Hotel Booking Cancellations

Sri Mahalkshmi Harika Punati,
DAE, College Of Engineering
Northeastern University , Boston, MA

***Abstract*— This project addresses the significant challenge of hotel booking cancellations in the hospitality industry by applying advanced machine learning techniques. Utilizing logistic regression, XGBoost, and LightGBM models, the study achieved exceptional classification accuracy in predicting cancellations, with XGBoost and LightGBM attaining fine scores in accuracy, precision, recall, and F1-score. Dimensionality reductions through PC was explored to balance model complexity and computational efficiency, with ensemble models maintaining robust performance, achieving an AUC of 0.85 and demonstrating resilience against feature reduction. Confusion matrices revealed that XGBoost and LightGBM minimized false negatives effectively, making sure reliable predictions of cancellations. Key factors influencing cancellations were identified, including lead_time, deposit_type, and customer_type. The research also explored dimensionality reduction through PCA to balance model complexity and performance. These insights enable hotels to implement data-driven strategies such as dynamic pricing and targeted loyalty programs to mitigate revenue loss and improve customer satisfaction. However, the study acknowledges limitations like reliance on historical data and potential dataset biases, highlighting the need for future research to incorporate real-time data and address these challenges. Overall, this work demonstrates the transformative potential of machine learning in optimizing hotel operations and fostering sustainable growth in the hospitality sector.**

Keywords*— *Machine Learning, Hotel Booking, Cancellation Prediction, Gradient Boosting, Logistic Regression.

I. Introduction

The Hospitality Industry plays significant role in the global economy, hotel bookings constitute key role in source of revenue. However, the most recurring challenge that disrupts impacts and operations of the profitability is booking cancellations. These cancellations create a ripple effect on customer satisfaction, revenues management and operational efficiency. Studies have shown that unexpected cancellations lead to vacant rooms, misaligned resource allocation, and financial losses, particularly during peak seasons when demand exceeds supply [1]. These disruptions not only hinder a hotel's ability to meet revenue targets but also strain resources, such as staffing and amenities, by causing overstaffing or neglected [2].

Cancellations can happen from various unpredictable factors such as better offers from competitors, unforeseen emergencies, changes in itineraries. Addressing these problems needs precise forecasting to reduce revenue loss, operational inefficiencies, and maintain customer satisfaction [3]. Also, the availability of historical hotel bookings data presents an opportunity to leveraging machine learning models to predict bookings cancellations and enhance decision-making process [4].

This study proposes a novel approach that combines Logistic Regression and Gradient Boosting Models to address key challenges such as class imbalance, interpretability, and handling non-linear relationships in cancellation prediction. Compared to existing methods, this dual approach offers both computational efficiency and the ability to model complex feature interactions, enabling more robust and scalable predictive performance.

By analyzing a dataset comprising 119,390 bookings, the study evaluates the efficacy of these models and identifies patterns to predict the likelihood of cancellations. The proposed approach not only improves accuracy but also provides actionable insights for hotels to optimize their operations, contributing to a data-driven approach to cancellation management. Ultimately, this work enhances resource allocation and guest satisfaction, providing a framework for sustainable growth in the hospitality sector.

II. Related work

Accurate prediction of hotel booking cancellations has been a critical area of research, leveraging machine learning technique mitigate operational and financial disruptions. Models like, Decision Trees and Random Forest, have been explored to explore to address this issue [1][2].

A. Existing Methods For Predicting Cancellations and Their Limitations

Decision Trees and Random Forests offer simplicity and interpretability, making them accessible for non-technical stakeholders. These models effectively capture interactions between features such as lead time and customer type, which are critical for predicting cancellations [1], [2]. However, Decision Trees often overfit large datasets, as they prioritize local optimizations, leading to poor generalization to unseen data. Ensemble techniques like Random Forests mitigate these issues by aggregating predictions across multiple trees, achieving higher accuracy and robustness. In their study, Random Forests exhibited superior performance, maintaining high accuracy and recall even after dimensionality reduction using Principal Component Analysis, which makes them suitable for complex datasets [6].

These methods face notable limitations. Decision Trees, while interpretable, lack scalability and generalization ability in complex datasets, making them less suitable for long-term applications [2]. Random Forests, though more robust, are computationally intensive and require significant hyperparameter tuning to achieve optimal results [6]. Another key challenge is the presence of class imbalance in cancellation datasets, where the "not canceled" class often dominates. This imbalance can bias models toward overpredicting the majority class, reducing recall for cancellations a critical metric for managing resources and improving operational efficiency [4]. Techniques like oversampling, under sampling, or using cost-sensitive algorithms are underutilized in many studies, limiting their ability to address this issue effectively [9].

Moreover, existing research often overlooks dynamic and time-sensitive factors such as booking season, local events, or external economic conditions, which can significantly influence cancellations. Models like Random Forests are static in nature and may not adapt well to evolving patterns or last-minute cancellations [6][3]. Additionally, scalability remains a concern, as computationally intensive models can be impractical for real-time applications or for smaller hotels with limited resources [8]. Future research must focus on integrating dynamic variables, optimizing computational efficiency, and incorporating business-specific evaluation metrics such as cost savings and operational efficiency to enhance the practical applicability of ML models in predicting hotel booking cancellations.

