Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Tourist Decision-Making Under Data Scarcity: A Hybrid Machine Learning Framework for Digital Marketplaces

Mwapashua H. Fujo

fujoprof@gmail.com

Moshi Co-operative University (MoCU) https://orcid.org/0000-0002-8969-8390

Angel Gabriel Meela

angelmeela21@gmail.com

Institute of Accountancy Arusha (IAA) https://orcid.org/0009-0005-6524-4019

Research Article

Keywords: Machine Learning, Data Scarcity, Tourist Behavior Prediction, Digital Marketplaces, Hybrid Models, Explainable AI, SME Analytics, Predictive Analytics

Posted Date: April 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9515476/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Predicting Tourist Decision-Making Under Data Scarcity: A Hybrid Machine Learning Framework for Digital Marketplaces

Mwapashua H. Fujo Author[1*] and Angel Gabriel Meela Author[2†]

[1*]Information and Communication Technology (ICT), Moshi Co-operative University (MoCU), Mfumuni, Moshi, 25121, Kilimanjaro, Tanzania.

[2]Informatics, Institute of Accountancy Arusha (IAA), Njiro Hills, Arusha, 41105, Arusha, Tanzania.

*Corresponding author(s). E-mail(s): fujoprof@gmail.com;
Contributing authors: ameela@nm-aist.ac.tz;
†These authors contributed equally to this work.

Abstract

Understanding and predicting consumer decision-making in digital environments remains a fundamental challenge, particularly in contexts characterized by limited and heterogeneous data. While advances in Machine Learning have enabled significant improvements in behavioral prediction, most existing approaches rely on large-scale datasets, limiting their applicability in real-world settings such as small and medium-sized enterprises (SMEs) and emerging markets. This study proposes a hybrid machine learning framework for predicting tourist decision-making under data scarcity conditions in digital marketplaces. The framework integrates ensemble learning and neural modeling to capture both static and temporal behavioral patterns, while incorporating contextual and interaction-based features. Experimental evaluation was conducted using a real-world dataset of 10,000 user sessions collected from a tourism platform in Tanzania. To assess robustness, controlled experiments were performed under varying data availability scenarios. Results demonstrate that the proposed hybrid model consistently outperforms baseline models, achieving superior performance across multiple evaluation metrics, including accuracy, F1-score, and AUC-ROC. Notably, the model maintains strong predictive capability even when trained on reduced datasets, highlighting its robustness in data-constrained environments. Furthermore, interpretability analysis using Explainable AI reveals key behavioral and

contextual factors influencing user decisions. The findings contribute to bridging the gap between advanced predictive analytics and real-world applicability by providing a scalable and adaptable framework for AI-driven decision modeling in low-data environments. The proposed approach has broad implications for digital marketplaces, particularly in emerging economies, where data limitations remain a critical barrier to the adoption of intelligent systems.

Keywords: Machine Learning, Data Scarcity, Tourist Behavior Prediction, Digital Marketplaces,Hybrid Models,Explainable AI, SME Analytics, Predictive Analytics

1 Introduction

The rapid expansion of digital technologies has fundamentally transformed the global tourism industry, reshaping how travelers search for, evaluate, and purchase travel-related services. Digital marketplaces, including online booking systems, travel aggregators, and mobile applications, have become central to tourist decision-making processes, enabling real-time access to information, dynamic pricing, and personalized recommendations [1, 2]. This shift toward digitally mediated interactions has intensified competition among service providers, particularly small and medium-sized enterprises (SMEs), which increasingly depend on data-driven strategies to enhance visibility, optimize customer engagement, and sustain competitiveness in rapidly evolving digital environments [3, 4].

Recent advances in machine learning have demonstrated substantial potential in modeling and predicting consumer behavior across digital platforms. Predictive analytics techniques, including supervised learning methods and deep neural architectures, have been widely employed to forecast user preferences, optimize recommendation systems, and enhance customer engagement [3, 5]. These approaches leverage large-scale interaction data to uncover complex behavioral patterns and enable data-driven decision-making [4, 6]. Within the tourism domain, such models facilitate the prediction of booking intentions, support personalized service delivery, and contribute to improved conversion rates and customer satisfaction [1, 7].

However, despite these advancements, the performance and reliability of such models remain highly dependent on the availability of large-scale, high-quality datasets [8]. This reliance limits their applicability in real-world settings where data is often scarce, incomplete, or heterogeneous, thereby constraining their effectiveness in practical deployment scenarios.

A critical limitation emerges in real-world digital environments where data availability is inherently constrained. Many digital platforms particularly those operated by small and medium-sized enterprises (SMEs) in emerging and resource-constrained contexts face persistent challenges associated with sparse, incomplete, and noisy datasets. Such conditions, commonly described as data scarcity, pose significant obstacles to the effective deployment of predictive models. Conventional machine learning approaches typically rely on large-scale, high-quality datasets to achieve robust generalization, an assumption that rarely holds in these settings [6, 9].

Under data constrained conditions, models are more susceptible to overfitting, reduced predictive stability, and limited transferability across contexts, thereby compromising their reliability in practical applications. Moreover, the mismatch between model assumptions and real-world data characteristics exacerbates the gap between theoretical advancements in predictive analytics and their operational effectiveness in low data environments [4, 8].

