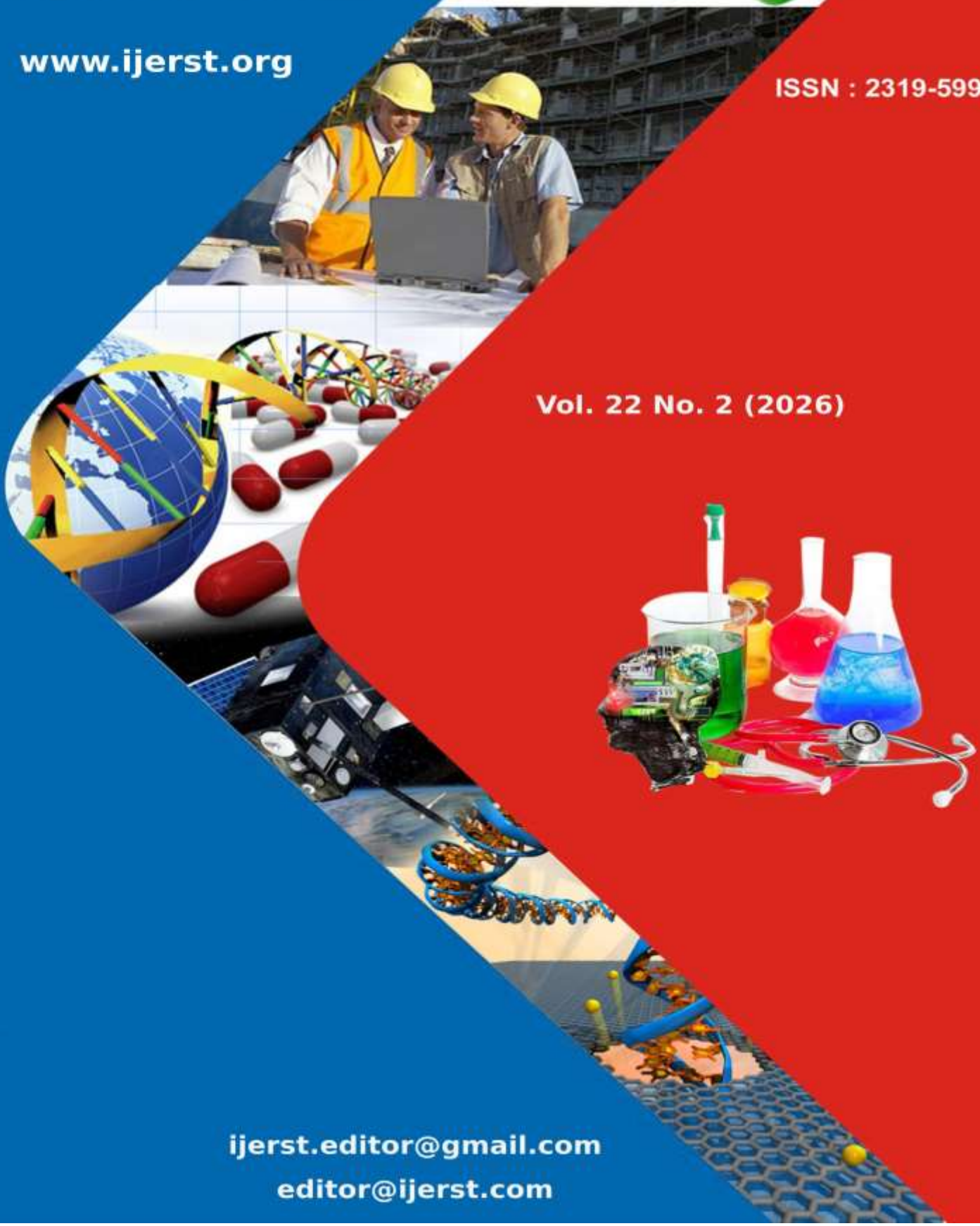
International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

PCA-Enabled Ensemble Learning for Accurate and Explainable Hotel Booking Cancellation Forecasting

1CH.Srinivasulu Reddy, Associate Professor, CSE Department, PBR Visvodaya Institute of Technology and Science, Kavali-A.P

2M.Mokshitha, M.Tech, Computer Science and Engineering, PBR Visvodaya Institute of Technology and Science, Kavali-A.P