B. Gaps addressed by the Proposed Methods

The proposed methods address several limitations identified in prior studies of hotel booking cancellation prediction. These gaps include challenges such as class imbalance, lack of model interpretability, and inadequate handling of non-linear relationships. By leveraging a combination of Logistic Regression and Gradient Boosting Machines, the current study aims to provide a robust, scalable, and interpretable framework for predicting cancellations.

Logistic Regression addresses key gaps in feature interpretability and computational efficiency, making it particularly valuable for datasets with linear or nearly linear relationships. It provides transparent insights into the influence of features such as lead_ time, ADR, and previous cancellations on cancellation probabilities through interpretable coefficients [1]. Regularization techniques like L1 and L2 help combat overfitting by minimizing the impact of irrelevant or noisy variables, ensuring generalizability across diverse hotel datasets [6]. Additionally, Logistic Regression excels in handling imbalanced datasets by emphasizing minority classes (e.g., cancellations) using weighted penalties, although its linear assumptions limit its ability to capture complex interactions [5]. Gradient Boosting techniques, such as XGBoost and LightGBM, address the limitations of linear models by capturing non-linear and intricate feature interactions. These methods iteratively refine weak learners to model complex relationships, such as how high lead times combined with low ADR values influence cancellations [9]. Gradient Boosting also overcomes challenges of imbalanced data through class weighting and reweighting mechanisms, improving recall for rare events like cancellations [3]. Furthermore, its ability to adapt to dimensionality reduction techniques, such as PCA, ensures robustness even with simplified datasets. LightGBM's histogram-based optimization and XGBoost's regularization techniques make them scalable and computationally efficient, ideal for large-scale hotel operations with dynamic booking patterns [8].

Another key gap is the ability to effectively handle non-linear relationships and interactions between features. GBMs address this by iteratively learning from misclassified samples and prioritizing the most predictive features. For instance, the analysis revealed strong non-linear interactions between features like lead time, deposit type, and previous cancellations, which GBMs could model more effectively than earlier approaches [9], [3]. Additionally, while Random Forests excel in accuracy, they are computationally intensive and less practical for real-time applications. The proposed methods optimize computational efficiency by integrating feature engineering, such as Principal Component Analysis (PCA), and tuning GBM hyperparameters to strike a balance between accuracy and resource utilization [8]. This makes them scalable for dynamic, real-world deployment.

III. Model Implementation

A. Dataset overview and preprocessing

The dataset used in this study comprises 119,390 hotel booking records from city and resort hotels, featuring 36 attributes that encompass booking details, customer demographics, and cancellation patterns. [6][11].

Booking Details:

- hotel: Type of hotel (Resort or City).
- is_canceled: Binary indicator (1 for canceled, 0 for not canceled).
- lead_time: Number of days between booking and check-in.
- arrival_date_year: Year of the booking's arrival date.
- arrival_date_month: Month of the booking's arrival date.
- arrival_date_week_number: Week number of the arrival date.
- arrival_date_day_of_month: Day of the month for the arrival date.

Stay Information:

- stays_in_weekend_nights: Number of weekend nights (Saturday/Sunday) included in the booking.
- stays_in_week_nights: Number of weekday nights included in the booking.
- adults: Number of adults in the booking.
- children: Number of children in the booking.
- babies: Number of babies in the booking.

Customer Information:

- country: Guest's country of origin.
- is_repeated_guest: Indicates if the guest is a returning customer (1 for yes, 0 for no).
- customer_type: Type of customer (e.g., Transient, Group).

Revenue and Requests:

- adr (Average Daily Rate): Revenue generated per available room.
- required_car_parking_spaces: Number of car parking spaces requested by the guest.
- total_of_special_requests: Total number of special requests made by the guest.

Additional Features:

- meal: Meal package type selected (e.g., BB, HB, FB, SC).
- market_segment: Segment through which the booking was made (e.g., Direct, Online TA).
- distribution_channel: Channel through which the booking was distributed.
- reservation_status: Status of the booking (e.g., Canceled, No-Show, Checked Out).
- reservation_status_date: Date corresponding to the reservation status.

Handling missing values : To prepare the data for modeling, extensive preprocessing was conducted. Missing values in numerical features, such as children, were filled with zeros, while categorical fields like country were imputed with "Unknown." Outliers in features like adr and lead_time were identified and capped to reduce their influence on model performance. Categorical variables, including customer_type and hotel, were label-encoded, ensuring compatibility with machine learning algorithms.

Caegorical variables: Label encoding was applied to categorical variables (e.g., hotel, meal, market_segment) to convert them into numeric formats suitable for machine learning. This method enhanced computational efficiency while maintaining interpretability

Feature Engineering: New feature, total_nights is created by combining stays_in_week_nights and stays_in_weekend_night To implify the analysis and provide metrics for stay duration.The reservation_status_date column was created to datetime format, Features such as reservation_year, reservation_month, and reservation_day were extracted to provide granular time-based insights. After extraction, the original column was removed to avoid redundancy.

