Predictive Analysis for Hotel Booking Cancellations and Guest Behavior

1st Rakesh Nandan Anantharam
Engineering Science - Data Science
University at Buffalo - SUNY)
Buffalo, United States
ranantha@buffalo.edu

2nd Afreen
Engineering Science - Data Science
University at Buffalo - SUNY)
Buffalo, United States
afreen@buffalo.edu

3rd Kousalya Kethavath
Engineering Science - Data Science
University at Buffalo - SUNY
Buffalo, United States
kousalya@buffalo.edu

***Abstract*—This project analyzes hotel booking demand data to identify key factors influencing guest bookings and predict cancellations, leveraging data analysis and machine learning for optimized booking management. The analysis is based on a dataset that includes information such as booking lead time, hotel type, market segment, booking cancellation and total stay duration. Through extensive exploratory data analysis (EDA), we identified patterns such as the dominance of small group bookings and the relatively limited impact of lead time on guest numbers. Feature engineering was employed to enhance the dataset, creating new variables such as total stay, lead time categories between hotel type and stay duration. This project provides insights that can assist hotel managers in optimizing operations, managing inventory, and catering to demand fluctuations, especially for group and last-minute bookings.**

I. Problem Statement

The hospitality industry is significantly impacted by the ability to predict customer behavior, especially when it comes to understanding booking demand and cancellations. Predicting hotel booking demand and cancellation rates enables hotels to optimize their inventory management, pricing strategies, and revenue streams. Moreover, understanding the factors influencing cancellations can help hotels implement proactive measures to minimize revenue loss and operational inefficiencies.

Cancellations in the hospitality industry lead to vacant rooms, lost revenue, and challenges in resource allocation. Unpredictable cancellations can result in over-staffing, wasted resources, and lost opportunities to accommodate other guests. This problem is particularly pressing during peak travel seasons or special events when maximizing occupancy is critical for financial success. Therefore, understanding the factors behind cancellations is essential to help hotels better manage their operations, avoid losses, and improve customer satisfaction.

In this project, the objective is to conduct an Exploratory Data Analysis (EDA) on the Hotel Booking Demand dataset and build a model to predict whether a booking will be canceled. By leveraging customer demographics, booking details, and seasonal patterns, this analysis will provide insights into key trends and behaviors, enabling hotels to anticipate cancellations more effectively. Accurate predictions will allow hotels to implement strategies such as overbooking, targeted marketing for at-risk customers, or dynamic pricing to mitigate cancellation impacts. The insights gained from this project can contribute to broader applications in customer behavior analytics and operational optimization, making it a valuable tool for improving hotel management strategies.

The key questions to address are:

- What factors contribute to booking cancellations?
- Can we accurately predict whether a hotel booking will be canceled using available data?
- How do seasonal trends and customer preferences influence hotel booking demand?

II. Data Sources

A. Hotel Booking Demand Dataset (Kaggle):

This dataset contains detailed information about hotel bookings, including variables such as booking lead time, hotel type, market segment, and the number of guests. It is publicly available on Kaggle.

B. Source:

The dataset used in this project was sourced from Kaggle, https://www.kaggle.com/datasets/ahmedsafwatgb20/hotel-bookingscsv. The dataset contains 119,391 rows and 32 features, which is sufficient for conducting exploratory data analysis and building predictive models. Some of the key features include:

- Hotel: Resort Hotel or City Hotel.
- Lead Time: Number of days between booking and arrival.
- Stay Duration: Number of nights stayed (weekend and weekdays).
- Market Segment: How the booking was made (Direct, Corporate, Groups, etc.).
- Total Guests: Number of adults, children, and babies.

III. Data Cleaning and Processing

A. Removing duplicates:

We found and removed 31,994 duplicate rows to avoid bias in our analysis.

B. Handling missing values:

Certain columns had missing values, which were handled as follows:

- 'children' column had 4 missing values, which were filled with the most frequent value.
- 'country' column had 488 missing values, were replaced with the most frequent value (mode).
- 'agent' and 'company' columns had many missing values, so they were filled with 0 (assuming no agent or company involvement).

C. Handling Inconsistent Data:

The dataset included rows where both adults, children, and babies were zero, representing invalid entries. Such records were removed using the filter.After removing these invalid rows, the dataset was updated.

D. Negative ADR (Average Daily Rate):

The Average Daily Rate (ADR) represents the average amount paid per room per day by a guest. In this dataset, there is an anomalous minimum ADR value of -6.38, which is not realistic and likely represents a data entry error. Additionally, there is an unusually high ADR value of 5400, which is a significant outlier compared to the rest of the data. To ensure the accuracy of the analysis, these erroneous and outlier values are removed from the dataset.

Similarly, unusually high values for children and babies were removed due to their potential as outliers. Additionally, bookings that include babies but no adults are illogical and likely reflect data entry errors.

E. Dropping Invalid Rows:

Bookings where the number of adults, children, and babies were all zero were considered invalid, as it implies there were no guests for those bookings. These rows were removed from the dataset. After dropping these invalid rows, the dataset contained 119,388 records with 31 columns.

F. Handling Inconsistent Data:

The dataset included rows where both adults, children, and babies were zero, representing invalid entries. Such records were removed using the filter.After removing these invalid rows, the dataset was updated.