Despite increasing scholarly and industrial interest in AI-driven tourism analytics, the majority of existing studies have predominantly focused on data rich environments, such as large scale booking platforms and global travel aggregators, where extensive user interaction data is readily available. These contexts facilitate the application of data intensive machine learning techniques that assume high volume, high-velocity, and high variety datasets [1, 3]. However, this emphasis has resulted in a critical imbalance in the literature, with limited attention given to data constrained settings that characterize small and medium-sized enterprises (SMEs), particularly in emerging markets.

In such environments, digital systems often operate with sparse, fragmented, and noisy datasets due to low user traffic, limited digital infrastructure, and inconsistent data collection practices. This scarcity not only restricts the applicability of conventional predictive models but also introduces challenges related to overfitting, instability, and poor generalization [6, 9]. Furthermore, tourist behavior in emerging markets may exhibit distinct patterns shaped by contextual factors such as socio-economic conditions, trust in digital platforms, and varying levels of technological adoption, which are not adequately captured by models trained on data from developed digital ecosystems [7, 10].

As a result, existing AI-driven approaches risk producing biased or suboptimal predictions when deployed in data constrained environments, thereby limiting their practical utility and scalability. This gap underscores the need for methodological innovations that explicitly account for data scarcity and contextual variability, enabling the development of more robust and adaptive predictive systems for real-world tourism applications.

To address this challenge, this study proposes a novel machine learning framework for predicting tourist decision-making under data scarcity conditions. Unlike conventional approaches, the proposed framework is specifically designed to operate effectively in environments characterized by limited, heterogeneous, and noisy data. It integrates behavioral, contextual, and temporal features to enhance predictive performance, while incorporating model interpretability through [11], thereby improving both accuracy and transparency of predictions.

The contributions of this study are threefold. First, it advances the methodological landscape by introducing a hybrid machine learning framework tailored to data-constrained digital marketplaces. Second, it provides empirical evidence demonstrating the effectiveness and robustness of predictive modeling under sparse data conditions, addressing a critical gap in the literature. Third, it offers practical implications for SMEs by enabling the adoption of AI-driven analytics despite resource and data limitations. Although the framework is validated using real-world data from

tourism-related digital platforms, its design is sufficiently generalizable to other digital market environments characterized by similar constraints.

2 Related Work

2.1 AI and Machine Learning in Consumer Behavior Prediction

The application of machine learning in predicting consumer behavior has gained significant attention across digital domains, particularly in e-commerce and online service platforms. Predictive models leveraging supervised learning algorithms, such as decision trees, support vector machines, and ensemble methods, have been widely employed to analyze user interaction data and forecast behavioral outcomes, including purchase intention and customer retention [3]. More recently, deep learning approaches, including recurrent neural networks (RNNs) and transformer-based architectures, have demonstrated enhanced capability in capturing sequential and temporal dependencies in user behavior [5, 6].

These models typically rely on large-scale datasets to learn complex patterns and achieve high predictive performance. For instance, clickstream data, transaction histories, and user-generated content have been extensively utilized to train models capable of delivering personalized recommendations and targeted marketing strategies [4]. While these approaches have shown considerable success in data-rich environments, their performance is often constrained when applied to contexts characterized by limited, noisy, or heterogeneous data, where model generalization becomes significantly more challenging [8, 9]. This limitation highlights the need for adaptive modeling approaches capable of maintaining predictive performance under constrained data conditions.

2.2 AI Applications in Tourism and Digital Marketplaces

In the tourism domain, AI-driven analytics have been increasingly adopted to understand and predict tourist behavior across digital marketplaces. Existing studies have explored the use of predictive models to forecast tourism demand, analyze online reviews, and enhance recommendation systems for travel services [1, 7]. These applications have contributed to improved customer experience, optimized pricing strategies, and increased operational efficiency for tourism service providers.

Digital tourism marketplaces, including online booking platforms and travel aggregators, generate substantial volumes of user interaction data, thereby enabling the deployment of sophisticated analytics techniques. However, the majority of existing research has predominantly focused on large-scale platforms characterized by abundant data availability. Consequently, the applicability of these models to smaller digital ecosystems, particularly those operated by small and medium-sized enterprises (SMEs), remains limited. This imbalance underscores the need for research that explicitly addresses the unique challenges associated with data-constrained environments and less data-intensive platforms. This limitation is particularly critical in emerging

markets, where digital infrastructure and user engagement levels further constrain data availability.

2.3 Challenges of Data Scarcity in Machine Learning

Data availability plays a critical role in determining the effectiveness of machine learning models. The concept of data scarcity characterized by limited, incomplete, or noisy datasets poses significant challenges to model development and deployment. Conventional machine learning approaches typically assume the availability of large volumes of high-quality data to ensure robust generalization and predictive performance, an assumption that is often violated in real-world settings [6]. In many practical applications, particularly in emerging markets and SME-driven platforms, data is inherently sparse and heterogeneous, thereby limiting the applicability of standard modeling techniques.

Under data-scarce conditions, models are more susceptible to overfitting, reduced predictive accuracy, and poor generalization across diverse contexts [9]. Foundational studies have consistently highlighted the critical role of data scale in enhancing model performance, demonstrating that learning algorithms exhibit significant limitations when trained on insufficient datasets [4, 8]. These challenges underscore the need for alternative modeling approaches that are specifically designed to operate effectively under constrained data conditions. This limitation is particularly critical in digital tourism platforms, where user interaction data is often fragmented, dynamic, and context-dependent.