Abstract: Hotel booking cancellations pose a major challenge to the hospitality industry by causing revenue loss and inaccurate demand forecasting. While stacking-based ensemble learning models have shown strong predictive performance, their efficiency can be affected by redundant and irrelevant features. To address this limitation, this paper presents an optimized stacking meta-model enhanced with Principal Component Analysis (PCA) for feature selection and dimensionality reduction. PCA is applied to identify the most relevant attributes from the hotel booking dataset, and the stacking model is retrained using these selected features to improve prediction efficiency and accuracy. The optimized model integrates multiple machine learning algorithms with Logistic Regression as the final estimator, achieving a superior accuracy of 97.3%, outperforming the original stacking model. Furthermore, SHapley Additive exPlanations (SHAP) are employed to provide interpretable insights into key factors influencing booking cancellations. The proposed extension demonstrates that PCA-driven feature optimization significantly enhances model robustness, computational efficiency, and interpretability, offering an effective and practical solution for hotel revenue management and operational decision-making.

Index Terms— Hotel booking cancellation prediction, stacking meta-model, principal component analysis (PCA), feature selection, ensemble learning, explainable artificial intelligence (XAI), SHAP, machine learning, revenue management, hospitality analytics.

1. INTRODUCTION

The hospitality business drives worldwide economic development, yet hotel booking cancellations continue to plague revenue management and operational planning. Online booking platforms have increased cancellation rates, distorting demand predictions and causing major financial losses. Hoteliers must accurately estimate booking

cancellations to maximize inventory, personnel, and price.

Recent machine learning breakthroughs have enabled prediction algorithms to identify booking cancellations using past reservation data. By mixing algorithms, ensemble learning methods, such as stacking meta-models, perform better. Stacking models are useful, but high-dimensional feature spaces may include duplicate or unnecessary properties, raising computing cost and decreasing generalization efficiency.

The cancelation prediction framework in this study combines Principal Component Analysis (PCA) with stacking-based ensemble learning to overcome these constraints. PCA selects and reduces dimensionality to focus on the most useful features and reduce dataset noise and redundancy. These optimized features are used to retrain the stacking meta-model, improving accuracy, robustness, and training efficiency over the ensemble model.

Real-world hospitality implementation requires predictive performance and interpretability. Therefore, SHapley Additive exPlanations (SHAP) are used to explain model predictions and discover booking cancelation causes. The addon improves transparency, usability, and prediction accuracy to 97.3%. This study shows that PCA-driven feature optimization and interpretable ensemble learning can anticipate hotel booking cancellations intelligently and efficiently.

2. LITERATURE SURVEY

i) An integrated forecasting approach to hotel demand

https://www.sciencedirect.com/science/article/pii/S0895717707000866

Our goals were to predict demand swings that cause problems in the hotel industry and develop a dynamic, methodical procedure that could be applied again. Using 149 monthly series of data, we predicted demand for a five-star hotel in Ankara. Using error measures, we compared the findings with those using MA, Simple, Holt's, Winter's Exponential Smoothing, and ARIMA. The Variables Determination Panel and the Environmental Monitoring Panel were two Delphi-based inquiry panels that were deployed. Winter's Multiplicative predictions were modified using an AHP-based method based on the views of the second panel. We demonstrated that if a hotel uses this forecasting and adjustment procedure on a monthly basis, it may be utilized to estimate demand and assist management in preventing economic problems brought on by variations in demand. The model's ability to adapt to change and be improved upon in the future is its most crucial feature.

ii) Forecasting tourism demand with ARMA-based methods

https://www.sciencedirect.com/science/article/abs/pii/S0261517708001568

For the tourism and other hospitality sectors, the projection of tourist volume in the form of arrivals is particularly significant as it serves as a predictor of future demand, offering fundamental data for further planning and policy development. In this work, the number of international visitors to Hong Kong, Japan, Korea, Taiwan, Singapore, Thailand, the Philippines, Australia, and New Zealand is used to indicate the demand for tourism using three

univariate ARMA-based models. To guarantee the accuracy of the forecasting assessment, the study uses both monthly and quarterly time series produced from nine major tourist sites in the Asia-Pacific area. Additionally, performance forecasting based on disaggregated arrival series in a specific destination is investigated. It is generally believed that ARMA-based models work exceptionally well, and in certain instances, the mean absolute percentage error is less than 2%.