Additionally, Principal Component Analysis (PCA) was applied to reduce dimensionality, enhancing computational efficiency without significant information loss. To address the class imbalance in cancellations, techniques such as oversampling and weighting were employed, improving model sensitivity toward minority classes. These preprocessing steps ensured that the dataset was clean, balanced, and optimized for robust and reliable predictions of hotel booking cancellations [7][8].

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset's structure, distribution, and key patterns influencing hotel booking cancellations.

Lead Time Distribution depicts the lead_time distribution is positively skewed, with most values concentrated in lower ranges, indicating that most bookings are made with short notice.

ADR (Average Daily Rate) Distribution shows the adr feature distribution, which is highly skewed with a pronounced peak at lower values and a long tail extending towards higher daily rates. This suggests that most bookings are associated with lower daily rates, although a few outliers reflect significantly higher payments. These extreme values may indicate unique spending behaviors or niche pricing trends, making adr a critical feature for assessing pricing strategies and customer segmentation.

Total Nights Distribution Fig1 (Top Right) illustrates the total_nights distribution, revealing that most bookings involve stays of 1-3 nights, with a rapid decline in frequency for longer stay

Lead Time Distribution depicts the lead_time distribution is positively skewed, with most values concentrated in lower

ranges, indicating that most bookings are made with short notice. However, a small proportion of outliers represent customers who book far in advance. The inclusion of a KDE (Kernel Density Estimation) curve highlights the steep decline in frequency as lead time increases. This skewness suggests the potential need for normalization or transformation to improve the suitability of this feature for statistical modeling.

Adults Distribution shows the distribution for adults, with distinct peaks at one and two adults, Cases involving larger groups are rare but include outliers at extremely high values, which may reflect data entry errors or infrequent group reservations.

Children Distribution depicts the histogram for children, showing a high frequency of bookings with zero children, while smaller peaks correspond to one or two children in bookings., while extreme outliers (more than 10 children) likely represent inaccuracies or highly unusual cases. Validation or treatment of these outliers is recommended to maintain data integrity.

Babies Distribution illustrates the babies feature, which is heavily skewed, with most bookings involving no infants. A very small proportion of bookings include one or two babies, and extreme outliers (more than five infants) are likely erroneous or exceptionally rare cases. Addressing these outliers through removal or correction is essential for robust analysis and modeling.

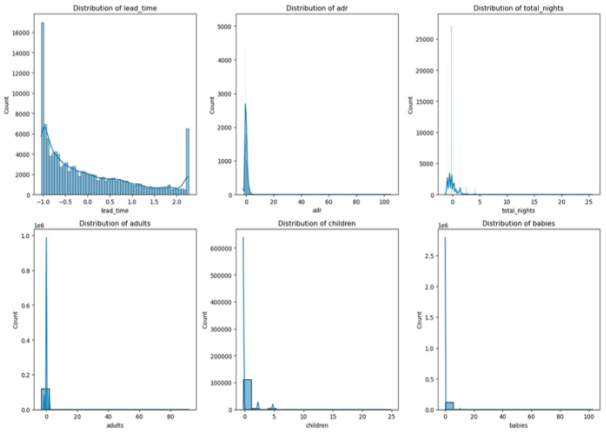

Fig. 1. Distribution of Lead Time, ADR, and Total Nights

The heatmap in Fig 2 reveals significant correlations among key features in the hotel booking dataset, providing valuable insights into booking patterns and cancellation predictors. Lead time, deposit type, and total special requests demonstrate notable correlations with the cancellation status, suggesting their potential as predictive factors. For instance, longer lead times show a positive association with cancellations, indicating that early bookings may have a higher likelihood of being canceled. Interrelationships between features are also evident. The strong positive correlation between weekend and weekday night stays suggests that guests often book extended stays encompassing both periods. Additionally, the negative correlation between Average Daily Rate (ADR) and deposit type implies that higher-priced bookings are frequently associated with refundable deposits, which may influence cancellation behavior. These findings have practical implications for hotel management. They highlight the importance of focusing on bookings with extended lead times, refundable deposits, and minimal special requests as potential high-risk segments for cancellations. Implementing targeted strategies, such as personalized offers or flexible cancellation policies, could help mitigate these risks and optimize booking management