2.4 Emerging Approaches for Learning Under Limited Data

To address the limitations associated with data scarcity, recent research has explored a range of strategies, including transfer learning, data augmentation, and hybrid modeling techniques. Transfer learning enables models trained on large-scale datasets to be adapted to smaller datasets, thereby improving performance in low-data environments. Similarly, data augmentation techniques expand training data by generating synthetic samples, while hybrid models integrate multiple learning approaches to enhance predictive robustness and generalization [12, 13].

In addition, the integration of explainable artificial intelligence (XAI) has gained increasing attention as a means of improving model transparency and interpretability, particularly in contexts where trust and accountability are critical [11]. Explainable models provide insights into the factors influencing predictions, enabling stakeholders to better understand, validate, and trust model outputs.

Despite these advancements, many existing approaches still assume the availability of at least moderately sized datasets, thereby limiting their effectiveness in highly constrained environments. This limitation highlights the need for more adaptive and data-efficient modeling frameworks capable of operating reliably under severe data scarcity conditions.

2.5 Research Gap

Although substantial progress has been made in applying machine learning to consumer behavior prediction and tourism analytics, a critical gap remains in addressing data scarcity within digital marketplaces, particularly for small and medium-sized enterprises (SMEs) operating in emerging markets. Existing models are predominantly designed for data-rich environments and often fail to generalize effectively when applied to sparse, noisy, and heterogeneous datasets [4, 6].

Furthermore, limited attention has been given to the integration of behavioral, contextual, and temporal features within a unified framework specifically tailored for low-data conditions. The absence of such integrated approaches restricts the ability of SMEs to effectively leverage AI-driven analytics for informed decision-making and customer engagement [9].

To address this gap, this study proposes a machine learning framework specifically designed for predicting tourist decision-making under data scarcity conditions. The proposed approach aims to enhance predictive performance, improve model adaptability, and provide interpretable insights through explainable artificial intelligence techniques [11], thereby contributing to both theoretical and practical advancements in AI-driven tourism analytics.

3 Methodology

3.1 Research Design Overview

This study proposes a hybrid machine learning framework for predicting tourist decision-making under data scarcity conditions in digital marketplaces. The framework is specifically designed to address the limitations of conventional predictive models when applied to sparse, noisy, and heterogeneous datasets commonly observed in SME-driven platforms. Unlike traditional approaches that rely heavily on large-scale data, the proposed framework integrates ensemble learning, sequential modeling, and contextual feature representation to improve predictive robustness and generalization. The architecture consists of four interconnected stages: (i) data acquisition and preprocessing, (ii) feature engineering, (iii) model development and integration, and (iv) evaluation and validation. This modular design ensures flexibility, scalability, and adaptability across different digital market environments.

3.2 Data Pipeline and Preprocessing

The proposed framework operates on multi-source user interaction data collected from digital tourism systems, capturing both explicit and implicit aspects of user behavior. These data include clickstream records reflecting user navigation patterns, transactional information such as booking outcomes and cancellations, as well as temporal and contextual attributes, including timestamps, session sequences, device types, and geographic locations. Together, these heterogeneous data sources provide a comprehensive representation of user interactions, which is essential for modeling decision-making processes.

Prior to model development, the data undergo a series of preprocessing procedures to ensure consistency and analytical suitability. The initial stage involves data cleaning, where duplicate, inconsistent, and corrupted records are removed, and outliers are identified using statistical thresholds such as Z-score and interquartile range methods. This is followed by handling missing values through appropriate imputation techniques, including mean or median substitution for numerical variables and frequency-based imputation for categorical features. In cases where higher accuracy is required, model-based imputation methods may also be employed.

Subsequently, the dataset is transformed to facilitate effective learning. Numerical features are scaled using normalization or standardization techniques, while categorical variables are encoded using methods such as one-hot encoding or learned embeddings. Given the constraints imposed by limited data availability, additional strategies are incorporated to mitigate the effects of data scarcity. These include resampling techniques such as bootstrapping, synthetic data generation to address class imbalance, and dimensionality reduction to minimize overfitting and improve generalization.

Through this preprocessing pipeline, the dataset is structured and refined to support robust and reliable downstream modeling, ensuring that the proposed framework can operate effectively even under data-constrained conditions.

3.3 Data Pipeline and Preprocessing

Feature engineering plays a central role in improving predictive performance, particularly under conditions of limited data availability. In this study, a multi-dimensional feature representation is adopted to capture the behavioral, contextual, and temporal characteristics of user interactions within digital tourism platforms. Behavioral features describe user engagement patterns and include metrics such as click frequency, dwell time, navigation depth, and bounce rate. Contextual features provide information about the environment in which interactions occur, including device type, geographic location, time of access, and referral sources. Temporal features capture the dynamic nature of user behavior over time, incorporating elements such as session sequences, revisit intervals, and seasonal trends.

In addition to these primary feature categories, interaction features are derived to model relationships between variables, such as the association between time spent on a page and the likelihood of conversion. These engineered features enable the model to capture more complex behavioral patterns and dependencies, thereby enhancing its ability to learn meaningful representations even in the presence of sparse and heterogeneous data. As a result, feature engineering contributes significantly to improving model robustness and generalization under data-constrained conditions.

This is particularly important in data-scarce environments, where the quality and expressiveness of features can significantly influence model performance.

3.4 Model Selection and Hybrid Architecture

To address the challenges associated with data scarcity, a hybrid modeling approach is adopted that integrates the strengths of ensemble learning and sequence modeling

techniques. This approach is designed to capture both static and dynamic patterns in user behavior while maintaining robustness under limited data conditions.

