Contents

Hotel Booking Cancellation 2
Problem Statement: 2
Objectives: 2
Dataset Columns Description 2
Summary Statistics of Numerical Features: 4
Summary Statistics for Categorical Features: 5
Data Preprocessing: 7
Dropped features: 7
Feature transformation: 7
Handling missing values: 7
Handling the noisy data: 7
Data Encoding 7
Balance of Target Feature group representation: 8
Splitting the dataset: 9
Model Building: 9
Decision Tree Model: 9
Random forest model 12
XGBoost Model: 15
Model Comparison: 18
Recommendation: 19

Hotel Booking Cancellation

Problem Statement:

This project aims to develop a predictive model to determine if a hotel booking will be cancelled. The dataset contains numerous features related to bookings collected from a large set of users. The challenges include data preprocessing tasks like handling missing values, feature selection and addressing data noise. We will train multiple models, evaluate their performance using appropriate metrics, and interpret the models by identifying the key features influencing hotel booking cancellations.

Objectives:

- Explore the dataset
- Preprocessing the dataset
- Model building
- Evaluation and Comparison
- Recommendation

Dataset Columns Description

Index	Variable	Description
1	hotel	Type of hotel (Resort Hotel, City Hotel)
2	is_canceled	Reservation cancellation status (0 = not canceled, 1 = canceled)
3	lead_time	Number of days between booking and arrival
4	arrival_date_year	Year of arrival
5	arrival_date_month	Month of arrival
6	arrival_date_week_number	Week number of the year for arrival
7	arrival_date_day_of_month	Day of the month of arrival
8	stays_in_weekend_nights	Number of weekend nights (Saturday and Sunday) the guest stayed or booked
9	stays_in_week_nights	Number of week nights the guest stayed or booked
10	adults	Number of adults
11	children	Number of children
12	babies	Number of babies
13	meal	Type of meal booked (BB, FB, HB, SC, Undefined)
14	country	Country of origin of the guest
15	market_segment	Market segment designation
16	distribution_channel	Booking distribution channel
17	is_repeated_guest	If the guest is a repeat customer (0 = not repeated, 1 = repeated)
18	previous_cancellations	Number of previous bookings that were canceled by the customer

19	previous_bookings_not_canceled	Number of previous bookings that were not canceled by the customer
20	reserved_room_type	Type of reserved room
21	assigned_room_type	Type of assigned room
22	booking_changes	Number of changes made to the booking
23	deposit_type	Type of deposit made (No Deposit, Refundable, Non Refund)
24	agent	ID of the travel agent responsible for the booking
25	company	ID of the company responsible for the booking
26	days_in_waiting_list	Number of days the booking was in the waiting list
27	customer_type	Type of customer (Transient, Contract, Transient-Party, Group)
28	adr	Average Daily Rate
29	required_car_parking_spaces	Number of car parking spaces required
30	total_of_special_requests	Number of special requests made
31	reservation_status	Last reservation status (Check-Out, Canceled, No-Show)
32	reservation_status_date	Date of the last reservation status
33	name	Guest's name
34	email	Guest's email address
35	phone-number	Guest's phone number
36	credit_card	Last four digits of the guest's credit card

Table 1 : Feature Descriptions

- Number of Entries: The dataset comprises 119,390 entries.

- Columns: It features 36 columns, each representing different aspects of hotel bookings.

Data Types:

- The majority of the columns, 16 in total, are of the object data type, typically used for strings or categorical data.
- Another 16 columns are of the int64 data type, indicating integer values.
  The remaining 4 columns are of the float64 data type, which usually represent decimal values.

Missing Values:

- The 'children' column has 4 missing entries.
- The 'country' column has 488 missing entries.
- The 'agent' column has 16,340 missing entries.
- The 'company' column has a substantial number of missing entries, totalling 112,593. These missing values might need to be addressed, depending on the context of the analysis or the model we intend to build.