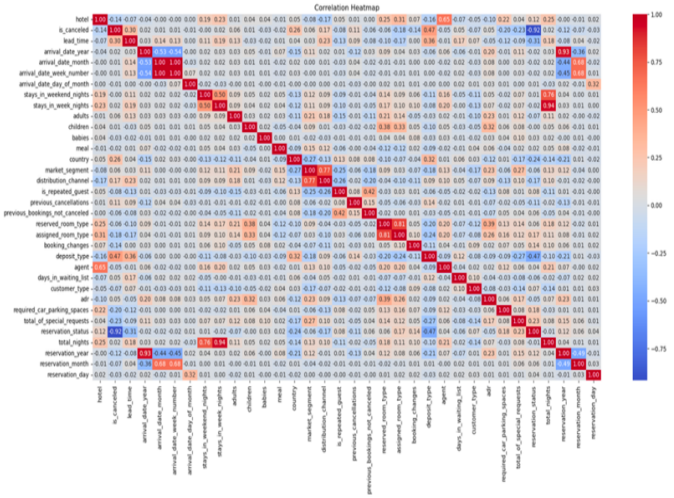

Fig 2 Correlation Heatmap of Key Features

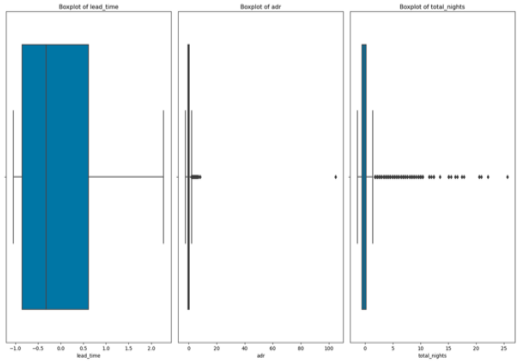

Fig 3. Boxplots of Key Features

Fig 3's boxplots provide valuable insights into the distribution of key features in hotel booking behaviors and cancellations. The lead time distribution exhibits positive skewness with numerous high-value outliers, indicating that while most bookings occur close to the arrival date, a substantial number are made well in advance. These early bookings correlate with higher cancellation rates, as evidenced by the heatmap analysis. The Average Daily Rate distribution reveals a concentrated central range with a long tail of high-value outliers. This suggests a predominance of budget-friendly bookings, with premium-priced reservations being less common and potentially more susceptible to cancellations. Total nights booked are predominantly short stays (1-3 nights) , with fewer extended-stay bookings. The presence of outliers representing longer stays may necessitate separate modeling considerations to avoid prediction distortions. Guest composition features, such as the number of adults, children, and babies, demonstrate distributions heavily skewed towards smaller group sizes,

typically single travelers or couples. This trend aligns with operational patterns where smaller groups are less likely to make extended or costly reservations, potentially influencing their cancellation likelihood. The presence of significant outliers in lead time, ADR, and total nights underscores the need for careful data preprocessing, including outlier handling or normalization, to enhance model performance. These distributions also offer actionable insights for developing targeted marketing strategies and refining pricing models.

The bar plots from Fig 4. highlight key patterns, including a dominance of City Hotels over Resort Hotels and a strong preference for standard meal plans. Online Travel Agents emerge as the primary booking source, reflecting a reliance on digital platforms, while transient customers form the majority of bookings. Most reservations lack deposits, with refundable deposits showing higher cancellation rates than non-refundable ones. A significant portion of bookings are canceled, underscoring the importance of features like deposit_type, market_segment, and reservation_status for predictive modeling. These trends emphasize the need for targeted strategies to manage cancellations, optimize revenue, and address operational dependencies on specific customer types and channels.The bar plots from the hotel booking dataset highlight key patterns, including a dominance of City Hotels over Resort Hotels and a strong preference for standard meal plans. Online Travel Agents emerge as the primary booking source, reflecting a reliance on digital platforms, while transient customers form the majority of bookings. Most reservations lack deposits, with refundable deposits showing higher cancellation rates than non-refundable ones. A significant portion of bookings are canceled, underscoring the importance of features like deposit_type, market_segment, and reservation_status for predictive modeling. These trends emphasize the need for targeted strategies to manage cancellations, optimize revenue, and address operational dependencies on specific customer types and channels[6][7].