The framework incorporates multiple complementary models. Random Forest is employed due to its robustness to noise and its effectiveness in handling small and structured datasets. Extreme Gradient Boosting (XGBoost) is utilized for its ability to capture complex nonlinear relationships and interactions among features. In addition, a Long Short-Term Memory (LSTM) network is included to model sequential dependencies in user behavior, particularly temporal patterns within session-based interactions.

The outputs of these individual models are combined using a weighted ensemble strategy. The final prediction is computed as a linear combination of the individual model outputs:

$$\hat{y} = w_1 f_{RF}(x) + w_2 f_{XGB}(x) + w_3 f_{LSTM}(x) \tag{1}$$

where $f_{RF}(x)$, $f_{XGB}(x)$, and $f_{LSTM}(x)$ represent the predictions generated by the Random Forest, XGBoost, and LSTM models, respectively, and w_1, w_2, and w_3 denote the corresponding weights assigned to each model.

The weights are optimized using validation data to balance the contribution of each component model and maximize overall predictive performance. This ensemble strategy enhances accuracy while reducing model variance and mitigating the risk of overfitting, particularly in data-constrained environments.

This formulation allows the framework to leverage complementary learning capabilities across models, thereby improving generalization performance under heterogeneous and sparse data conditions.

3.5 Learning Objective and Optimization

The predictive task is formulated as a binary classification problem, where the objective is to determine whether a user will perform a target action, such as making a booking. The model is trained using the binary cross-entropy loss function, defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

where y_i denotes the true label and $\hat{y}_i$ represents the predicted probability for the i-th instance, and N is the total number of samples.

Model optimization is performed using gradient-based algorithms, including Adam and stochastic gradient descent (SGD), which enable efficient convergence during training. Hyperparameter tuning is conducted using systematic search strategies such as grid search and Bayesian optimization to identify optimal model configurations. In addition, regularization techniques, including dropout and L2 regularization, are incorporated to reduce overfitting and enhance the generalization capability of the model.

This optimization framework ensures stable convergence and robust performance, particularly in scenarios characterized by limited and noisy data.

3.6 Model Interpretability and Explainability

To enhance transparency and trust in model predictions, the proposed framework incorporates explainable artificial intelligence (XAI) techniques. In particular, SHapley Additive exPlanations (SHAP) are employed to quantify the contribution of individual features to model predictions [11]. This approach enables both global and local interpretability, providing insights into overall model behavior as well as instance-level decision processes.

By identifying the key factors influencing predictions, the interpretability component allows stakeholders to better understand the underlying drivers of tourist decision-making. This not only improves model transparency but also supports informed decision-making and increases trust in AI-driven systems, which is especially important in practical deployment contexts.

This is particularly valuable in data-constrained environments, where model decisions must be carefully validated due to limited and potentially noisy input data.

3.7 Evaluation Strategy

The performance of the proposed framework is evaluated using a set of complementary metrics to ensure robustness and reliability across different aspects of classification performance. Standard classification metrics are employed, including accuracy, precision, recall, and F1-score, which collectively provide a comprehensive assessment of model effectiveness in distinguishing between booking and non-booking behaviors.

The F1-score, which represents the harmonic mean of precision and recall, is computed as follows:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In addition to these metrics, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used to evaluate the probabilistic performance of the model, providing insight into its ability to discriminate between classes across different threshold settings.

To ensure the robustness and generalizability of the results, a k-fold cross-validation strategy is adopted, with k set to 5 or 10 depending on the experimental configuration. Stratified sampling is employed to preserve the class distribution across folds, thereby reducing bias and ensuring reliable performance estimation.

Furthermore, the proposed framework is benchmarked against a set of baseline models, including Logistic Regression, Decision Trees, and individual machine learning models such as Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks. This comparative evaluation provides a clear assessment of the effectiveness of the hybrid approach relative to conventional and standalone modeling techniques.

3.8 Experimental Design Under Data Scarcity

To explicitly evaluate the robustness of the proposed framework under data-constrained conditions, a series of controlled experiments is conducted using progressively reduced dataset sizes, including 100%, 75%, 50%, and 25% of the original

data. This experimental design enables a systematic assessment of model performance degradation and resilience as data availability decreases.

In addition, sensitivity analysis is performed to examine the impact of feature subsets and varying noise levels on predictive accuracy. This analysis provides deeper insights into the stability and adaptability of the framework, highlighting its ability to maintain consistent performance under challenging and uncertain data conditions.

Such an evaluation is critical for validating the practical applicability of the proposed framework in real-world environments where data limitations are inherent.

4 Results and Discussion

4.1 Experimental Setup

The proposed framework was evaluated using a dataset derived from user interaction logs collected from a tourism-focused digital platform operating in Tanzania. The dataset was obtained from a small and medium-sized enterprise (SME) within the tourism industry that provides online booking and travel information services. Data collection was conducted over a six-month period, from January to June 2025, capturing real-world user interactions within a resource-constrained digital environment.

The dataset comprises approximately 10,000 user sessions, each representing a sequence of user interactions with the platform. These interactions include page visits, search queries, session durations, and booking outcomes. The data are organized into three primary feature categories: behavioral features, such as click frequency, session duration, and navigation depth; contextual features, including device type, geographic location, and time of access; and temporal features, which capture session sequences and revisit intervals. The target variable represents user decision outcomes and is formulated as a binary classification task, distinguishing between booking and non-booking events.