Summary Statistics of Numerical Features:

	count	mean	std	min	25%	50%	75%	max
lead_time	119390.0	104.011416	106.863097	0.00	18.00	69.000	160.0	737.0
arrival_date_week_number	119390.0	27.165173	13.605138	1.00	16.00	28.000	38.0	53.0
arrival_date_day_of_month	119390.0	15.798241	8.780829	1.00	8.00	16.000	23.0	31.0
stays_in_weekend_nights	119390.0	0.927599	0.998613	0.00	0.00	1.000	2.0	19.0
stays_in_week_nights	119390.0	2.500302	1.908286	0.00	1.00	2.000	3.0	50.0
adults	119390.0	1.856403	0.579261	0.00	2.00	2.000	2.0	55.0
children	119386.0	0.103890	0.398561	0.00	0.00	0.000	0.0	10.0
babies	119390.0	0.007949	0.097436	0.00	0.00	0.000	0.0	10.0
previous_cancellations	119390.0	0.087118	0.844336	0.00	0.00	0.000	0.0	26.0
previous_bookings_not_canceled	119390.0	0.137097	1.497437	0.00	0.00	0.000	0.0	72.0
booking_changes	119390.0	0.221124	0.652306	0.00	0.00	0.000	0.0	21.0
days_in_waiting_list	119390.0	2.321149	17.594721	0.00	0.00	0.000	0.0	391.0
adr	119390.0	101.831122	50.535790	-6.38	69.29	94.575	126.0	5400.0
required_car_parking_spaces	119390.0	0.062518	0.245291	0.00	0.00	0.000	0.0	8.0
total_of_special_requests	119390.0	0.571363	0.792798	0.00	0.00	0.000	1.0	5.0

Table 2: Numerical Features Stats

Inference:

- Lead Time: The average duration between booking and arrival is around 104 days, with a range from 0 to 737 days.
- Arrival Date Week Number: The average week number of the year for arrival is approximately 27.17, ranging from 1 to 53.
- Arrival Date Day of Month: The average day of the month for arrival is about 15.80, with a range from 1 to 31.
- Stays in Weekend Nights: Guests typically stay for around 0.93 weekend nights on average, with the maximum stay being 19 weekend nights.
- Stays in Week Nights: Guests generally stay for about 2.5-week nights on average, with stays ranging from 0 to 50-week nights.
- Adults: The average number of adults per booking is around 1.86, with the number ranging from 0 to 55 adults.
- Children: On average, there are about 0.1 children per booking, with some bookings having up to 10 children.
- Babies: The average number of babies per booking is very low, approximately 0.008, with some bookings including up to 10 babies.
- Previous Cancellations: Guests have cancelled about 0.09 times on average in the past, with some guests having up to 26 previous cancellations.

- Previous Bookings Not Cancelled: On average, guests have not cancelled about 0.14 bookings in the past, with some guests having up to 72 previous bookings that were not cancelled.
- Booking Changes: Bookings are changed on average about 0.22 times, with some bookings being changed up to 21 times.
- Days in Waiting List: The average duration a booking spends on the waiting list is around 2.32 days, with some bookings waiting up to 391 days.
- ADR: The Average Daily Rate is approximately 101.83, with rates ranging from -6.38 (possibly indicating errors or special cases) to 5400.
- Required Car Parking Spaces: On average, guests require about 0.06 parking spaces, with some bookings needing up to 8 spaces.
- Total of Special Requests: Guests make an average of about 0.57 special requests, with some making up to 5 requests.

Summary Statistics for Categorical Features:

	count	unique	top	freq
hotel	119390	2	City Hotel	79330
is_canceled	119390	2	0	75166
arrival_date_year	119390	3	2016	56707
arrival_date_month	119390	12	August	13877
meal	119390	5	BB	92310
country	119390	178	PRT	48590
market_segment	119390	8	Online TA	56477
distribution_channel	119390	5	TA/TO	97870
is_repeated_guest	119390	2	0	115580
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
agent	119390	334	9.0	31961
company	119390	353	nan	112593
customer_type	119390	4	Transient	89613
reservation_status	119390	3	Check-Out	75166
reservation_status_date	119390	926	2015-10-21	1461
name	119390	81503	Michael Johnson	48
email	119390	115889	Michael.C@gmail.com	6
phone-number	119390	119390	669-792-1661	1
credit_card	119390	9000	**********4923	28

Table 3: Categorical Features Stats

Inference:

- Hotel: The dataset includes two types of hotels, with "City Hotel" being the most common, appearing 79,330 times out of 119,390 entries.
- Is Cancelled: This column has two unique values (0 for not cancelled and 1 for cancelled). The value "0" (not cancelled) is the most frequent, occurring 75,166 times.
- Arrival Date Year: There are three distinct years in the dataset, with 2016 being the most common year of arrival, recorded 56,707 times.
- Arrival Date Month: All 12 months are represented, with August being the most frequent month of arrival, observed 13,877 times.
- Meal: Five unique meal types are booked, with "BB" being the most common, appearing 92,310 times.
- Country: There are 178 different countries of origin, with "PRT" (Portugal) being the most frequent, appearing 48,590 times.
- Market Segment: The dataset includes eight different market segments, with "Online TA" being the most common, appearing 56,477 times.
- Distribution Channel: There are five unique booking distribution channels, with "TA/TO" being the most frequent, appearing 97,870 times.
- Is Repeated Guest: This column has two unique values (0 for not repeated and 1 for repeated), with "0" (not repeated) being the most common, appearing in the majority of entries.
- Reserved Room Type and Assigned Room Type: Various room types are present, with some types being more frequent than others.
- Deposit Type: Three types of deposits are recorded, with "No Deposit" being the most common, appearing 104,641 times.
- Agent: There are 334 unique agents, with agent '9.0' being the most frequent, appearing 31,961 times.
- Company: There are 353 unique companies, but 'nan' (indicating missing values) is the most frequent, occurring 112,593 times, suggesting a high percentage of missing values in the 'company' column.
- Customer Type: Four unique customer types are present, with "Transient" being the most common, appearing 89,613 times.
- Reservation Status: There are three unique reservation statuses, with "Check-Out" being the most common, appearing 75,166 times.
- Reservation Status Date: There are 926 unique values, with '2015-10-21' being the most frequent, appearing 1,461 times.
- Name: There are 81,503 unique names, with 'Michael Johnson' being the most frequent, appearing 48 times.
- Email: There are 115,889 unique email addresses, with 'Michael.C@gmail.com' being the most frequent, appearing 6 times.
- Phone Number: There are 119,390 unique phone numbers, indicating nearly every guest has a unique phone number.

- Credit Card: There are 9,000 unique credit card numbers, with '**4923' being the most frequent, appearing 28 times.

Data Preprocessing:

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Dropped features:

- Features such as reservation_status, reservation_status_date, and assigned_room_type are directly linked to the target variable, is_canceled. Including these features in the model would cause data leakage. Therefore, it's crucial to exclude these features to develop a predictive model that can accurately forecast cancellations rather than simply label them
- We also further drop the features which we consider as high cardinality features, which are 'name', 'country', 'agent', 'company', 'email', 'phone-number', 'credit_card'

Feature transformation:

- From the feature ‘name’, we create a new feature which represents the number of bookings made by each guest.

Handling missing values:

After dropping the above discussed features, we verify the dataset for null values and we discover that children column has 4 missing values. We impute the missing values by the mode of the feature.

Handling the noisy data:

While performing the initial statistically summary, we notice that feature ‘adr’ and ‘adults’ had some out of pattern data,

- Adr: There is 1 record with negative value, which doesn’t make sense
- Adults: There are 403 bookings where the adult column as 0 entries, this could be an entry error.

We replace the anomaly in ‘adr’ with mean and in case of ‘adults’ feature, we decide to remove those 403 records.

Data Encoding

- We encode the nominal features without any intrinsic order using one-hot encoding.
- We do not perform any encoding for numerical features
- We do label encoding for categorical features which has meaningful order.