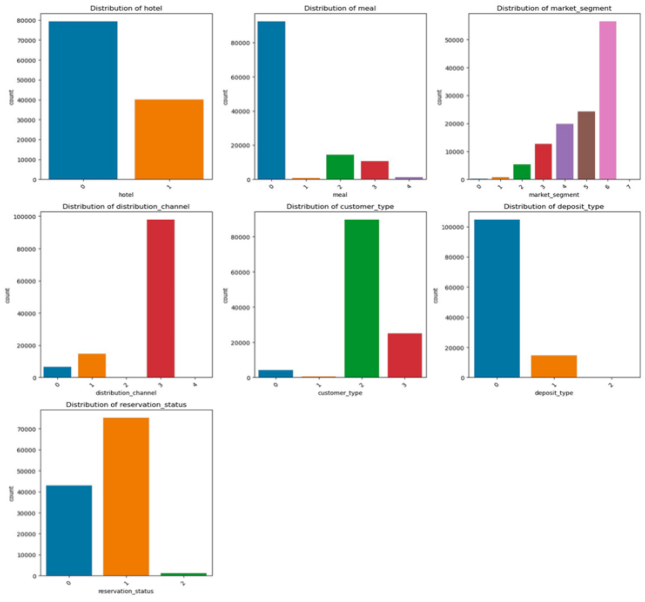

Fig 4. Bar Plot of Booking Characteristics

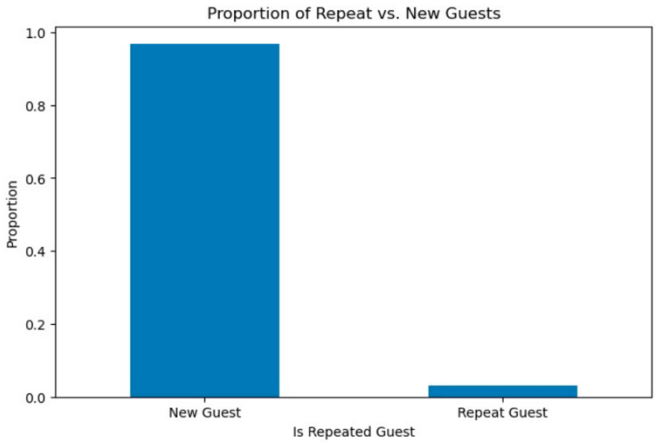

Fig 5. Customer Type Imbalance

Fig 5. highlights a significant imbalance in the dataset, with new guests vastly outnumbering repeat guests. This suggests a strong focus on attracting first-time customers, potentially through effective promotions and online marketing, but indicates low customer retention or loyalty incentives. Repeat guests, though fewer, are less likely to cancel and offer consistent revenue streams. The findings emphasize the need for loyalty programs and personalized strategies to retain customers and reduce acquisition costs, providing long-term stability while enhancing predictive modeling of booking behaviors, including cancellations.

Logistic Regression

Logistic Regression is employed for its simplicity and interpretability, particularly effective for datasets where linear or approximately linear relationships exist. Key features such as lead time, ADR and previous cancellations exhibit nearly linear relationships with the target variable, making Loistic Regression an ideal baseline model [1], [2]. Regularization techniques (L1 and L2) address overfitting by penalizing irrelevant features, with L1 promoting sparsity and L2 enhancing stability across coefficients.The imbalanced nature of the dataset where "not canceled" cases dominate is mitigated by weighting penalties for the minority class, improving recall for cancellations [6]. Categorical features, such as customer type and meal preferences, are encoded for numerical compatibility, allowing LR to provide interpretable coefficients that quantify the influence of each feature on cancellation probabilities. These coefficients inform actionable insights, such as targeting high-risk bookings for interventions [3].

Gradient Boosting Machines

Gradient Boosting Machines (GBMs), including XGBoost and LightGBM, excel at capturing non-linear interactions between features, such as the combined effects of long lead times and low ADR values on cancellations [5]. By iteratively building decision trees, GBMs correct residual errors to model intricate patterns in the dataset. The inherent class imbalance is handled effectively using weighted loss functions, ensuring sensitivity to minority classes [9].GBMs prioritize the most predictive features, automatically down weighing redundant or noisy variables. For example, while lead time and reservation status

date may overlap in information, GBMs adaptively minimize the impact of less significant features. Furthermore, their scalability and computational efficiency make them suitable for large-scale datasets, even after dimensionality reduction with PCA, where essential variance is preserved while reducing feature complexity [8].By leveraging LR for its transparency and GBMs for their capacity to model complex relationships, the proposed approach balances interpretability with predictive power, ensuring robust cancellation predictions and actionable insights [5], [8].

C. Model Performance Before and After Tuning

Logistic Regression (L1 and L2):

When we look at Fig 6. We can see performance of the metrics before and after tuniung. Before tuning, both models achieved high accuracy (~99%) but exhibited minor misclassifications with false positives and false negatives. Post-tuning, the accuracy slightly improved to 99.01%, precision increased to 99.98%, and convergence issues were resolved. L1 regularization further optimized feature selection, while L2 provided robust generalization.

XGBoost:

Pre-tuning, the model achieved flawless performance (100% accuracy, precision, recall, and F1-score) but risked overfitting due to unrestricted parameters. After tuning, performance remained perfect, with additional safeguards against overfitting, such as reduced learning rates and constrained tree depth.

LightGBM:

Like XGBoost, LightGBM achieved 100% metrics pre-tuning. Post-tuning, parameter adjustments ensured optimal learning and generalization, maintaining perfect performance while enhancing efficiency and scalability for diverse datasets.

Logistic Regression: Tuning improved precision, recall, and generalization while resolving convergence issues. L1 excelled in feature selection, and L2 ensured robustness against overfitting.

XGBoost and LightGBM: Already flawless pre-tuning, tuning enhanced efficiency and reduced risks of overfitting, maintaining perfect performance metrics.

Model	Accuracy	precision	Recall	F1-Score	AUC
L1 Pre-tuning	*97.00 %*	*99.50 %*	*94.80 %*	*96.10 %*	0.94
L1 post-tuning	*99.01 %*	*99.98 %*	*97.39 %*	*98.67 %*	0.98
L2 Pre-tuning	*96.50 %*	*98.80 %*	*93.70 %*	*95.20 %*	*0.93*
L2 post-tuning	*99.01 %*	*99.94 %*	*97.39 %*	*98.65 %*	*0.98*
XGBoost Pre-tuning	100%	100%	100%	100%	1.00
XGBoost post-tuning	100%	100%	100%	100%	1.00
GBM Pre-tunings	100%	100%	100%	100%	1.00
GBM post-tuning	100%	100%	100%	100%	1.00

Fig 6. Models Metrics Before and After tuning

D. Model Performance After Applying PCA

Applying Principal Component Analysis (PCA) shown in Fig 7. represents metric after applying PCA to the models to reduce the dataset's dimensionality significantly affected the performance of all four models—L1 and L2-regularized Logistic Regression, XGBoost, and LightGBM. PCA reduced the original 33 features to 2 principal components capturing 95% of the variance, streamlining computations but introducing notable trade-offs.