G. Creating New Features:

- A total_stay feature was created by summing the number of stays_in_weekend_nights and stays_in_week_nights to capture the total duration of a guest's stay
- A has_children feature was introduced to identify bookings that involved children or babies. If either children or babies was greater than zero, this feature was set to 1, otherwise 0.
- A long_wait feature was created to indicate whether a booking involved a waiting period. If days_in_waiting_list was greater than zero, this feature was set to 1:

H. Log Transformation

By applying the logarithmic transformation to both ADR and lead time, we enhance the model's ability to predict cancellations more effectively.

- ADR directly impacts the hotel's revenue management and pricing strategy. By transforming ADR, we ensure that extreme high-end bookings don't dominate the model's focus, allowing more accurate predictions across a wider range of bookings.
- Lead time is an important factor because customers who book far in advance may be more likely to cancel. Log transforming this feature enables the model to learn from subtle patterns in the data without being skewed by a few very long lead times.

I. Checking for Data Imbalance:

To examine whether there is an imbalance in the dataset between bookings that were canceled and those that were not, the percentage distribution of the is_canceled column was calculated. A bar plot was generated to visualize the distribution (Fig. 1).

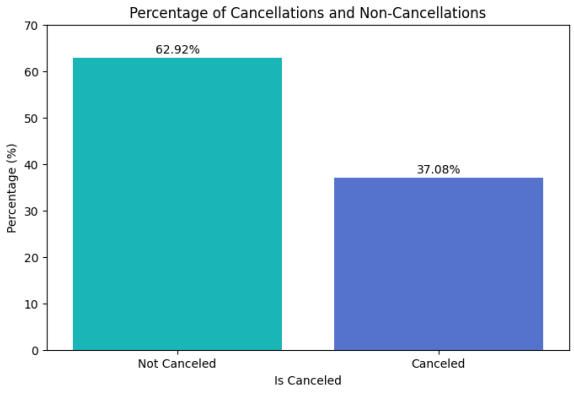


Fig. 1. Percentage of Cancellations and Non-Cancellations

IV. Exploratory Data Analysis

A. Number of Bookings for Each Hotel Type

Observations:

- The bar chart shows that City Hotels have significantly more bookings (79,330) than Resort Hotels (40,060). This suggests that City Hotels attract a larger audience.
- This disparity indicates that City Hotels dominate in terms of customer demand, as reflected by the larger proportion of bookings.

Interpretations:

- City Hotels likely attract a more frequent traveler demographic, such as business or short-term visitors, contributing to their consistently higher booking rates.
- Resort Hotels, with fewer overall bookings, may rely on seasonal peaks or longer vacation stays, potentially pointing to a more specialized or time-sensitive

customer base. This suggests opportunities for targeted marketing during off-peak periods to balance the booking trends (Fig. 2).

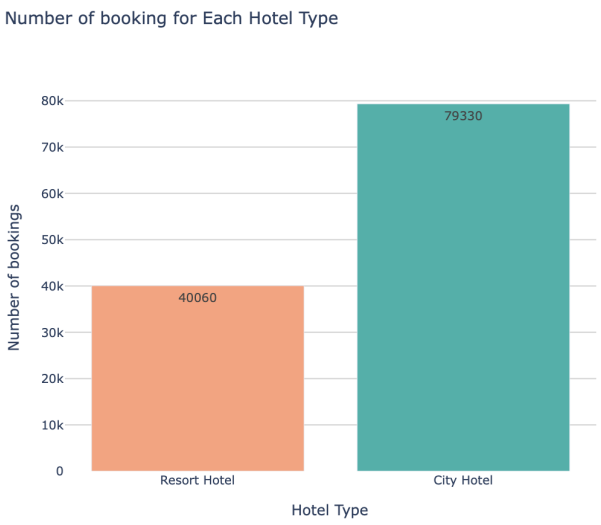


Fig. 2. Number Of Bookings vs Hotel Type

B. Number of Bookings by Country:

Observations:

- The map highlights that most bookings come from European countries, with Portugal being the largest contributor, followed by other nearby regions. Non-European countries show much lower booking volumes, with minimal representation from Africa and South America.

Interpretations:

- The high number of bookings from Portugal suggests that the hotel primarily attracts domestic travelers or those from nearby European countries. This trend could be due to proximity, cultural ties, or regional tourism habits.
- The low booking counts in other continents, particularly in Africa and South America, suggest that the hotel's customer base may have less global reach and is more regionally concentrated (Fig. 3).

C. Number of Arrivals Over Time

The time series graph below illustrates the trend of hotel arrivals from July 2015 to August 2017. This chart shows the number of arrivals for both City Hotels and Resort Hotels, revealing clear seasonal fluctuations.

Observations:

- The City Hotel shows significant fluctuations in the number of bookings/arrivals over time, with peaks in mid-2016 and mid-2017, surpassing 4000 bookings.

