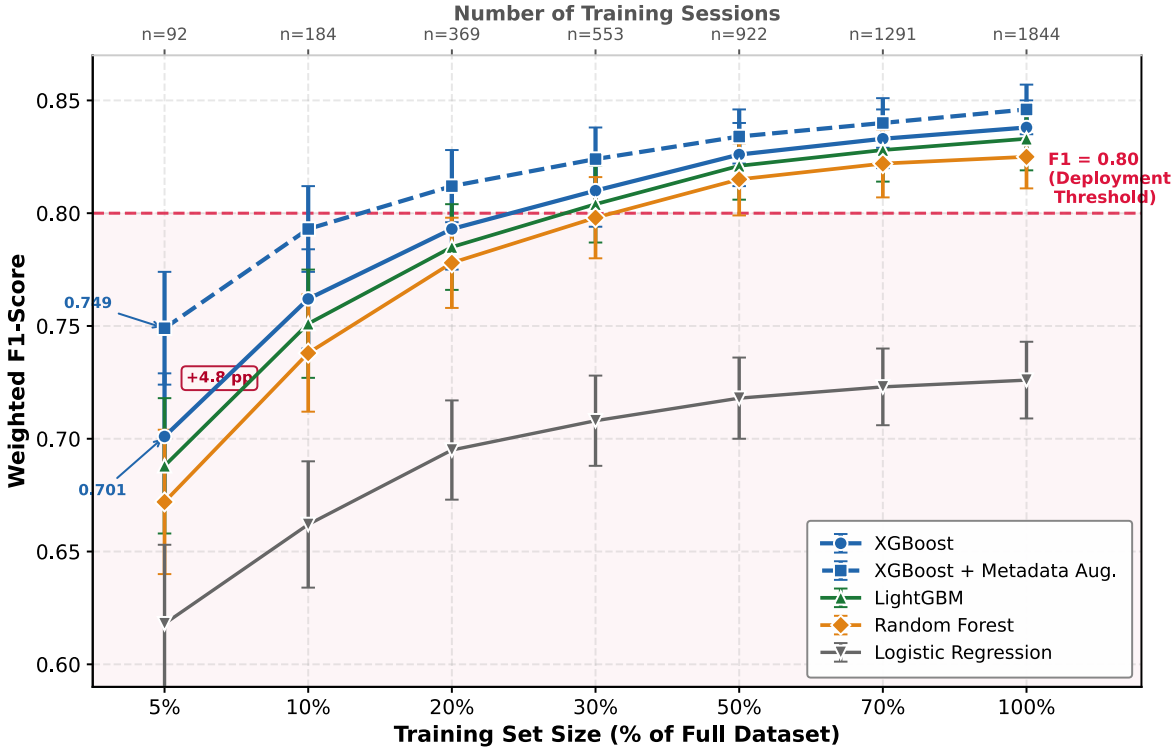

Figure 3 displays a line chart with training set size (x-axis, expressed as percentage of the full training set: 5%, 10%, 20%, 30%, 50%, 70%, 100%) against weighted F1-score (y-axis, range 0.60–0.85). Five lines are plotted: XGBoost (solid blue), XGBoost with metadata augmentation (dashed blue), LightGBM (solid green), RF (solid orange), and Logistic Regression (solid gray). Error bars represent standard deviation across five random subsamples. The chart reveals that all ensemble methods degrade gracefully, with XGBoost maintaining F1 > 0.78 at 20% training data. Metadata augmentation provides the largest relative benefit at 5% and 10% thresholds, narrowing the F1 gap between 10% and full training data from 0.076 to 0.045 for XGBoost. The performance ordering among algorithms remains stable across all data sizes. A horizontal dashed red reference line marks the F1 = 0.80 operational deployment threshold.