Balance of Target Feature group representation:

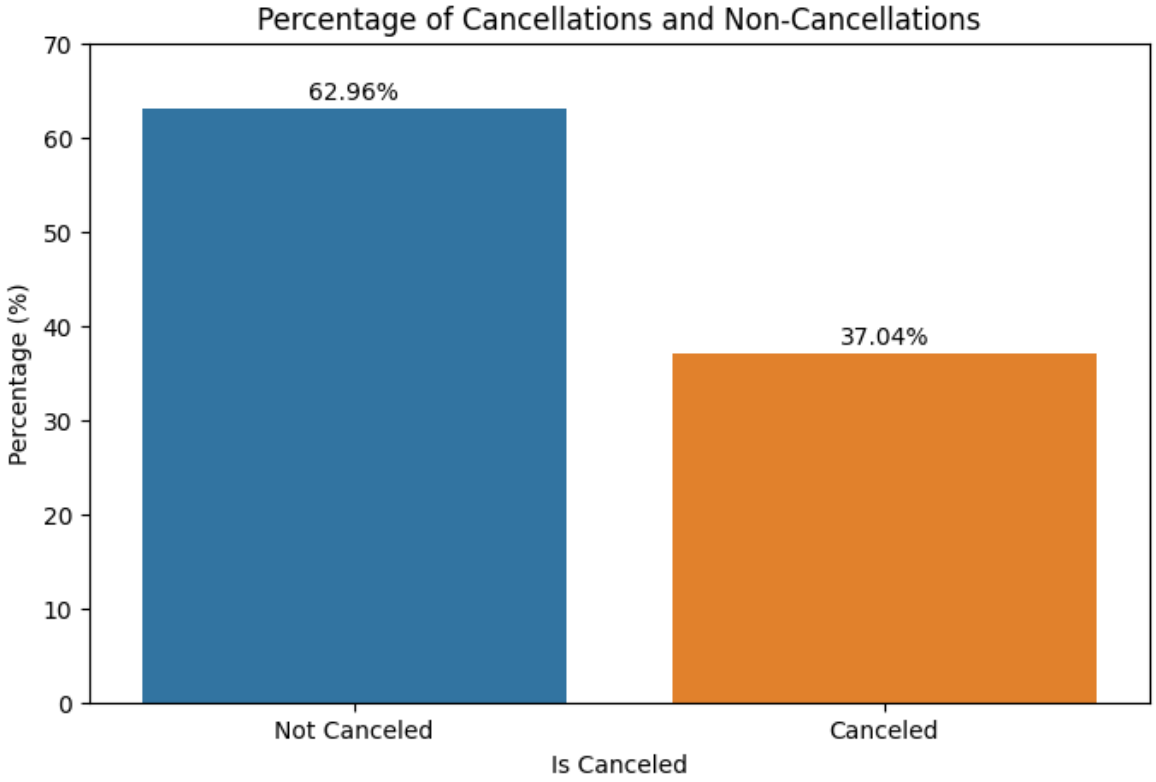


Figure 1: Target Group Balance

Inference:

The bar graph depicts the balance between customers who have 'Non cancelled' and 'Cancelled'. There are about 37% of customers who have cancelled the booking and 63% of customers who have not cancelled the bookings. We notice there is an imbalance in target feature, however the imbalance is not as bad as a 90:10 ratio, since we have a decent 60:40 split approximately.

Splitting the dataset:

We split the dataset into train and test groups with a ratio of 80:20 and we stratify this process to maintain the balance in proportion between two groups.

Model Building:

Decision Tree Model:

A decision tree classifier is a supervised machine learning algorithm used for classification tasks. It splits data into subsets based on feature values, forming a tree structure where each node represents a feature, branches represent decision rules, and leaf nodes represent class labels. It's easy to interpret and can handle both numerical and categorical data, but it can overfit complex datasets.

Confusion Matrix:

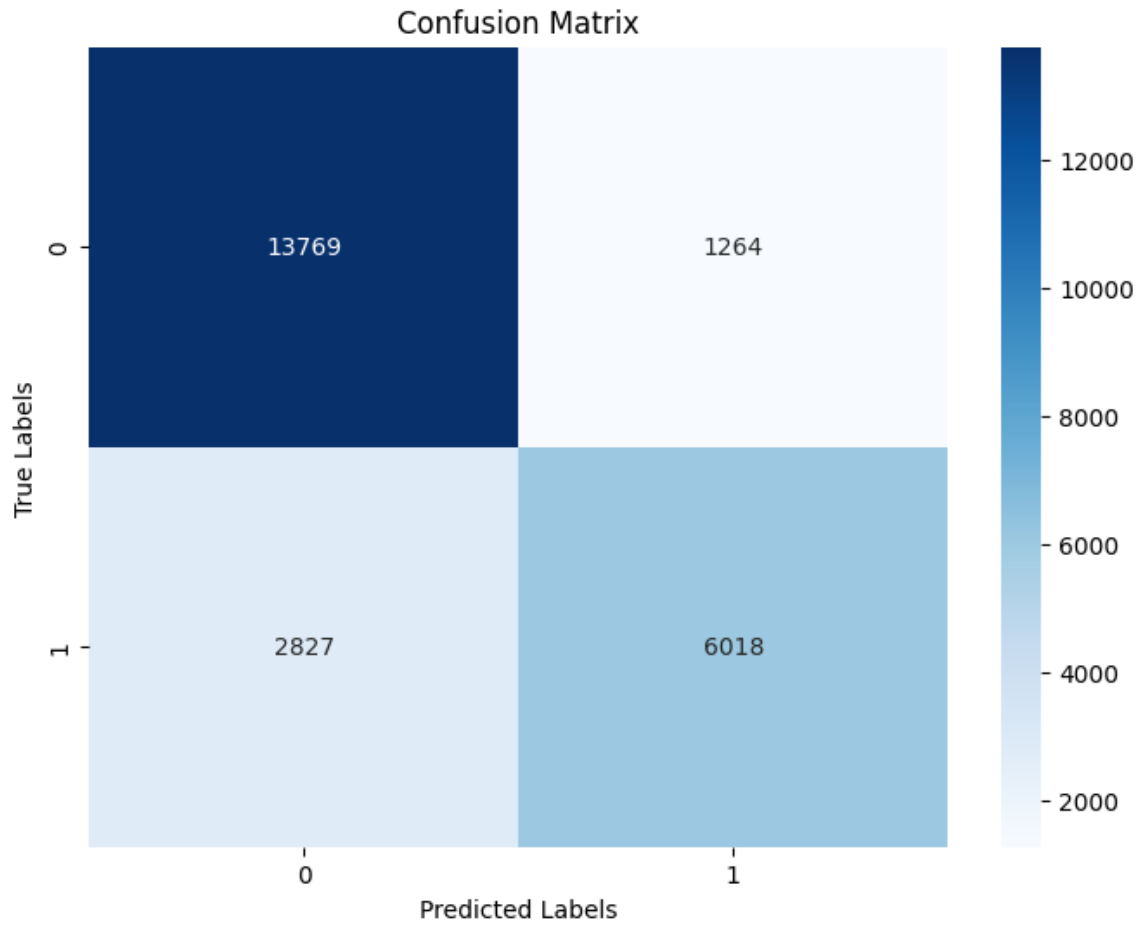


Figure 2: Confusion Matrix of DT Model

Inference: From this confusion matrix, we can infer the following:

- **True Negatives (TN)**: 13769
    - The model correctly predicted 13,769 bookings as not canceled.
- **False Positives (FP)**: 2827

- The model incorrectly predicted 2,827 bookings as canceled when they were not.

- **False Negatives (FN)**: 1264
  - The model incorrectly predicted 1,264 bookings as not canceled when they were actually canceled.

- **True Positives (TP)**: 6018
  - The model correctly predicted 6,018 bookings as canceled.