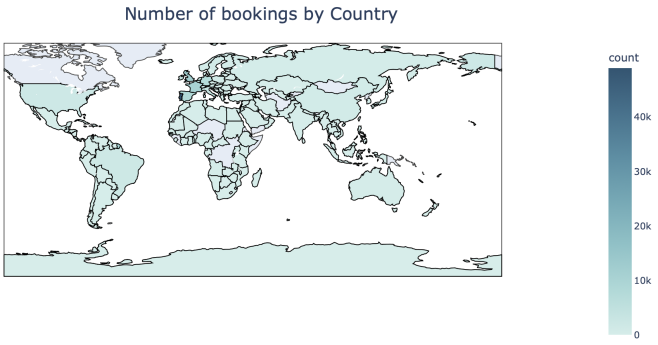


Fig. 3. Bookings by Country

- The Resort Hotel exhibited relatively stable booking volumes, with arrivals consistently ranging between 1000 and 2000 throughout the observed period.

Interpretation:

- The higher variability in City Hotel bookings may suggests that its demand was more influenced by short-term factors, such as event-driven or business-related stays, contributing to more dynamic booking trends.
- In contrast, the steadier bookings for Resort Hotels indicate a more predictable customer base, likely associated with planned vacations or extended stays, leading to a more consistent number of arrivals (Fig. 4).

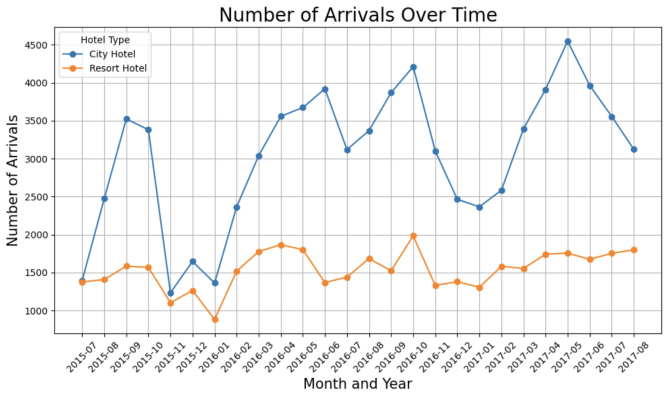


Fig. 4. No. of Arrivals over time

D. Number of Guests per Booking

Observations:

- The histogram shows that the vast majority of bookings are for 1 to 2 guests, with the distribution heavily skewed toward lower guest counts.
- The total guest count rarely exceeds 10, and bookings with more than 10 guests are outliers.

Interpretation:

- Most hotel bookings are for small groups or individuals, which is typical for both business and leisure travelers
- The infrequent occurrence of large groups indicates that hotels, on average, cater to smaller groups of guests. The high frequency of small bookings suggests that individual travelers or couples make up a significant portion of the hotel's clientele. This pattern could affect resource planning, where hotels may focus more on accommodating single or double bookings rather than larger group bookings (Fig. 5).

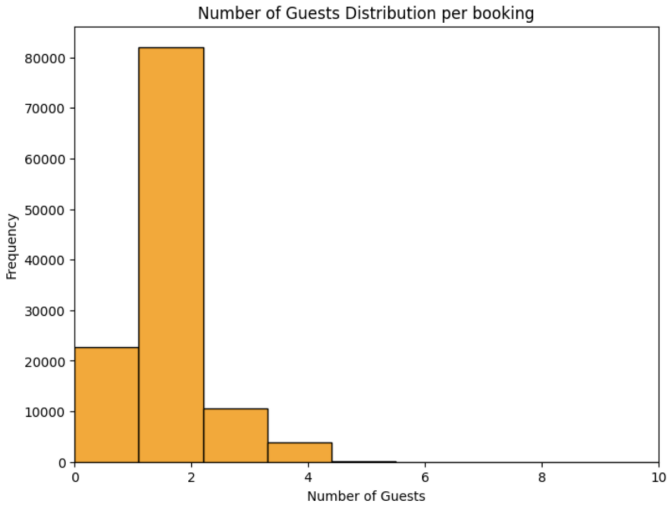


Fig. 5. Number of guests distribution

E. Number of Booking by Lead time Category

Observations:

- The bar chart shows the distribution of bookings across different lead time categories.
- The Very Short lead time category (0-30 days) has the highest number of bookings, followed by the Short and Medium categories.

Interpretation:

- Most bookings are made within 0-90 days before the stay, suggesting that guests typically plan their trips on shorter notice.
- Hotels may not need to heavily focus on very long-term booking strategies since the demand for such bookings is relatively low (Fig. 6).

F. Correlation for Numerical Features

Observations:

- Total Guests is highly correlated with adults and children but has little correlation with babies.
- Stays in Weekend Nights and Stays in Week Nights show a strong positive correlation, meaning that bookings often include both weekend and weekday nights.

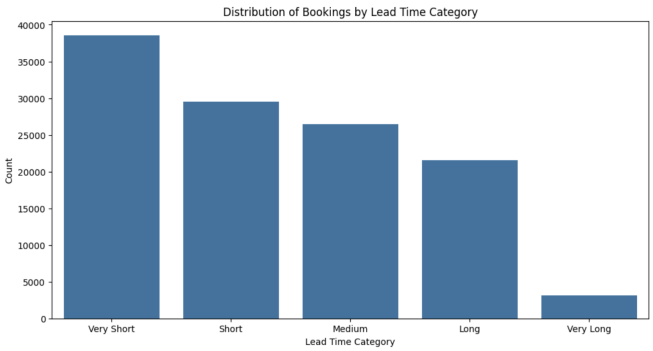


Fig. 6. Bookings by Lead Time

- Lead Time shows minimal correlation with the number of guests, indicating that how far in advance a booking is made has little impact on the total guest count.
- ADR (Average Daily Rate) has some correlation with total guests, but it's relatively weak, suggesting that larger groups may sometimes pay more, but it's not a strong determining factor.