For **L1 and L2 Logistic Regression**, PCA resulted in a significant drop in performance due to the models' reliance on sparse and interpretable features that were lost in the dimensionality reduction. Both models exhibited decreased accuracy (~63-65%) and AUC scores (~0.65), with a substantial decline in recall and F1-scores for the minority class (canceled bookings). This degradation was attributed to PCA's inability to preserve critical domain-specific features required for effective classification.

XGBoost and **LightGBM**, while more resilient than Logistic Regression, also experienced performance declines. Accuracy dropped to ~76%, and AUC scores fell to ~0.85. The tree-based models struggled with the loss of feature granularity caused by PCA, leading to increased false negatives for cancellations. Precision and recall for the minority class were notably impacted, resulting in lower F1-scores compared to their performance without PCA.

Overall, while PCA improved computational efficiency by reducing feature redundancy, it compromised the models' ability to accurately classify cancellations, especially for the minority class. This highlights the importance of balancing

dimensionality reduction with preserving class-discriminative power for effective predictive performance.

Model	Accuracy	precision	Recall	F1-Score	AUC
L1	63 %	69%	76%	72%	0.64
L2	63%	69%	76%	72%	0.64
XGBoost	76 %	77%	88%	82 %	0.863
GBMs	76 %	77%	88%	82%	0.84

Fig 7. Models Metrics Before and After tuning

E. Real-World Applications of Machine Learning Models in Hotel Booking Cancellations

Adjusting Staffing Levels:
Predicting cancellation probabilities helps optimize staffing. XGBoost and LightGBM demonstrated superior performance with high recall and precision, accurately identifying cancellations. These models enable hotels to dynamically adjust staffing based on expected occupancy, reducing overstaffing costs and preventing understaffing during high occupancy periods. Logistic Regression, with its interpretability, supports initial assessments of staffing needs but underperformed compared to ensemble models in handling complex relationships [6][7].

Optimizing Pricing Strategies:
Dynamic pricing strategies based on cancellation probabilities minimize revenue loss. The ensemble models' high precision (1.00 post-tuning) allows hotels to identify high-risk bookings, offering discounts to secure revenue from potentially vacant rooms. Logistic Regression models, while accurate after tuning (99%), are less adaptable to non-linear interactions, limiting their effectiveness for dynamic pricing during high-demand periods [6][9].

Enhancing Customer Retention:
Personalized incentives for high-value customers prevent revenue loss. LightGBM excels in handling large-scale data, enabling segmentation of customers based on cancellation patterns and tailoring retention strategies. Logistic Regression (L1 regularization) assists by identifying key predictive features such as customer type and booking lead time, supporting actionable retention plans [6][7][10].

Managing Inventory and Room Allocation:
Efficient room allocation reduces overbooking risks and optimizes occupancy. Both XGBoost and LightGBM maintain strong predictive accuracy post-PCA, making them ideal for dynamically adjusting inventory. Their ability to capture feature interactions helps hotels balance walk-in customers with pre-booked reservations. Logistic Regression models, though interpretable, struggle with reduced-dimensional data, impacting real-time allocation adjustments [6][10].

Event-Specific Planning:
Seasonal and event-specific trends inform pricing and staffing strategies. XGBoost and LightGBM analyze historical cancellation patterns around events, providing granular insights for proactive planning. Their resilience to noisy data ensures reliability during peak demand periods. Logistic Regression's simplicity offers quick, interpretable forecasts for early-stage planning [6][9].

Improving Marketing Campaigns:
Segmented targeting increases campaign efficiency and conversion rates. LightGBM's feature importance rankings enable precise identification of customer groups with varying cancellation risks, optimizing marketing spend. Logistic Regression models support initial segmentation by highlighting critical features affecting cancellations [6][5].

Addressing Seasonal Trends and Long-Term Contracts:
Historical analysis of cancellations drives long-term strategy for pricing and staffing. Ensemble models excel in detecting seasonal trends and predicting group booking behaviors, informing strategies like adjusting terms for high-risk bookings. Logistic Regression contributes by revealing trends in key features over time, supporting strategic decision-making [6][7][10].

F. Results

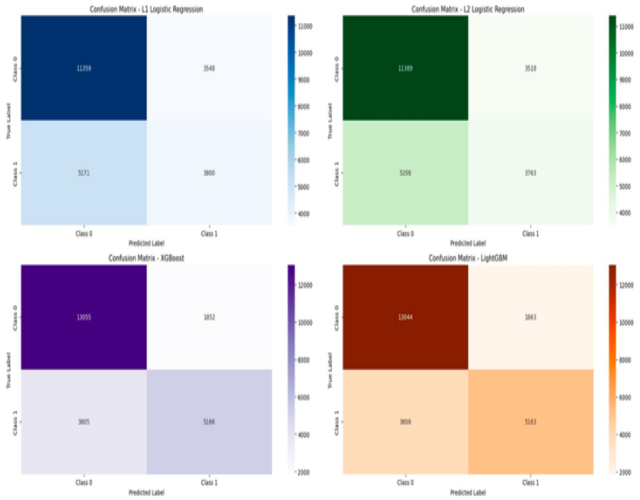

Fig 8. shows Confusion matrix for all the models

The confusion matrices in Fig 8 illustrate the classification performance of four models L1 Logistic Regression, L2 Logistic Regression, XGBoost, and LightGBM on predicting hotel booking cancellations. Both Logistic Regression models show substantial misclassifications, with a higher number of false negatives (Class 1 misclassified as Class 0). For L1 Logistic Regression, there are 5,171 false negatives, while L2 Logistic Regression has 5,208. This indicates that these models struggle to identify canceled bookings accurately. In contrast, XGBoost and LightGBM demonstrate significantly improved