At the 10% training data threshold (n=184 sessions), XGBoost without augmentation achieved F1 of 0.762, while the metadata-augmented variant reached 0.793 — a gain of 3.1 percentage points. At 5% (n=92), the augmentation benefit increased to 4.8 percentage points (0.701 → 0.749). These results indicate that structured metadata features partially compensate for the scarcity of behavioral data, though performance remains below the full-data baseline (F1 = 0.838), underscoring the importance of progressive data accumulation.

## 5. Discussion and Future Work

### 5.1. Practical Implications for Cultural Institutions

The findings of this study carry several practical implications for cultural institutions operating virtual exhibition platforms. The consistent superiority of tree-based ensemble methods over linear baselines, achieved without requiring specialized hardware or extensive computational resources, positions these algorithms as accessible tools for institutions with limited technical infrastructure. The identification of session duration ratio and interaction density as dominant predictive features provides actionable guidance for platform designers: real-time monitoring of these two metrics alone could enable lightweight early-warning dashboards for visitor disengagement without deploying the full 38-feature pipeline. The hybrid recommendation approach, which improves content matching through a relatively simple weighted combination of collaborative and content-based signals, can be implemented using standard open-source libraries (Surprise, LightFM) without proprietary infrastructure.

The engagement-weighted recommendation variant offers a mechanism for cultural institutions to balance personalization with curatorial intent. Institutions aligned with the National Endowment for the Arts (NEA) and the Institute of Museum and Library Services (IMLS) strategic priorities of expanding cultural participation may find particular value in monitoring the ILD (Intra-List Diversity) metric alongside accuracy measures, as high personalization accuracy that narrows content exposure could inadvertently reinforce existing cultural preference patterns rather than broadening visitor horizons. The cold-start mitigation results suggest that newly established virtual museums can achieve operational-level prediction performance (F1 > 0.80) once approximately 500 user sessions have been accumulated, provided that structured artwork metadata is incorporated as supplementary features during the initial deployment phase.

The modest magnitude of performance differences between algorithms (XGBoost vs. LightGBM F1 difference of 0.5 percentage points) carries an additional practical message: for resource-constrained institutions, the choice between these two algorithms is less critical than the investment in feature engineering quality and data collection infrastructure. Practitioners should prioritize comprehensive behavioral logging over algorithmic sophistication in early-stage platform development.

### 5.2. Limitations

Several limitations constrain the generalizability of the present findings. The dataset reflects a single platform deployment with a specific artwork collection and audience demographic profile. Cross-platform validation on datasets from institutions with diverse collection types (e.g., science museums, historical archives) is needed to assess whether the identified patterns of feature importance hold across cultural domains. The three-class engagement operationalization simplifies a continuous behavioral spectrum; future work could explore regression-based approaches or ordinal classification to preserve finer-grained distinctions.

The cold-start evaluation used a simulated data-reduction protocol rather than a genuine new-platform deployment, which may overestimate transfer effectiveness. Genuine cross-institution transfer studies — training on data from one museum and evaluating on a different institution's visitors — would provide stronger evidence for practical transferability. The recommendation evaluation did not incorporate long-term engagement outcomes (return visits, sustained interest) due to the six-month collection window; longitudinal studies spanning multiple visit cycles would better characterize the sustained impact of personalized recommendations on cultural engagement depth.

## References

[1]. Sylaiou, S., Dafiotis, P., Koukopoulos, D., Koukoulis, C., Vital, R., Antoniou, A., & Fidas, C. A. (2024). From physical to virtual art exhibitions and beyond: Survey and some issues for consideration for the metaverse. Journal of Cultural Heritage, 66, 86–98. https://doi.org/10.1016/j.culher.2023.11.002

[2]. Mihelj, S., Leguina, A., & Downey, J. (2019). Culture is digital: Cultural participation, diversity and the digital divide. New Media & Society, 21(7), 1465–1485. https://doi.org/10.1177/1461444819831196

[3]. Deng, N., Zhang, N., & Jiang, X. (2025). Transformation in the digital age: Factors influencing visitor engagement in virtual museums. ACM Journal on Computing and Cultural Heritage, 18(3), 1–23. https://doi.org/10.1145/3726873