Interpretation:

- Number of Guests is driven largely by the number of adults and children rather than babies, which suggests that most hotel rooms are designed for adults or older children.
- The strong correlation between weekend and weekday stays suggests that bookings often span both periods, possibly indicating longer stays over multiple days (Fig. 7).

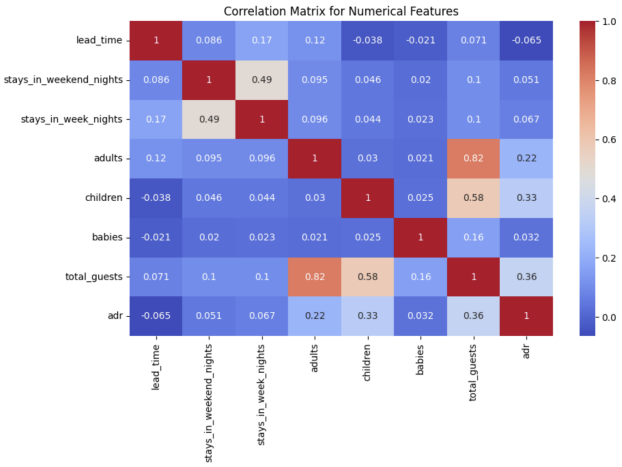


Fig. 7. Correlation for Numerical Features

G. No. of Guests - Meal type

Observations:

- The vast majority of guests selected the BB (Bed & Breakfast) meal type, with significantly fewer choosing

the other options such as FB (Full Board), HB (Half Board), and SC (Self Catering).
- The number of guests does not appear to have a significant impact on the meal type chosen, as the distribution of total guests across meal types remains consistent.

Interpretation:

- The BB (Bed & Breakfast) meal option is the most popular choice among guests, regardless of the number of people in the booking. This suggests that guests prioritize breakfast as part of their stay, but do not necessarily require additional meal services like full or half board.
- The consistency in guest numbers across meal types indicates that meal preferences are likely driven by factors other than group size, such as personal preferences or package offerings by the hotel (Fig. 8).

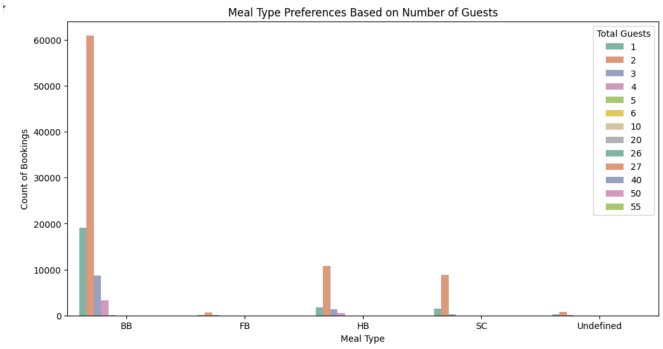


Fig. 8. No. of Bookings by Meal Type

H. Cancellations by Room Mismatch

Observations:

- The chart shows that when there is no room mismatch (False), a significant number of bookings were not canceled, but a substantial portion still resulted in cancellations.
- When there is a room mismatch (True), fewer bookings were made, and most of these were not canceled, with very few resulting in cancellations.

Interpretation:

- The data suggests that room mismatches do not significantly increase cancellation rates. Even when rooms are assigned correctly, cancellations occur frequently, indicating that factors other than room mismatch, such as booking preferences or external circumstances, may drive the decision to cancel (Fig. 9).

I. Cancellation Trend by Lead Time

Observations:

- The box-plot shows that bookings with longer lead times (number of days between booking and check-in)

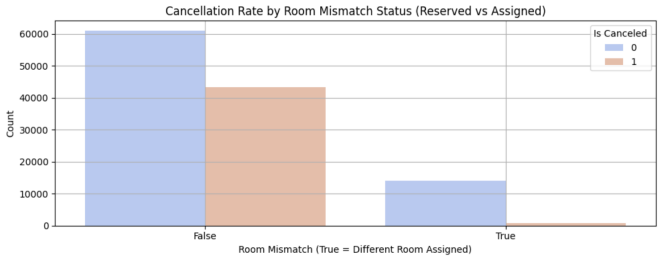


Fig. 9. Cancellations by Room Mismatch

are more likely to result in cancellations, as indicated by the higher median lead time for canceled bookings.
- Canceled bookings have a broader spread of lead times, including many outliers, while non-canceled bookings tend to have shorter and more consistent lead times.

Interpretation:

- Longer lead times appear to be associated with a higher likelihood of cancellations, suggesting that customers who book further in advance may be more prone to changing or canceling their reservations.
- Non-canceled bookings are generally made closer to the check-in date, indicating that shorter-term bookings may be more reliable and less prone to cancellation (Fig. 10).