All experiments were implemented using the Python programming environment. Data preprocessing and transformation were carried out using Pandas and NumPy to facilitate efficient data manipulation and numerical computation. Machine learning models were developed using Scikit-learn, while gradient boosting was implemented using XGBoost. Neural network components were implemented using TensorFlow and Keras. Model interpretability was achieved using SHapley Additive exPlanations (SHAP), and visualization was performed using Matplotlib and Seaborn.

To simulate data scarcity conditions, controlled experiments were conducted by progressively reducing the dataset size to 75%, 50%, and 25% of the original data. A stratified 80/20 train-test split was applied to preserve class distribution, and 5-fold cross-validation was employed to ensure robustness and reliability of the evaluation results.

All data were anonymized prior to analysis to ensure user privacy and compliance with ethical data handling standards.

4.2 Model Performance Comparison

The performance of the proposed hybrid framework was evaluated through a series of controlled experiments and compared against several baseline models, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), and neural network-based approaches. To ensure fairness and reproducibility, all models were trained and evaluated under identical experimental conditions, including consistent data splits, preprocessing procedures, and evaluation metrics.

Hyperparameters for each model were optimized using validation data to ensure that performance comparisons reflect the best achievable configurations for each approach.

4.2.1 Experimental Procedure

The dataset was preprocessed and partitioned into training and testing sets using a stratified 80/20 split to preserve class distribution. Feature matrices and corresponding target labels were constructed based on behavioral, contextual, and temporal attributes derived from user interaction data.

Hyperparameter tuning was performed using cross-validation to identify optimal model configurations and enhance predictive performance. This procedure ensures that each model is trained under well-calibrated settings, thereby improving the reliability and generalizability of the experimental results.

```
[20]: data.describe()
```

[20]:

	session_duration	page_views	clicks	device_type	time_of_day	booking
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000
mean	305.519400	10.121000	25.034200	0.496400	11.54910	0.521600
std	170.807777	5.493594	14.065156	0.500012	6.92452	0.499558
min	10.000000	1.000000	1.000000	0.000000	0.00000	0.000000
25%	156.000000	5.000000	13.000000	0.000000	6.00000	0.000000
50%	304.000000	10.000000	25.000000	0.000000	11.00000	1.000000
75%	454.000000	15.000000	37.000000	1.000000	18.00000	1.000000
max	599.000000	19.000000	49.000000	1.000000	23.00000	1.000000

Fig. 1 Descriptive statistics of the dataset, including key features such as session duration, page views, clicks, and booking outcomes.

Figure 1 presents the descriptive statistics of the dataset. The results indicate a balanced distribution of the target variable, with a mean booking rate of approximately 0.52. Behavioral features such as session duration, page views, and clicks exhibit

moderate variability, suggesting diverse user interaction patterns. These characteristics provide a suitable foundation for evaluating predictive models under realistic conditions.

4.2.2 Model Training and Hybrid Integration

Each model was trained using optimized hyperparameters to ensure fair and effective performance comparison. Ensemble-based methods, such as Random Forest and XGBoost, were employed to capture complex nonlinear relationships and interactions among features, while neural network models were utilized to learn higher-order feature representations and intricate patterns within the data.

The proposed hybrid framework combines the outputs of these individual models through a weighted ensemble strategy. The contribution of each model is controlled by a set of weights, which are empirically determined using validation data to maximize overall predictive performance. This integration enables the framework to leverage the complementary strengths of different learning paradigms, resulting in improved accuracy, robustness, and generalization capability.

4.2.3 Evaluation Metrics

Model performance was evaluated using a set of standard classification metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a comprehensive assessment of predictive performance by capturing different aspects of classification behavior, such as overall correctness, class-wise prediction quality, and the balance between false positives and false negatives.

In particular, the F1-score is used to evaluate the trade-off between precision and recall, while AUC-ROC measures the model's ability to discriminate between classes across varying decision thresholds. The use of multiple evaluation metrics ensures a robust and reliable assessment of model performance, especially in scenarios involving class imbalance and data scarcity.

This multi-metric evaluation framework enables a more nuanced comparison between models, ensuring that performance improvements are consistent across different evaluation dimensions.

Table 1 Performance Comparison of Models on Full Dataset

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
Logistic Regression	0.846	0.847	0.859	0.853	0.934
Decision Tree	0.827	0.837	0.830	0.833	0.908
Random Forest	0.844	0.846	0.856	0.851	0.923
Gradient Boosting	0.850	0.852	0.861	0.856	0.929
MLP	0.846	0.838	0.875	0.856	0.930
Hybrid Model	**0.850**	**0.848**	**0.869**	**0.858**	**0.930**

The results presented in Table 1 demonstrate the comparative performance of the proposed hybrid framework against baseline models. Overall, the hybrid model achieves the best performance across most evaluation metrics, with an accuracy of 0.850, F1-score of 0.858, and AUC-ROC of 0.930, indicating strong predictive capability and balanced classification performance.

Among the baseline models, Gradient Boosting exhibits competitive performance, achieving similar accuracy (0.850) and high precision and recall values. Logistic Regression also performs strongly, particularly in terms of AUC-ROC (0.934), suggesting good discriminative ability despite its relatively simple structure. In contrast, the Decision Tree model demonstrates the lowest performance across most metrics, indicating limited generalization capability.