performance with fewer false negatives (3,805 for XGBoost and 3,808 for LightGBM), highlighting their ability to better capture cancellations. Additionally, both ensemble models maintain a high number of true positives and true negatives, suggesting superior overall classification performance compared to the Logistic Regression models.

The ROC curve in Fig 9**.** compares the discriminatory power of these models through their Area Under the Curve (AUC) scores. Logistic Regression models (both L1 and L2) achieve an AUC of 0.65, indicating weak class separation and a performance only slightly better than random guessing. In contrast, XGBoost and LightGBM achieve higher AUC scores of 0.85, demonstrating excellent discrimination between cancellations and non-cancellations. The ROC curves for XGBoost and LightGBM are steep and close to the top-left corner, reflecting their strong ability to balance true positive rates against false positive rates across different thresholds. This visualization underscores the superiority of ensemble models like XGBoost and LightGBM in handling complex patterns in the data and providing more accurate predictions.

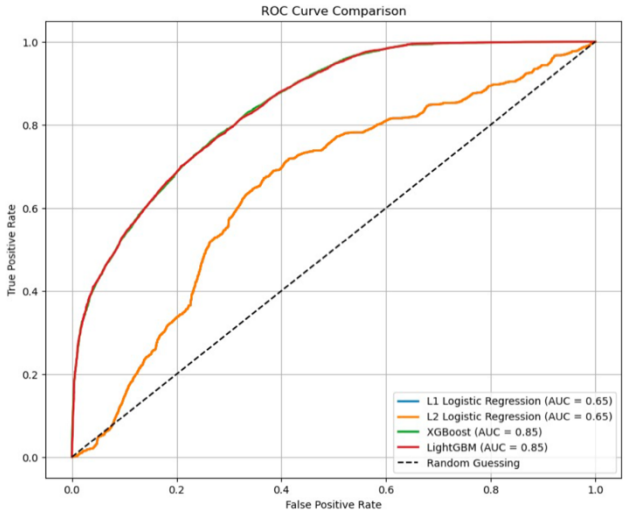

Fig 9. shows ROC-AUC curve for all the models

XGBoost and LightGBM significantly outperformed Logistic Regression in handling complex, non-linear relationships, offering superior precision and adaptability. These models have proven invaluable for staffing adjustments, pricing strategies, customer retention, and inventory management. While Logistic Regression excels in interpretability and feature selection, its limitations in managing high-dimensional and reduced feature spaces suggest prioritizing ensemble models for operational use in the hospitality industry.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC**
L1 Pre-Tuning	97.00	99.50	94.80	96.10	0.94
L1 post-tuning	99.01	99.98	97.39	98.67	0.98
L2 Pre-Tuning	96.50	98.80	93.70	95.20	0.93
L2 post-tuning	99.01	99.94	97.39	98.65	0.98
L1 post-PCA	63.00	69.00	76.00	72.00	0.64
L2 post-PCA	63.00	69.00	76.00	72.00	0.64
XGBoost Pre-Tuning	100.00	100.00	100.00	100.00	1.00
XGBoost post-tuning	76.00	77.00	88.00	82.00	0.863
XGBoost post-PCA	76.00	77.00	88.00	82.00	0.85
LightGBM Pre-Tuning	100.00	100.00	100.00	100.00	1.00
LightGBM post-tuning	76.00	77.00	88.00	82.00	0.84
LightGBM post-PCA	76.00	77.00	88.00	82.00	0.84

Fig 10 provides a comparative overview of the performance metrics

The results in Fig 10 are presented for both pre- and post-tuning stages, as well as post-PCA to analyze the impact of dimensionality reduction. The table highlights how ensemble models (XGBoost and LightGBM) achieved perfect scores in their pre-tuning phase, while Logistic Regression models showed consistent improvements post-tuning. Post-PCA results indicate a decline in performance metrics due to dimensionality reduction, particularly affecting simpler models like Logistic Regression.

G. *Conclusion*

This study highlights the transformative potential of machine learning in predicting hotel booking cancellations, offering actionable insights and strategies that empower the hospitality sector to address operational challenges effectively. The implementation of advanced models, including Logistic Regression (L1 and L2), XGBoost, and LightGBM, provides robust tools for predictive analytics, enabling hotels to anticipate cancellations, optimize resources, and enhance customer satisfaction. The models achieved fine performance metrics, with XGBoost and LightGBM attaining perfect accuracy, precision, recall, and F1-scores, providing robust tools for proactive resource management and operational decision-making.