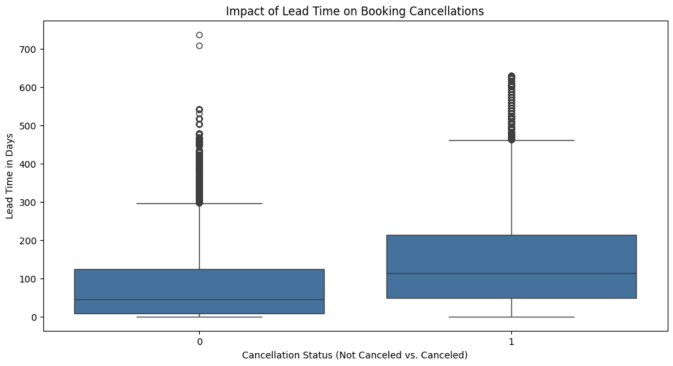


Fig. 10. Cancellations - Lead Time

J. Cancellation Trend by Market Segment

Observations:

- The Online Travel Agents (Online TA) market segment shows the highest number of bookings, but it also has a significant cancellation rate, with a large portion of bookings getting canceled.
- Direct bookings and groups also show notable cancellation rates, but their overall booking volumes are lower compared to Online TA.
- Corporate and Offline TAs have relatively fewer cancellations, indicating more stable booking patterns in these segments.

Interpretation:

- The Online Travel Agents (Online TA) segment experiences both the highest number of bookings and the highest cancellation rate, highlighting this market segment as a key area where cancellations are most prevalent (Fig. 11).

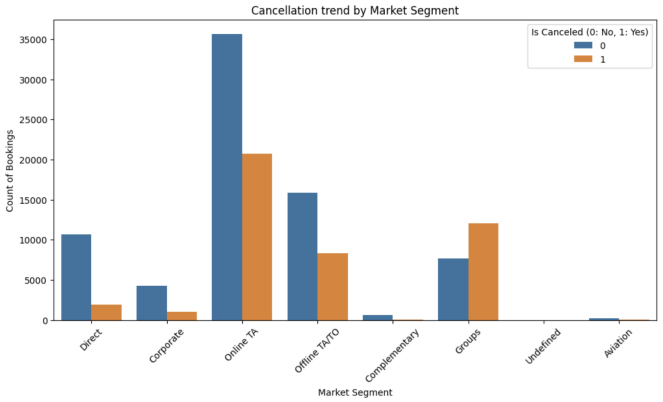


Fig. 11. Cancellations - Market Segment

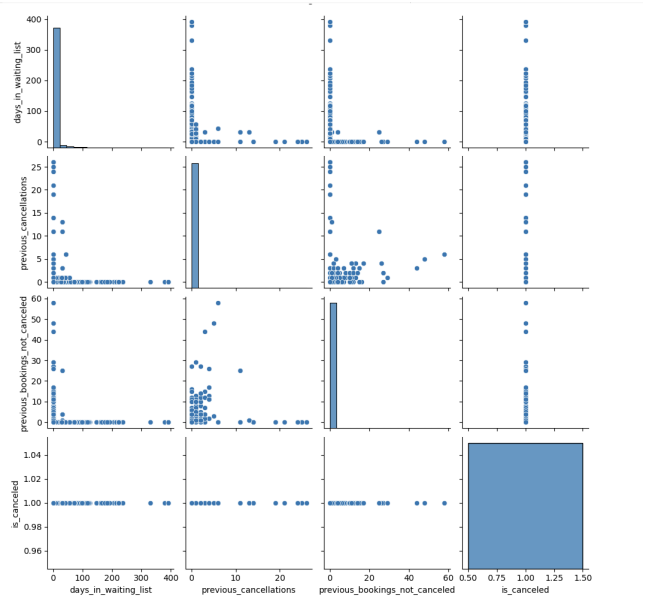


Fig. 12. Cancellations - Pair Plot

K. Cancellation Trend by Market Segment

Observations:

- Guests with a history of previous cancellations or non-canceled bookings are more likely to cancel future bookings, as shown by the clustering of data in these variables.
- Waiting list duration also appears to be associated with cancellations, as bookings that stay longer on the waiting list are more prone to being canceled.

Interpretation:

- Patterns in guest behavior, such as frequent cancellations or modifications, highlight the need for strategies that address these repeat behaviors to reduce future cancellations.
- Managing waiting list duration's and improving communication during this time could help minimize cancellations, especially for guests with longer wait times (Fig. 12).

V. Models/Algorithms

To prevent data leakage and build a robust model for predicting cancellations, the following features should be removed:

- reservation_status directly reveals whether a booking was canceled, as 'Canceled' implies is_canceled = 1 and 'Check-Out' implies is_canceled = 0. Including this feature would lead to data leakage.
- reservation_status_date: indicates when the reservation status was last updated. It directly correlates with whether the booking was canceled (cancellation occurs before the arrival date), leading to data leakage.
- arrival_date_year: represents specific years, which may bias the model toward past data and reduce its generalization to future bookings.
- assigned_room_type feature indicates the type of room given to a guest after the booking is confirmed. Since it is determined post-booking, it cannot be used to predict cancellations. Thus, it is considered irrelevant and should be excluded.