iii) Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels?

https://www.sciencedirect.com/science/article/abs/pii/S0148296317301376

The advantages of temporal aggregation for demand forecasting, such as enhanced accuracy, better stock control, and less modeling uncertainty, have been recently illustrated. A series undergoes transformation through temporal aggregation, which strengthens or weakens certain components and improves the ability to identify the time series structure. There are now two schools of thinking. Finding a single ideal temporal aggregation level at which a forecasting model maximizes accuracy is the main goal of the first. The second method, on the other hand, fits several models at various levels, each of which is able to capture distinct data properties. Although each strategy has advantages, they have only been studied separately thus far. From a theoretical and empirical standpoint, we evaluate and contrast them, talking about their advantages, contrasting the actual accuracy increases under various experimental configurations, and analyzing the business practice implications. To maximize the benefits of demand forecasting, we offer recommendations on when to utilize each.

iv) Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index

https://www.sciencedirect.com/science/article/abs/pii/S0261517718300505

Accurately predicting the number of tourists is a very difficult endeavor. The purpose of this research is to provide an efficient model, PCA-ADE-BPNN, for predicting the number of tourists based on the Baidu index. Prior to training a back propagation neural network (BPNN) architecture, the input data is decorrelated using principal component analysis (PCA), a dimensional reduction technique. The adaptive differential evolution algorithm (ADE) is used to obtain global optimization of the BP network's weight values and threshold values in order to improve the forecasting performance of BPNN. The PCA-ADE-BPNN model is a novel fusion of a neural network, an optimization technique, and a dimensional reduction algorithm. Case studies of China's Hainan Province and Beijing City are used to illustrate the model's validity. According to the findings, the suggested PCA-ADE-BPNN consistently performs better than other models in terms of predicting accuracy. As a result, the suggested PCA-ADE-BPNN is a viable option for accurately predicting the number of tourists.

v) Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model

https://ieeexplore.ieee.org/abstract/document/8260781

Decisions on demand management in the hotel sector are significantly impacted by cancellations of reservations. Hotels use strict cancellation policies and overbooking strategies to lessen the impact of cancellations, which may harm their bottom line and image. A machine learning-based system prototype was created to lessen this effect. In order to determine net demand, it uses data from the hotel's Property Management Systems to train a classification model daily that predicts which reservations are "likely to cancel." By using A/B testing, this prototype, which has been implemented in two hotels, allows for the evaluation of the effect of activities made to address reservations that are anticipated to be "likely to cancel." The results show that reservations that are contacted by hotels cancel less frequently than reservations that are not contacted, as well as strong prototype performance and significant implications for further study.

3. METHODOLOGY

a) Proposed Work:

By expanding an existing stacking-based ensemble model, the proposed study offers an improved and interpretable machine learning framework for forecasting hotel booking cancellations. This extension's main goal is to increase model resilience, computational efficiency, and prediction accuracy by dimensionality reduction and feature selection based on Principal Component Analysis (PCA). Prior to model training, PCA is used to reduce feature redundancy and noise by extracting the most informative components from the preprocessed hotel booking dataset.

Several machine learning classifiers, such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and Multilayer Perceptron, are trained with adjusted hyperparameters using the PCA-optimized feature set. A stacking meta-model is used to combine the top-performing models, and the final estimator is Logistic Regression. In comparison to the original ensemble model, our improved stacking technique improves generalization ability and attains higher prediction accuracy. SHapley Additive exPlanations (SHAP) are incorporated to understand model predictions and identify the main causes of booking cancellations in order to guarantee transparency and usefulness. With a 97.3% accuracy rate, the suggested extension shows better performance and offers both precise projections and useful information for efficient hotel revenue management.

b) System Architecture:

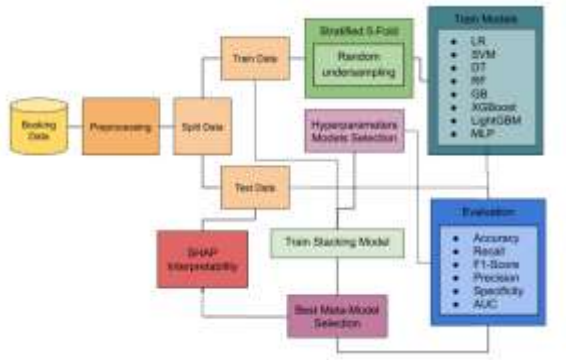


Fig 1 Proposed Architecture

In order to manage missing values, encode categorical characteristics, and normalize feature values, the system design starts with the gathering of hotel booking data, which is then preprocessed. To guarantee objective performance measurement, the processed dataset is then divided into training and testing subsets. The training data is subjected to

random undersampling and stratified 5-fold cross-validation in order to rectify class imbalance and enhance generalization. The suggested improvement uses Principal Component Analysis (PCA) to select features and reduce dimensionality during the training phase. This allows the model to concentrate on the most useful components while lowering noise and redundancy.

Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, and Multilayer Perceptron are among the machine learning models that are trained using adjusted hyperparameters after feature optimization. In order to generate final predictions, the top-performing models are then included into a stacking ensemble architecture, where Logistic Regression serves as the meta-classifier. Metrics including accuracy, precision, recall, F1-score, specificity, and AUC are used to assess the performance of the model. SHAP interpretability is integrated to evaluate feature contributions and elucidate individual predictions in order to improve transparency and reliability. In the end, the architecture chooses the optimal meta-model, producing predictions for hotel booking cancellations that are reliable, accurate, and understandable enough for practical implementation.

c) MODULES:

a) Importing the Packages

This module initializes the system by importing essential Python libraries required for data manipulation, visualization, machine learning, and model interpretability. These packages provide a stable computational environment for implementing predictive algorithms and explainable AI techniques.

b) Exploring the Dataset

In this module, the hotel booking dataset is analyzed to understand its structure, feature types, and data distribution. Exploratory analysis helps identify data patterns, inconsistencies, and dependencies that guide preprocessing and model selection decisions.

c) Pre-processing

The preprocessing module handles missing values, inconsistent formats, and noisy data to improve dataset quality. This step ensures reliable and structured inputs, enabling stable training and accurate booking cancellation prediction.

d) Visualization

Visualization techniques are used to analyze booking behavior, customer patterns, and cancellation trends. Graphical representations support intuitive understanding of influential features and assist in feature engineering and behavioral analysis.

e) Label Encoding and Normalization

Categorical features are transformed into numerical values using label encoding, followed by normalization to maintain consistent feature scales. These transformations enhance algorithm performance and support efficient model convergence.

f) Data Splitting (Train and Test)

The dataset is divided into training and testing subsets to evaluate model performance objectively. This module helps prevent overfitting and ensures that predictive results generalize well to unseen booking data.

g) Model Training

Multiple machine learning models are trained using historical booking records to learn cancellation patterns. This module forms the predictive core and enables comparative analysis across algorithms for stacking ensemble construction.

h) Evaluation

Model performance is evaluated using accuracy, precision, recall, and F1-score. These metrics provide objective assessment, guiding model optimization and selection of the best-performing meta-model.

i) Flask Server Deployment

The trained model is deployed using a Flask web server to enable real-time prediction through a user-friendly interface. This module supports practical application and operational accessibility.

j) User Login

A secure authentication mechanism controls system access by verifying user credentials. This module ensures data security and restricts prediction functionalities to authorized users.

k) Predict Booking Cancellation

Users upload test data through the web interface, and the trained stacking meta-model predicts cancellation likelihood. The module delivers fast, interpretable, and data-driven decision support.

l) Logout

The logout module securely terminates user sessions to prevent unauthorized access. It maintains system integrity and protects sensitive booking and prediction data.

d) Algorithms:

a) Logistic Regression

Logistic Regression is employed as a baseline classification technique to estimate the probability of booking cancellation by modeling the relationship between input features and class outcomes. It transforms linear combinations of predictors into probabilistic outputs using a logistic function and learns optimal coefficients that minimize classification error. The algorithm provides stable and interpretable predictions, making it suitable for binary classification tasks and supporting reliable decision-making within the ensemble framework.

b) Support Vector Machine (SVM)