[4]. Emerson, A., Henderson, N., Rowe, J., Min, W., Lee, S., Minogue, J., & Lester, J. (2020). Early prediction of visitor engagement in science museums with multimodal learning analytics. Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), 107–116. https://doi.org/10.1145/3382507.3418890

[5]. Wu, S. (2019). Forecasting museum visitor behaviors using deep learning. Proceedings of the IEEE International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 36–42. https://doi.org/10.1109/MLBDBI48998.2019.00015

[6]. Tian, J., Kim, D., Yang, S., Park, Y., & Kim, J. (2025). Enhancing museum visitor forecasting using deep learning and sentiment analysis: A transformer-based approach for sustainable management. PLOS ONE, 20(1), e0335623. https://doi.org/10.1371/journal.pone.0335623

[7]. Pavlidis, G. (2019). Recommender systems, cultural heritage applications, and the way forward. Journal of Cultural Heritage, 35, 183–196. https://doi.org/10.1016/j.culher.2018.06.003

[8]. Casillo, M., Colace, F., Conte, D., Lombardi, M., Santaniello, D., & Valentino, C. (2023). Context-aware recommender systems and cultural heritage: A survey. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3109–3127. https://doi.org/10.1007/s12652-021-03438-9

[9]. Raptis, G. E., Fidas, C., Katsini, C., & Avouris, N. (2019). A cognition-centered personalization framework for cultural-heritage content. User Modeling and User-Adapted Interaction, 29(1), 9–65. https://doi.org/10.1007/s11257-019-09226-7

[10]. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

[11]. Panos, F., Raptis, G. E., Katsini, C., & Katsanos, C. (2024). Towards identifying visitor types in virtual museums using spatial and interaction data. Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), 286–291. https://doi.org/10.1145/3631700.3664908

[12]. Konstantakis, M., Alexandridis, G., & Caridakis, G. (2020). A personalized heritage-oriented recommender system based on extended cultural tourist typologies. Big Data and Cognitive Computing, 4(2), Article 12. https://doi.org/10.3390/bdcc4020012

[13]. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[14]. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems (NeurIPS), 35, 507–520.

[15]. Zhang, T., Zhang, Z., Fan, Z., Luo, H., Liu, F., Liu, Q., Cao, W., & Li, J. (2023). OpenFE: Automated feature generation with expert-level performance. Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202, 41880–41901.

[16]. McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., & White, C. (2023). When do neural nets outperform boosted trees on tabular data? Advances in Neural Information Processing Systems (NeurIPS), 36, 76336–76369.

[17]. Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. Proceedings of the International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=cp5PvcI6w8_

[18]. Ferraro, A., Ferreira, G., Diaz, F., & Born, G. (2022). Measuring commonality in recommendation of cultural content: Recommender systems to enhance cultural citizenship. Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22), 567–572. https://doi.org/10.1145/3523227.3551476

[19]. Casillo, M., Colace, F., Conte, D., Lombardi, M., Santaniello, D., & Valentino, C. (2025). A context-aware recommender system-based framework for improving cultural experiences. User Modeling and User-Adapted Interaction, 36(1). https://doi.org/10.1007/s11257-025-09437-1

[20]. Philippopoulos, P. I., Drivas, I. C., Tselikas, N. D., Koutrakis, K. N., & Melidi, E. (2024). A holistic approach for enhancing museum performance and visitor experience. Sensors, 24(3), Article 966. https://doi.org/10.3390/s24030966

[21]. Trichopoulos, G., Konstantakis, M., Alexandridis, G., & Caridakis, G. (2023). Large language models as recommendation systems in museums. Electronics, 12(18), Article 3829. https://doi.org/10.3390/electronics12183829

[22]. Tsitseklis, K., Stavropoulou, G., Zafeiropoulos, A., Thanou, A., & Papavassiliou, S. (2023). RECBOT: Virtual museum navigation through a chatbot assistant and personalized recommendations. Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), 388–396. https://doi.org/10.1145/3563359.3596661