High Cardinality Categorical Features:

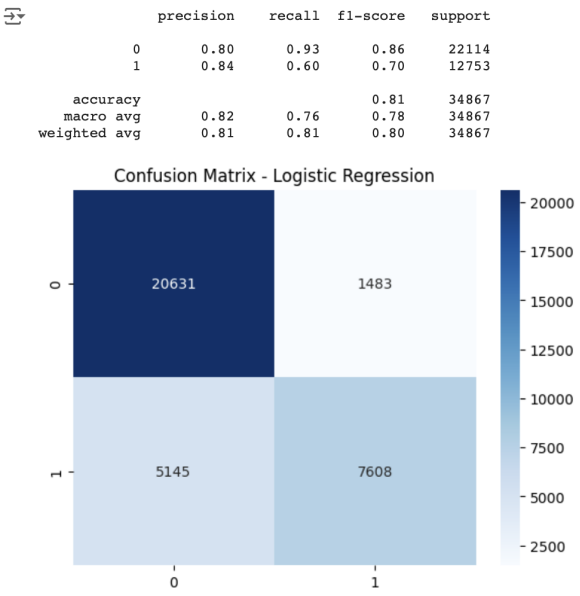
- country: The large number of country categories may not provide meaningful insights. Consider grouping countries into regions or keeping only the top frequent countries if needed.
- agent: Due to many unique categories and infrequent appearances, it may cause overfitting. Instead of using the original feature, you could create a new feature representing the number of bookings per agent, but it may not be very meaningful.

Therefore, we are omitting the above features along with some other features which are created for EDA or the features that reveal or are highly correlated with the target variable, is_canceled, to avoid data leakage and improve model generalization.

A. Logistic Regression

- Logistic Regression was chosen as the baseline model for predicting booking cancellations due to its suitability for binary classification tasks. It offers valuable insights into feature importance through model coefficients, enabling an understanding of the relationship between variables such as lead time and previous cancellations with the likelihood of a booking being cancelled.

- This algorithm works well when the relationship between predictors and the outcome is primarily linear, and the dataset contains a mix of categorical and numerical features. The model is easy to interpret and trains quickly, even on large datasets.
- Effectiveness: The Logistic Regression model achieved 81% accuracy and an F1-score of 0.70 for predicting cancellations. While it was a simple and interpretable choice, it struggled with recall for cancellations (class 1), indicating challenges in detecting rare events. Despite its limitations in handling non-linear relationships, its interpretability made it useful for identifying key predictors of cancellations. (Fig. 13).

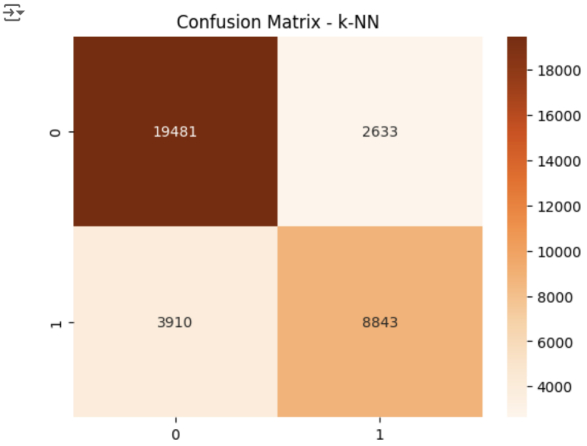


	precision	recall	f1-score	support
0	0.80	0.93	0.86	22114
1	0.84	0.60	0.70	12753
accuracy			0.81	34867
macro avg	0.82	0.76	0.78	34867
weighted avg	0.81	0.81	0.80	34867

Fig. 13. Logistic Regression - Confusion matrix

B. K-NN

- K-Nearest Neighbors (KNN) is selected for predicting booking cancellations due to its non-parametric, instance-based learning approach that relies on feature similarity. The model operates on the principle that bookings with similar characteristics, such as lead time, customer type, and market segment, are likely to yield similar outcomes regarding cancellations.
- This algorithm is particularly effective in scenarios where data points with similar features exhibit similar outcomes, allowing it to capture complex, non-linear relationships that simpler models might overlook. To address KNN's sensitivity to feature scales, numerical features were standardized during preprocessing.
- Effectiveness: The KNN model achieved 81% accuracy and an F1-score of 0.73 for predicting cancellations. Although it performed reasonably well, its computational cost due to the dataset size and lower recall for cancellations (class 1) limited its overall applicability. Despite being a simple and interpretable algorithm, these challenges hindered its effectiveness in detecting rare events compared to other models. (Fig. 14).



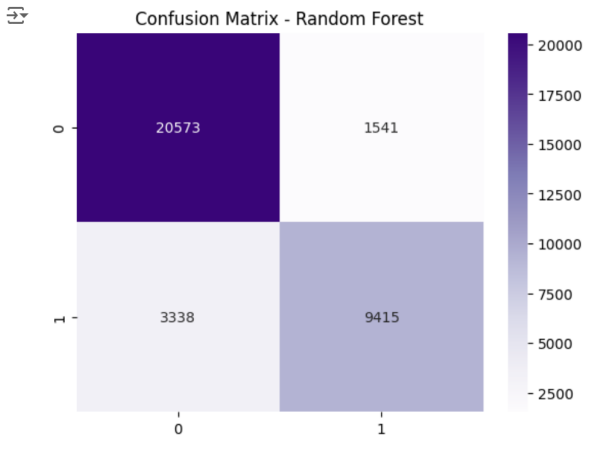
Classification Report :

	precision	recall	f1-score	support
0	0.83	0.88	0.86	22114
1	0.77	0.69	0.73	12753
accuracy			0.81	34867
macro avg	0.80	0.79	0.79	34867
weighted avg	0.81	0.81	0.81	34867