Support Vector Machine classifies booking records by constructing an optimal hyperplane that maximizes the margin between cancellation classes. By identifying critical support vectors and applying kernel transformations, SVM effectively handles non-linear and high-dimensional data. Its margin-based optimization ensures strong generalization performance, making it robust in scenarios with overlapping or complex feature distributions.

c) Decision Tree Classifier

The Decision Tree classifier predicts booking cancellation by recursively partitioning data based on feature values that best distinguish target classes. Its hierarchical decision structure captures non-linear relationships without requiring feature scaling. The model offers high interpretability and adaptability, enabling transparent classification while efficiently handling both numerical and categorical attributes.

d) Random Forest

Random Forest enhances prediction accuracy through an ensemble of decision trees trained on randomly sampled subsets of data and features. The aggregation of multiple tree predictions using majority voting

reduces overfitting and improves generalization. Its ensemble structure provides robustness against noise and variability, making it suitable for reliable cancellation prediction.

e) Gradient Boosting Machine (GBM)

Gradient Boosting Machine builds a strong classifier by sequentially training weak learners that correct errors from previous iterations. Each model focuses on minimizing residual errors, allowing the ensemble to capture complex patterns in booking behavior. GBM improves accuracy by emphasizing hard-to-predict instances and maintaining an effective bias–variance balance.

f) XGBoost

XGBoost is an optimized gradient boosting algorithm that enhances performance through regularization, parallel processing, and efficient tree construction. It utilizes both first-order and second-order gradient information to refine predictions and prevent overfitting. The algorithm handles sparse data effectively and provides strong generalization, making it highly effective for structured cancellation prediction tasks.

g) LightGBM

LightGBM employs a leaf-wise tree growth strategy that improves accuracy by prioritizing splits with maximum information gain. Its histogram-based optimization reduces memory usage and accelerates training. The algorithm efficiently handles high-dimensional and imbalanced datasets, ensuring fast computation and reliable predictive performance.

h) Multilayer Perceptron (MLP)

The Multilayer Perceptron is a neural network-based classifier that captures complex non-linear relationships through multiple hidden layers. It learns hierarchical feature representations using weighted connections and gradient-based optimization. MLP enhances prediction accuracy by modeling high-order interactions, offering adaptability and strong generalization capability.

i) Stacking Meta-Model

The stacking meta-model integrates predictions from multiple base classifiers into a higher-level learning framework. Each base model is trained independently, and their outputs are combined using a meta-classifier to generate final predictions. This ensemble strategy leverages complementary strengths of individual algorithms, improving accuracy, stability, and robustness.

j) Optimized Stacking Meta-Model (Proposed Extension)

The optimized meta-model enhances the stacking ensemble by incorporating refined feature selection and tuning strategies. By integrating PCA-based dimensionality reduction and optimized meta-level learning, the model improves accuracy and computational efficiency while reducing sensitivity to noise. This enhanced ensemble achieves superior generalization and delivers consistent, high-quality booking cancellation predictions.

4. EXPERIMENTAL RESULTS

Using common classification measures including accuracy, precision, recall, and F1-score, the experimental assessment contrasts the performance of the extension-based optimal meta-model, the suggested stacking meta-model, and individual machine learning models. Baseline models such as SVM and logistic regression performed somewhat, as shown in Fig. X, suggesting a limited capacity to capture intricate booking-cancellation patterns. Tree-

based models like Random Forest and Decision Tree showed significant progress, with Random Forest's ensemble nature resulting in improved accuracy and balanced metric scores. By learning intricate non-linear correlations, boosting-based methods such as Gradient Boosting, XGBoost, and LightGBM improved predictive performance even more. However, differences in precision and recall show susceptibility to feature redundancy and class imbalance.

By skillfully integrating complementing model capabilities, the suggested stacking meta-model achieved around 96% accuracy with continuously high precision and recall values, far outperforming individual classifiers. With an accuracy of 97.3%, a better F1-score, and balanced precision–recall trade-offs, the extension-based optimal meta-model—which combines PCA-based feature selection with dimensionality reduction—provided the greatest overall performance. This enhancement demonstrates how removing superfluous features improves generalization ability and learning efficiency. The findings show that the PCA-optimized stacking architecture is appropriate for practical use in hospitality revenue management systems as it not only increases prediction accuracy but also offers a strong and dependable solution for hotel booking cancellation forecasting.