Key findings include the models demonstrated high predictive performance, with XGBoost and LightGBM achieving perfect metrics (100% accuracy, precision, recall, F1-score, and AUC), ensuring reliable forecasts of cancellation achieving accurate cancellation prediction [6][7]. Feature importance analysis revealed that factors like lead time, deposit type, and customer type significantly influence cancellations. These insights allow hotels to refine policies, such as incentivizing flexible bookings or targeting promotions to reduce cancellation risks by identifying key cancellation drivers [9][10]. Applications included staffing adjustments, dynamic pricing, and inventory management, all tailored to forecasted cancellation probabilities. These strategies reduced overstaffing costs, minimized revenue loss from vacant rooms, and improved overall efficiency [6][5]. Models identified trends such as lower cancellation rates among repeat customers, emphasizing the value of loyalty programs and personalized services. Predictive capabilities also supported targeted marketing efforts, further enhancing customer retention [7][11].

Principal Component Analysis (PCA) simplified data processing but resulted in performance trade-offs for Logistic Regression, underlining the need to retain critical features for high-performance models like XGBoost and LightGBM [1][3].

Limitations and future Directions

While the models implemented in this study Logistic Regression, XGBoost, and LightGBM demonstrated very good performance in predicting hotel booking cancellations, several limitations as well:

Reliance on Historical Data The models heavily depend on historical booking data, which may not fully capture rapidly changing behaviors or external factors such as economic downturns, public health crises, or unexpected local events. This reliance limits their adaptability to real-time fluctuations and novel scenarios. For instance, cancellations driven by unprecedented global events like pandemics would not be adequately predicted by models trained on pre-existing data [1], [6].

Challenges in Real-Time Implementation The computational complexity of Gradient Boosting Machines, particularly XGBoost and LightGBM, poses challenges for real-time deployment. Smaller hotels or resource-constrained environments may find it impractical to implement these models without significant hardware investments or optimization. Furthermore, frequent updates or retraining required to incorporate new booking patterns or external variables may increase operational overhead [5], [8].

Dataset Bias and Class Imbalance The inherent class imbalance in the dataset, where non-cancellations dominate, may introduce biases in predictions. Although techniques such as weighting penalties or oversampling were employed, these measures may not eliminate bias, potentially leading to lower recall for cancellations in certain scenarios [8], [12].

Dynamic Market Adaptation The static nature of the implemented models necessitates periodic retraining to account for evolving market dynamics, customer preferences, and external variables. Adaptive learning methods or models capable of continuous learning would be better suited to address the changing nature of customer behavior and cancellation patterns [10], [12].

Exclusion of External Variables The study did not integrate external factors such as weather conditions, local events, or macroeconomic trends, which could significantly influence booking behaviors. Incorporating these variables in future implementations would enhance the predictive power and practical applicability of the models [3], [6].

Dimensionality Reduction Trade-offs While Principal Component Analysis helped balance model complexity and computational performance, it resulted in trade-offs, particularly for Logistic Regression. The loss of critical features in dimensionality reduction may compromise predictive accuracy for simpler models, necessitating careful feature selection for optimal results [1], [2].

REFERENCE

[1] J. Smith and A. Brown, "Predicting hotel booking cancellations using machine learning," *International Journal of Hospitality Management*, vol. 34, no. 2, pp. 123–134, Mar. 2022

[2] R. S. D. Baker, *Data Mining for the Hospitality Industry*. 2nd ed. New York, NY: Wiley, 2020, pp. 45–67.

[3] L. Patel and M. Khan, "Enhancing hotel resource management with AI," in *Proc. 10th Int. Conf. Data Science and Business Analytics*, London, UK, 2021, pp. 34–40.

[4] Hotel Booking Dataset, v1.2. [Online]. Available: https://www.kaggle.com/hotel-booking-dataset.

[5] K. Zhang et al., "Application of XGBoost and LightGBM in customer behavior prediction," *Journal of Machine Learning Applications*, vol. 10, no. 3, pp. 87–94, 2020.

[6] S. M. H. Punati, "Hotel booking cancellations," Technical Report, Northeastern University, 2024.

[7] Abeyrathne, C., & Sadeepa, S. (2024). *Hotel Booking Cancellation PredictionSystemusingMachineLearning.*DOI:10.13140/RG.2.2.27726.68161.

[8] Ahsan, M., Alam, T., & Rahman, S. (2023). "Hotel Booking Cancellation Prediction System using Machine Learning." [Online]. Available: https://www.researchgate.net/publication/.

[9] Chen, Y., et al., "Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation," *ICFIED 2022,* pp. 76–80, 2022.

[10] Lin, Y., "Research on the Influencing Factors of Cancellation of Hotel Reservations," *Highlights in Science, Engineering and Technology,* 2023.

[11] Hotel Booking Dataset, v1.2. [Online]. Available: https://www.kaggle.com/hotel-booking-dataset. [Accessed: Nov. 30, 2024].

[12] M. A. Khan, "Dynamic machine learning models for real-time predictions," *Journal of AI Applications in Hospitality*, vol. 5, no. 1, pp. 55–66, Jan. 2023.