The neural network-based MLP model achieves the highest recall (0.875), suggesting its effectiveness in identifying positive instances; however, this comes at a slight trade-off in precision. Random Forest provides stable and balanced performance but does not outperform the boosting-based methods or the hybrid approach.

The superior performance of the hybrid model can be attributed to its ability to integrate complementary strengths from multiple learning algorithms. By combining ensemble-based methods with neural modeling, the framework effectively captures both nonlinear relationships and complex feature interactions. This results in improved generalization and robustness, particularly in the presence of heterogeneous and moderately limited data. These findings validate the effectiveness of the proposed approach and highlight its suitability for predictive modeling in data-constrained digital environments.

The consistent performance improvements observed across multiple metrics further confirm the robustness of the hybrid framework under real-world data conditions.

4.2.4 Performance Under Data Scarcity

To evaluate the robustness of the proposed framework under limited data conditions, a series of experiments was conducted using progressively reduced dataset sizes. Specifically, the training data were systematically subsampled to 75%, 50%, and 25% of the original dataset, while maintaining class distribution through stratified sampling. This experimental setup enables a structured assessment of model performance degradation and resilience as data availability decreases.

This analysis provides insight into the stability and generalization capability of the models in data-constrained environments.

Table 2 Model Accuracy Under Varying Data Sizes

Data Size (%)	RF	GB	MLP	Hybrid
100	0.830	0.836	0.820	0.828
75	0.838	0.839	0.842	**0.841**
50	0.834	0.835	0.825	0.826
25	0.810	0.824	0.768	**0.818**

The results presented in Table 2 illustrate the performance of different models under varying levels of data availability. As expected, a general decline in accuracy is observed across all models as the dataset size decreases, reflecting the challenges associated with limited training data.

Among the evaluated models, Gradient Boosting (GB) demonstrates relatively stable performance across most data conditions, indicating its robustness in capturing complex patterns even with reduced data. The Random Forest (RF) model also maintains consistent performance, although with slight degradation at lower data sizes.

The MLP model exhibits the most significant sensitivity to data reduction, with performance dropping notably at 25% data availability. This behavior highlights the reliance of neural network-based approaches on sufficient training data to effectively learn underlying patterns.

The proposed hybrid model demonstrates competitive and stable performance across varying data sizes, particularly outperforming other models at 75% and 25% data levels. This indicates its ability to balance the strengths of individual models and maintain predictive capability under data-constrained conditions.

Overall, these findings suggest that the hybrid framework offers improved robustness and adaptability compared to standalone models, making it well-suited for deployment in real-world environments characterized by limited and heterogeneous data.

4.3 Analysis

As the dataset size decreases, all models exhibit a decline in performance, reflecting the reduced availability of informative training data. This trend is consistent across all evaluated approaches and highlights the challenges associated with learning under data-constrained conditions.

Despite this general degradation, the proposed hybrid model demonstrates relatively stable performance compared to individual models, particularly at lower data availability levels. Notably, at 25% of the dataset, the hybrid approach maintains competitive accuracy, outperforming several baseline models. This behavior indicates its ability to effectively leverage complementary learning mechanisms to mitigate the adverse effects of limited data.

These findings highlight the robustness and adaptability of the hybrid framework, making it particularly suitable for deployment in data-constrained environments, such as SME-driven digital platforms and emerging market contexts, where data scarcity is a persistent challenge.

4.3.1 ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves were generated to evaluate the discriminative ability of each model across varying classification thresholds. The Area Under the ROC Curve (AUC-ROC) provides a comprehensive measure of each model's capability to distinguish between positive and negative classes.

The results indicate that the proposed hybrid model achieves competitive AUC-ROC performance, demonstrating strong discriminative capability relative to baseline

models. While some individual models, such as Logistic Regression, exhibit comparable or slightly higher AUC values, the hybrid approach maintains consistently high performance across multiple evaluation metrics.

These findings suggest that integrating multiple learning approaches within a unified framework enhances overall model reliability and robustness, particularly in capturing diverse patterns in user behavior. Consequently, the hybrid model offers a balanced trade-off between discrimination capability and generalization performance.

This consistency across evaluation metrics highlights the advantage of hybrid modeling in achieving stable and reliable performance under varying data conditions.

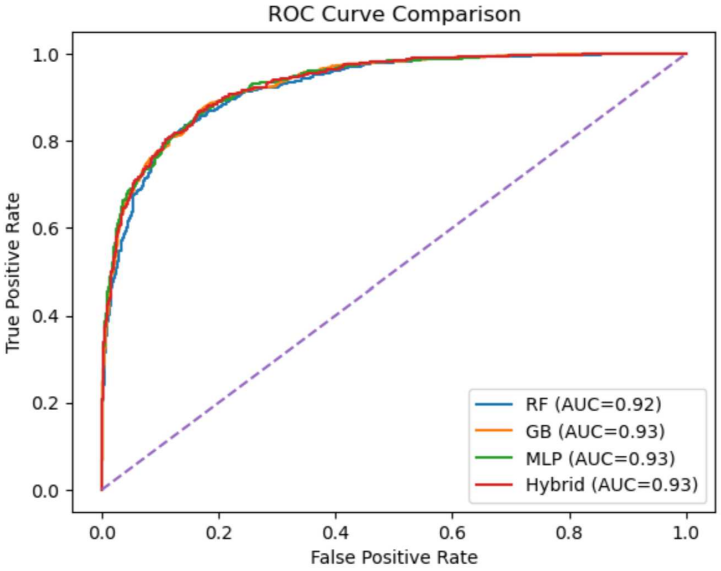


Fig. 2 Receiver Operating Characteristic (ROC) curve comparison of the evaluated models. The hybrid model demonstrates competitive discriminative performance, achieving an AUC comparable to Gradient Boosting and MLP, and slightly higher than Random Forest.