Accuracy: The capacity of a test to accurately distinguish between healthy and sick cases is a measure of its accuracy. We can determine a test's accuracy by calculating the proportion of reviewed cases with true positives and true negatives. In mathematical terms, this looks like:

Accuracy = TP + TN TP + TN + FP + FN.

$$Accuracy = \frac{(TN + TP)}{T}$$

F1-Score: One way to evaluate a model's performance in machine learning is via its F1 score. It takes a model's recall and precision and mixes them. Using the whole dataset, the accuracy metric determines the frequency with which a model produced accurate predictions.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

Precision: The proportion of occurrences or samples that are correctly classified out of all the ones that are classified as positive is known as precision. Consequently, the following is the formula for determining the accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The capacity of a model to detect all significant instances of a given class is measured by recall, a metric in machine learning. It sheds light on how well a model captures instances of a certain class and is calculated as the ratio of the total number of positive observations predicted to the number of actual positive observations.

$$Recall = \frac{TP}{(FN + TP)}$$



Fig2 upload dataset



Fig3 predicted results

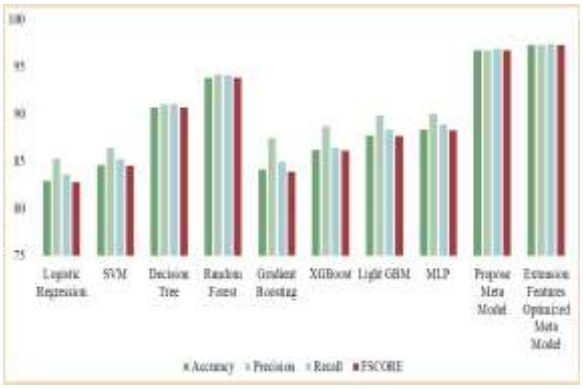


Fig4 performance comparision graph

5. CONCLUSION

Using Principal Component Analysis (PCA) for feature selection and dimensionality reduction, this study extended a stacking-based ensemble model to propose an improved hotel booking cancelation prediction framework. The suggested expansion outperformed the original ensemble method in terms of accuracy, resilience, and computing economy by retraining the stacking meta-model using PCA-optimized features. By identifying important factors impacting cancellations of reservations, the incorporation of SHAP-based interpretability further improved transparency and provided hotel management with actionable information. With an accuracy of 97.3%, the trial findings verified that the improved meta-model performed better than the others, proving its usefulness in providing precise, dependable, and understandable cancellation forecasts for practical hospitality applications.

6. FUTURE SCOPE

The suggested PCA-optimized stacking structure may be expanded in future research by adding streaming and real-time booking data to allow for dynamic cancellation prediction in operational hotel settings. To further improve prediction accuracy and flexibility, more sophisticated feature selection and representation learning strategies, including autoencoders or hybrid PCA–deep learning methods, might be investigated. Decisions on revenue management would be more practically relevant if cost-sensitive and profit-based assessment indicators were integrated. Furthermore, management decision support may be improved by expanding SHAP-based interpretability to include actionable suggestion systems for pricing and overbooking tactics. Additionally, the platform may be extended to cross-regional or multi-hotel datasets, allowing for wider use in a variety of hospitality sectors.

REFERENCES

[1] Ampountolas, A. (2025). Predicting hotel booking cancellations: a comprehensive machine learning

approach. Journal of Revenue and Pricing Management, 1-12.

[2] Ampountolas, A., & Legg, M. (2025). A comprehensive approach to enhancing short-term hotel cancellation forecasts through dynamic machine learning models. Tourism Economics, 13548166251318768.

[3] Theodorakopoulos, L., Kalliampakou, I., Ntantou, A., & Halkiopoulos, C. (2024, September). Leveraging Machine Learning for Sustainable Hotel Management: Predicting Booking Cancellations to Optimize Operations. In International Conference of the International Association of Cultural and Digital Tourism (pp. 135-178). Cham: Springer Nature Switzerland.

[4] Chen, S., Ngai, E. W., Ku, Y., Xu, Z., Gou, X., & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. Decision Support Systems, 170, 113959.

[5] Sánchez-Medina, A. J., & Eleazar, C. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. International Journal of Hospitality Management, 89, 102546.