Fig. 14. KNN - Confusion matrix

C. Random Forest

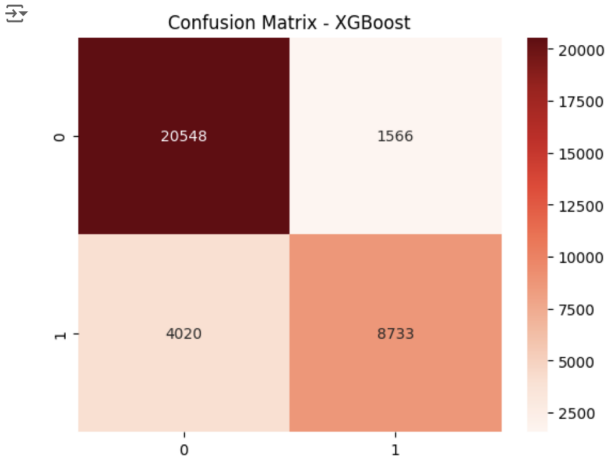
- Random Forest is chosen as the model for predicting booking cancellations due to its robustness as an ensemble method that combines predictions from multiple decision trees. This approach effectively handles both categorical and numerical variables, making it less prone to overfitting compared to individual decision trees, as it averages the predictions of multiple trees.
- It excels in managing complex interactions in large datasets, enhancing predictive accuracy and reducing variance. To optimize performance, we tuned parameters such as n_estimators (number of trees) and max_depth (tree depth) to control overfitting and computational costs.
- Effectiveness: The Random Forest model achieved 86% accuracy and an F1-score of 0.79 for predicting cancellations, demonstrating its strong performance. It effectively captured non-linear relationships and balanced precision, recall, and F1-score, making it the best choice for this task. Its ability to provide feature importance rankings added significant value, allowing us to identify the key factors influencing booking cancellations and enhancing our understanding of the problem. (Fig. 15).



Classification Report :

	precision	recall	f1-score	support
0	0.86	0.93	0.89	22114
1	0.86	0.74	0.79	12753
accuracy			0.86	34867
macro avg	0.86	0.83	0.84	34867
weighted avg	0.86	0.86	0.86	34867

Fig. 15. Random Forest - Confusion matrix



Classification Report :

	precision	recall	f1-score	support
0	0.84	0.93	0.88	22114
1	0.85	0.68	0.76	12753
accuracy			0.84	34867
macro avg	0.84	0.81	0.82	34867
weighted avg	0.84	0.84	0.84	34867

Fig. 16. XGBoost - Confusion matrix

D. XGBoost

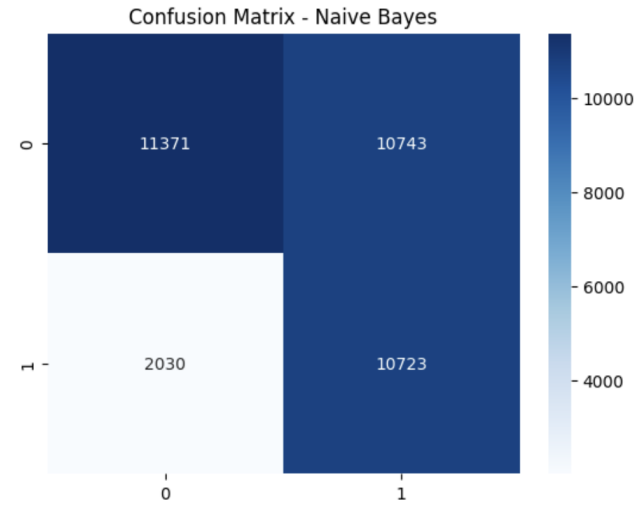
- XGBoost (XGBClassifier) is chosen for predicting booking cancellations due to its effectiveness as an ensemble method based on gradient boosting. It excels with structured data, efficiently handling both numerical and categorical features while performing well with imbalanced datasets, where cancellations are less frequent.
- The model utilizes multiple weak learners (decision trees) to minimize residual errors, effectively capturing complex interactions between features like lead time, previous cancellations, and customer type. To optimize performance, several hyperparameters, including n_estimators, learning_rate, and max_depth, were tuned using grid search.
- Effectiveness: The XGBoost model achieved 84% accuracy and an F1-score of 0.76 for predicting cancellations. While it performed well, it slightly lagged in recall for class 1 (cancellations), affecting its overall effectiveness compared to Random Forest. Despite this, XGBoost's ability to handle complex datasets and provide feature importance scores is invaluable for understanding the key factors influencing booking cancellations. (Fig. 16).

E. Naive Bayes

- Naive Bayes (GaussianNB) is selected for predicting booking cancellations due to its basis in Bayes theorem and the assumption that features are normally distributed. Despite its simplicity, it can be effective for certain classification problems, particularly when the assumptions of normality hold for some features.
- This model works well for this problem because it provides a probabilistic approach, offering insights into how strongly each feature influences the likelihood of cancellation. It is particularly useful when a simple, interpretable model is needed for moderately-sized datasets, despite containing both categorical and continuous variables.
- Effectiveness: The GaussianNB model achieved 63% accuracy and an F1-score of 0.63 for predicting cancellations. However, its reliance on the assumption of feature independence limited its performance, especially in complex datasets like this one, making it unsuitable for accurately predicting cancellations. Despite being computationally efficient and fast to train, its performance was the lowest among all models, highlighting its limitations in this context. (Fig. 17).