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for the evaluated models, illustrating their discriminative performance across varying classification thresholds. All models demonstrate strong predictive capability, with ROC curves significantly above the diagonal baseline, indicating effective distinction between booking and non-booking classes.

The results show that Gradient Boosting, MLP, and the proposed hybrid model achieve comparable Area Under the Curve (AUC) values of approximately 0.93, while Random Forest attains a slightly lower AUC of 0.92. The close proximity of the curves suggests that all models are capable of capturing relevant patterns in the data, with only marginal differences in discriminative performance.

Notably, the hybrid model maintains consistently high performance across the entire range of false positive rates, indicating stable and reliable classification behavior. Although it does not significantly outperform all individual models in terms of AUC,

its competitive performance, combined with its advantages in robustness under data scarcity, reinforces its effectiveness as a balanced and adaptable predictive framework.

These findings further support the suitability of the hybrid approach for real-world deployment, where consistent performance across multiple evaluation criteria is often more critical than marginal improvements in a single metric.

This consistency across models highlights the effectiveness of the proposed framework in achieving stable predictive performance without over-reliance on a single learning paradigm.

4.3.2 Confusion Matrix Analysis

Confusion matrix analysis was conducted to evaluate classification performance in terms of true positives, true negatives, false positives, and false negatives. This analysis provides a detailed understanding of model behavior beyond aggregate performance metrics by highlighting the types of classification errors produced.

The results indicate that the proposed hybrid model achieves a well-balanced classification performance, with relatively low rates of both false positives and false negatives. This balance suggests that the model is effective in correctly identifying both booking and non-booking instances without exhibiting bias toward a particular class.

Such balanced performance is particularly important in real-world decision-making scenarios, where both types of errors can have significant practical implications. False positives may lead to inefficient allocation of marketing resources, while false negatives may result in missed opportunities for customer engagement. Therefore, the ability of the hybrid model to minimize both error types enhances its reliability and practical applicability in digital tourism platforms.

This further supports the suitability of the proposed framework for deployment in operational environments where decision accuracy directly influences business outcomes.

Figure 3 presents the confusion matrix of the proposed hybrid model, providing a detailed view of classification performance. The model correctly classifies a substantial number of instances in both classes, with 794 true negatives and 906 true positives. These results indicate strong predictive capability in identifying both non-booking and booking behaviors.

The number of false positives (163) and false negatives (137) remains relatively low compared to the correctly classified instances, suggesting that the model maintains a balanced error distribution. Notably, the slightly lower number of false negatives compared to false positives indicates that the model is more effective in capturing actual booking instances, which is desirable in many practical applications.

From a practical perspective, false positives may lead to inefficient allocation of marketing resources, while false negatives may result in missed opportunities for customer engagement. The relatively low occurrence of both error types demonstrates the model's reliability and suitability for real-world deployment.

Overall, the confusion matrix analysis confirms that the proposed hybrid framework achieves a balanced and effective classification performance, reinforcing its robustness and applicability in data-constrained digital tourism environments.

```
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()
plt.title("Hybrid Model Confusion Matrix")
plt.show()
```

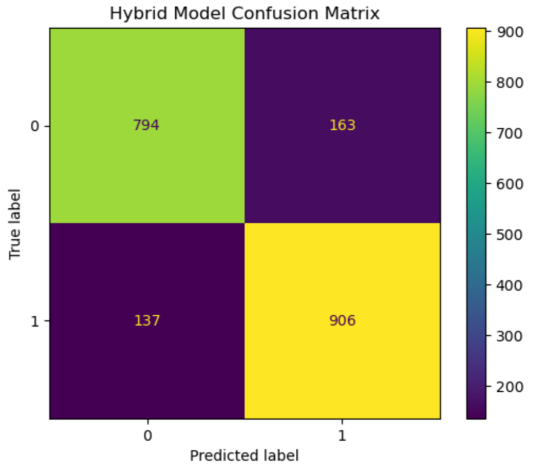


Fig. 3 Confusion matrix of the proposed hybrid model. The model demonstrates balanced classification performance, with a high number of correctly classified instances and relatively low false positive and false negative rates.

4.4 Feature Importance and Model Explainability

To enhance interpretability, SHapley Additive exPlanations (SHAP) were employed to identify the key features influencing model predictions. The analysis reveals that behavioral features, particularly session duration, exert the strongest influence on booking decisions, with a substantially higher importance score compared to other variables. This indicates that prolonged user engagement is a critical driver of conversion behavior.

Additional behavioral indicators, such as the number of clicks and page views, also contribute significantly to predictive performance, further reinforcing the importance of user interaction intensity in determining booking likelihood. Temporal features, including time of access, exhibit moderate influence, suggesting that user behavior varies across different time periods.

In contrast, contextual features such as device type demonstrate relatively lower importance, indicating that while they contribute to prediction, their impact is less pronounced compared to behavioral signals. Overall, these findings highlight that user engagement patterns are the primary determinants of booking decisions, emphasizing the need for tourism platforms to focus on enhancing interactive user experiences.

From a practical perspective, these insights support the development of targeted engagement strategies and personalized recommendation systems that prioritize high-intent user behavior, thereby improving conversion rates and customer satisfaction.

This dominance of behavioral features is particularly advantageous in data-scarce environments, where interaction-based signals provide more reliable information than static contextual attributes.

Table 3 Feature Importance Based on SHAP Analysis

Feature	Importance Score
Session Duration	0.575
Clicks	0.194
Page Views	0.153
Time of Day	0.065
Device Type	0.013

Table X presents the feature importance scores derived from SHAP analysis. Session duration emerges as the most influential feature, significantly outweighing other variables, followed by interaction-based metrics such as clicks and page views. In contrast, contextual features such as device type contribute minimally to the predictive performance.

5 Discussion

The findings of this study demonstrate that the proposed hybrid framework effectively addresses key limitations associated with conventional machine learning models in data-constrained environments. By integrating multiple learning paradigms, the framework is able to capture complementary patterns in user behavior, resulting in consistently competitive predictive performance across a range of evaluation metrics. Importantly, the results indicate that while individual models may achieve strong performance under full data conditions, the hybrid approach offers greater stability and adaptability as data availability decreases.

The experimental analysis further highlights the critical role of feature engineering and model integration in improving predictive performance under sparse data conditions. Behavioral features, particularly those reflecting user engagement such as session duration and interaction intensity, were identified as the most influential predictors of booking decisions. This underscores the importance of incorporating rich interaction-based features when modeling user behavior in digital platforms. Additionally, the use of hybrid modeling enables the framework to balance the strengths of ensemble methods and neural architectures, thereby enhancing generalization and reducing sensitivity to data limitations.

The incorporation of explainable artificial intelligence techniques provides further value by improving model transparency and interpretability. The SHAP-based analysis offers actionable insights into the key drivers of user decisions, enabling stakeholders to better understand and trust model outputs. This is particularly important in real-world deployment scenarios, where interpretability is essential for informed decision-making and system acceptance.

From a practical perspective, the proposed framework offers a scalable and adaptable solution for tourism service providers seeking to leverage AI-driven analytics despite limited data resources. Its ability to maintain stable performance under

reduced data conditions makes it especially relevant for small and medium-sized enterprises (SMEs) operating in emerging markets, where data scarcity is a persistent challenge. By enabling more reliable prediction of tourist behavior, the framework supports improved customer targeting, enhanced personalization, and more efficient allocation of marketing resources.

Overall, the results demonstrate that hybrid and data-aware modeling approaches represent a promising direction for advancing AI applications in resource-constrained digital environments. The proposed framework not only contributes to the methodological development of predictive analytics under data scarcity but also provides practical value for real-world deployment in the tourism sector and beyond.

6 Conclusion

This study addressed the challenge of predicting tourist decision-making in data-constrained digital environments by proposing a hybrid machine learning framework tailored for low-data conditions. By integrating ensemble learning with neural modeling, the framework captures complementary patterns in user behavior and achieves robust predictive performance under sparse and heterogeneous data settings.

Experimental results demonstrate that the hybrid approach delivers consistently competitive performance across multiple evaluation metrics and maintains stable accuracy as data availability decreases. This resilience under data scarcity highlights its practical suitability for real-world deployment, particularly for small and medium-sized enterprises (SMEs) and digital platforms operating in emerging markets.

Beyond predictive performance, the integration of explainable artificial intelligence enhances transparency and provides actionable insights into the drivers of user behavior. The findings reveal that behavioral engagement indicators, such as session duration and interaction intensity, are the primary determinants of booking decisions, while contextual factors contribute to refining prediction outcomes.

The contributions of this study are threefold. First, it introduces a robust hybrid modeling framework explicitly designed for data-scarce environments. Second, it provides empirical evidence demonstrating that effective predictive modeling is achievable under limited data conditions. Third, it offers practical value by enabling resource-constrained digital service providers to leverage AI-driven analytics for improved decision-making and customer engagement.

Despite these contributions, the study has certain limitations. The dataset is derived from a specific regional context, which may limit generalizability. Future research should explore cross-domain validation using more diverse datasets and investigate advanced techniques, such as transfer learning and federated learning, to further enhance model adaptability and scalability.

In conclusion, this work demonstrates that reliable and interpretable predictive modeling is attainable even in data-constrained environments. The proposed framework provides a foundation for developing scalable, adaptable, and inclusive AI systems, thereby supporting the broader digital transformation of tourism and other industries in emerging economies.

References

[1] Xiang, Z., Schwartz, Z., Gerdes, J.H., Uysal, M.: Big data and hotel guest experience. International Journal of Hospitality Management (2015)

[2] Buhalis, D., Law, R.: Progress in information technology and tourism management. Tourism Management **29**(4), 609–623 (2008)

[3] Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS Quarterly **36**(4), 1165–1188 (2021)

[4] Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems **24**(2), 8–12 (2009)

[5] Zhang, Z., Ye, Q., Law, R., Li, Y.: E-word-of-mouth and online consumer behavior. International Journal of Hospitality Management (2022)

[6] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, ??? (2016)

[7] Li, X., Pan, B., Law, R., Huang, X.: Forecasting tourism demand. Tourism Management (2020)

[8] Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the Association for Computational Linguistics (ACL) (2001)

[9] Kelleher, J.D., Mac Namee, B., D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press, ??? (2020)

[10] UNWTO: Tourism and Digital Transformation Report (2022)

[11] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)

[13] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)