F. Decision Tree

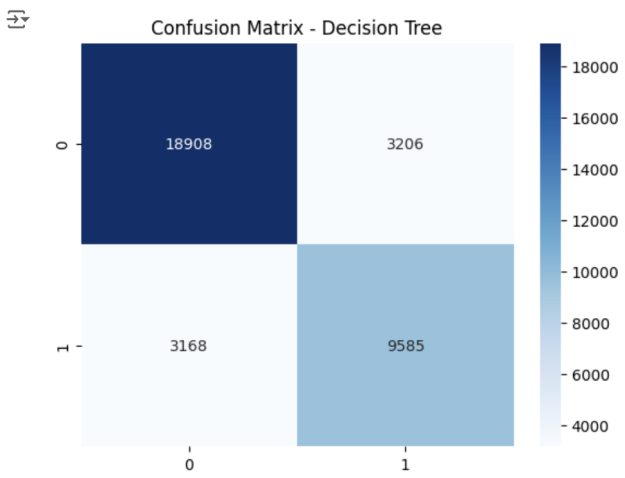
- The Decision Tree Classifier was selected for predicting booking cancellations due to its intuitive and interpretable nature, providing a clear pathway from features to the target variable. It effectively handles both categorical and numerical data, making it suitable for the mixed dataset of hotel bookings. The model captures complex decision rules related to factors like lead time, market segment, and customer type.
- This algorithm is beneficial for this problem because it can model non-linear relationships without extensive



Classification Report :

	precision	recall	f1-score	support
0	0.85	0.51	0.64	22114
1	0.50	0.84	0.63	12753
accuracy			0.63	34867
macro avg	0.67	0.68	0.63	34867
weighted avg	0.72	0.63	0.64	34867

Fig. 17. naive Bayes - Confusion matrix



Classification Report :

	precision	recall	f1-score	support
0	0.86	0.86	0.86	22114
1	0.75	0.75	0.75	12753
accuracy			0.82	34867
macro avg	0.80	0.80	0.80	34867
weighted avg	0.82	0.82	0.82	34867

Fig. 18. Decision Tree - Confusion matrix

preprocessing, splitting the data based on feature importance to identify key contributors to cancellations.

- Effectiveness: The Decision Tree model achieved 82% accuracy and an F1-score of 0.75 for predicting cancellations. However, it was prone to overfitting, which limited its generalizability. To mitigate this, parameters like max_depth and min_samples_split were tuned to control tree complexity. Despite these challenges, the model offered valuable insights into decision-making patterns. (Fig. 18).

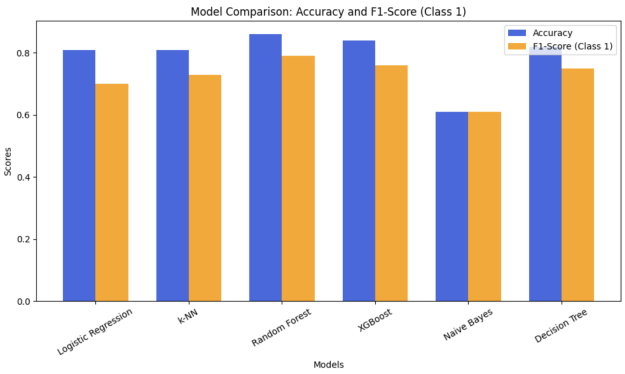


Fig. 19. Model Comparision

VI. Conclusion

Best Model: Random Forest emerged as the best model with the highest accuracy (86%) and f1-score (0.79) for predicting cancellations. It was effective at capturing the complex patterns in the data, making it the most suitable model for our problem statement.

This analysis provided valuable insights into hotel booking cancellations:

- Lead time, special requests, and previous cancellations were identified as major contributors to booking cancellations.
- Predicting cancellations: Random Forest was the most effective at accurately predicting cancellations, allowing hotels to anticipate cancellations and adjust pricing, marketing, and operations accordingly.

The analysis of hotel booking data provides valuable insights into customer preferences and booking behaviors. City Hotels are the preferred choice, receiving 79,330 bookings compared to 40,060 for Resort Hotels. This trend indicates a focus on frequent travelers, with a significant portion of bookings originating from European countries, particularly Portugal, highlighting a strong domestic market.

City Hotel bookings exhibit noticeable seasonal fluctuations, whereas Resort Hotels maintain relatively stable occupancy levels throughout the year. Most reservations are for 1 to 2 guests, suggesting a demand for smaller, short-term accommodations, with most bookings made within 0-90 days prior to check-in.

Further analysis reveals that guest numbers are primarily driven by adults and children, with the Bed & Breakfast meal option being the most popular. Cancellations are more common with longer lead times, although room mismatches do not significantly contribute to cancellations. The Online Travel Agents (OTA) segment shows the highest booking

and cancellation rates, indicating a potential area for targeted improvements.

These findings will enable hotel management to refine marketing strategies, optimize booking processes, and implement measures to improve customer satisfaction while reducing cancellation rates.

Project Contribution

All project members contributed equally to the work presented in this paper.