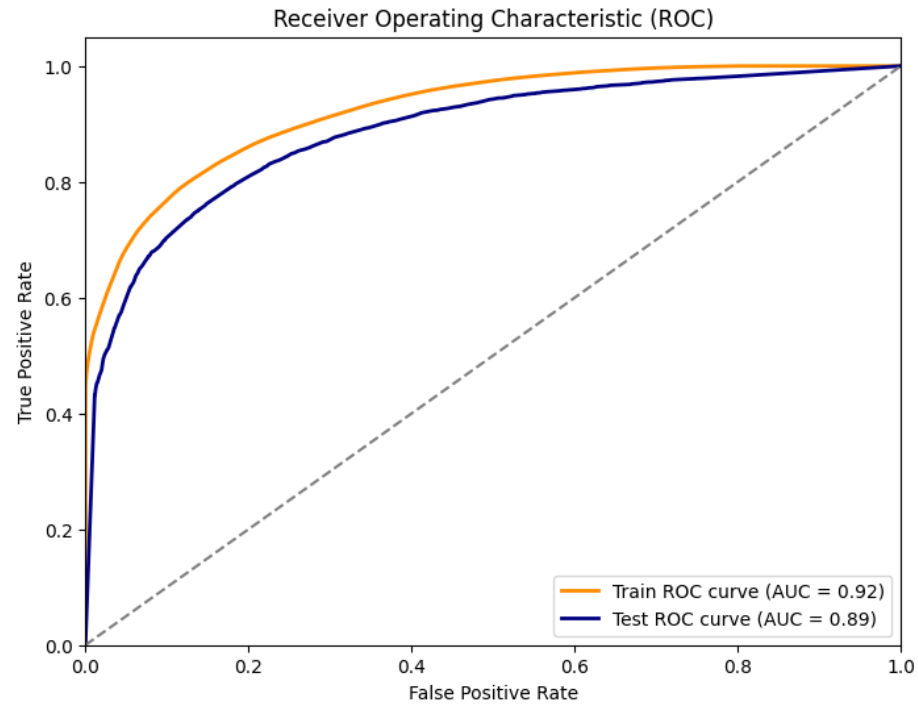
Classification Report:

	precision	recall	f1-score	support
0	0.83	0.92	0.87	15033
1	0.83	0.68	0.75	8845
accuracy			0.83	23878
macro avg	0.83	0.80	0.81	23878
weighted avg	0.83	0.83	0.82	23878

Inference:

- The model performs well overall with an accuracy of 83%.
- It is more effective at predicting non-cancellations (class 0) with higher recall (92%) compared to predicting cancellations (class 1) with recall (68%).
- Precision is the same (83%) for both classes, indicating that when the model predicts a class, it is equally likely to be correct.
- The F1-score, which balances precision and recall, is higher for class 0 (87%) than for class 1 (75%), indicating better performance in predicting non-cancellations.
- The model's performance can be considered balanced but shows room for improvement, especially in identifying cancellations (class 1) more accurately.

AUC-ROC Curve:




```
AUC for Training Set: 0.92
AUC for Testing Set: 0.89
```

Figure 3: AUC ROC DL MODEL

Inference:

AUC Scores:

- **Training Set**: 0.92
- **Testing Set**: 0.89

Key Inferences:

1. **Performance**: Excellent on both sets, with training AUC at 0.92 and testing AUC at 0.89.
2. **Generalization**: Minimal drop from training to testing indicates good generalization.
3. **Stability**: Small difference in AUC scores shows model stability.
4. **Practical Use**: High testing AUC (0.89) makes the model effective for predicting hotel booking cancellations.

Conclusion

The decision tree model demonstrates strong performance and reliability, making it effective for predicting hotel booking cancellations and aiding in operational planning.

Random forest model

A Random Forest classifier is an ensemble learning method based on decision trees. It creates multiple trees during training and predicts the mode (classification) or mean (regression) of the individual tree predictions. Random Forests introduce randomness by bootstrapping data samples and randomly selecting subsets of features for each tree split, reducing overfitting and enhancing generalization. They are robust, scalable, and effective for tasks like classification and regression across diverse datasets.

Confusion Matrix:

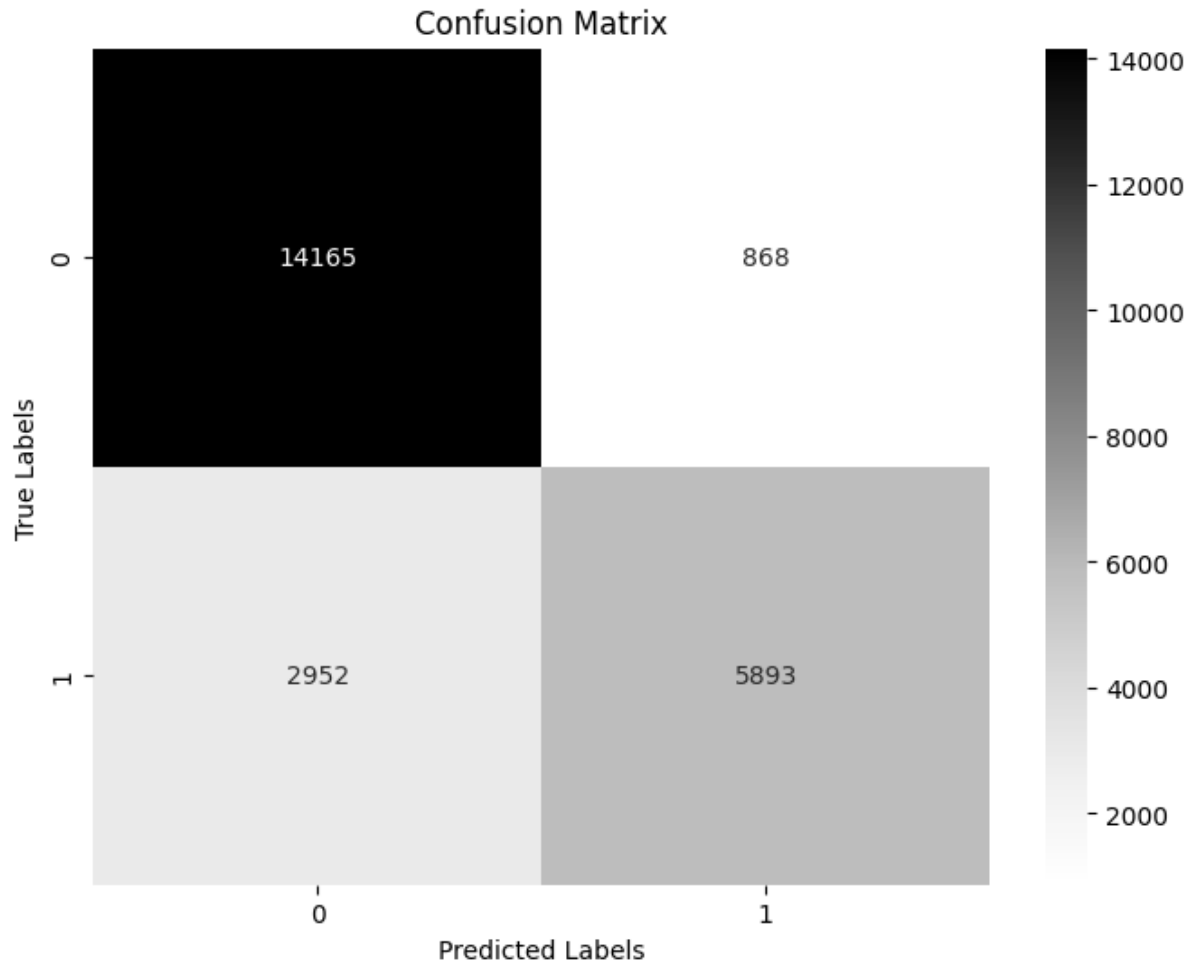


Figure 4: Confusion Matrix of RF Model

Inference:

1. **True Positives (TP)**: 5893
    - The model correctly predicted 5,893 bookings as canceled.
2. **False Positives (FP)**: 868
    - The model incorrectly predicted 868 bookings as canceled when they were not.

3. **False Negatives (FN)**: 2952
    - The model incorrectly predicted 2,952 bookings as not canceled when they were actually canceled.
4. **True Negatives (TN)**: 14165
    - The model correctly predicted 14,165 bookings as not canceled.

Classification Report:

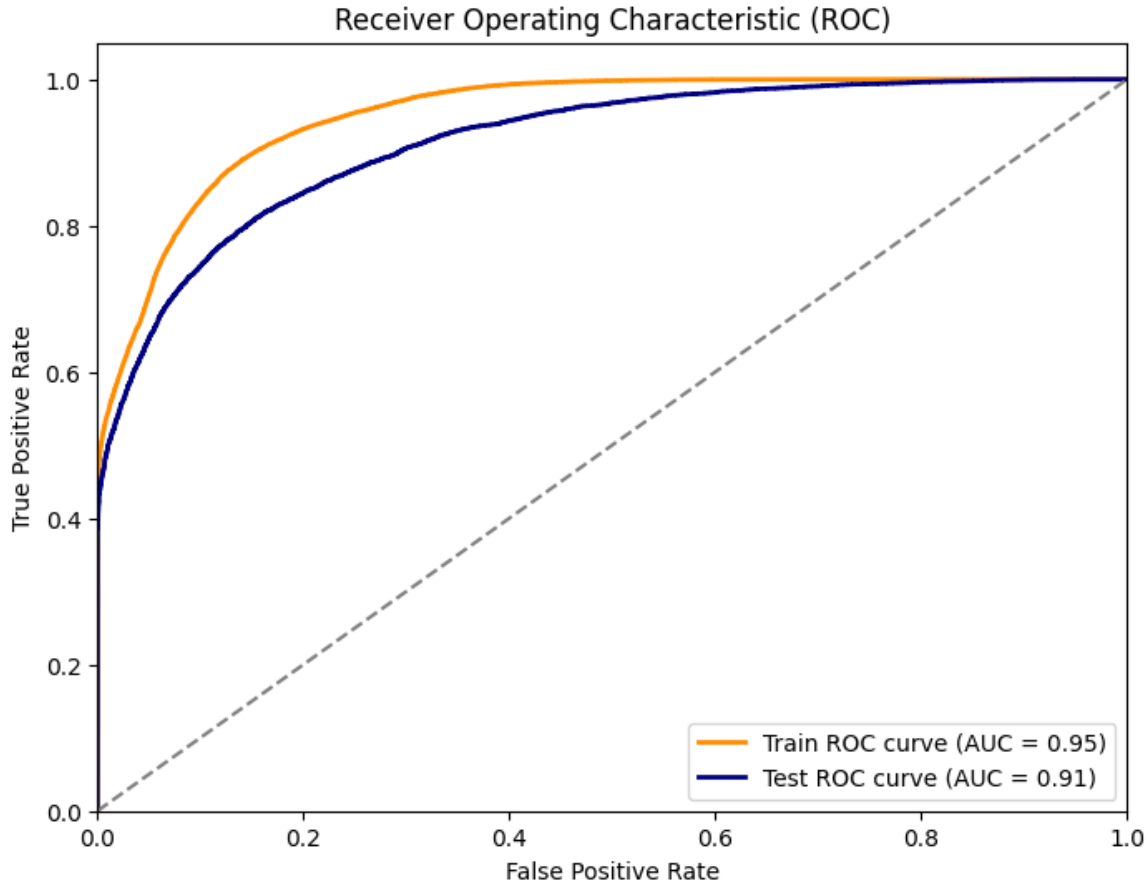
	precision	recall	f1-score	support
0	0.83	0.94	0.88	15033
1	0.87	0.67	0.76	8845
accuracy			0.84	23878
macro avg	0.85	0.80	0.82	23878
weighted avg	0.84	0.84	0.83	23878

Inference:

- The model demonstrates strong overall performance with an accuracy of 84%.
- It excels in predicting non-cancellations (class 0) with high recall (94%), indicating it effectively identifies most actual non-cancellations.
- For cancellations (class 1), the model shows lower recall (67%), suggesting it misses a significant number of actual cancellations.
- Precision is notably high for both classes: 83% for non-cancellations and 87% for cancellations, indicating that when the model predicts a class, it is generally correct.
- The F1-score is higher for non-cancellations (88%) compared to cancellations (76%), indicating better overall performance in predicting non-cancellations.

Overall, while the model shows robust performance in identifying non-cancellations, there is room for improvement in accurately predicting cancellations.

AUC – ROC Curve:




```
AUC for Training Set: 0.95
AUC for Testing Set: 0.91
```

Figure 5: AUC ROC curve RF Model

Inference:

Random forest model has produced better AUC score than DT Model. The random forest model demonstrates strong performance and reliability, making it effective for predicting hotel booking cancellations and aiding in operational planning than DT Model.

XGBoost Model:

XGBoost is a fast and efficient machine learning algorithm that uses gradient boosting to sequentially build decision trees, correcting errors along the way. It's known for its high performance in classification and regression tasks, incorporating regularization to prevent overfitting and handling large datasets effectively.

Confusion Matrix:

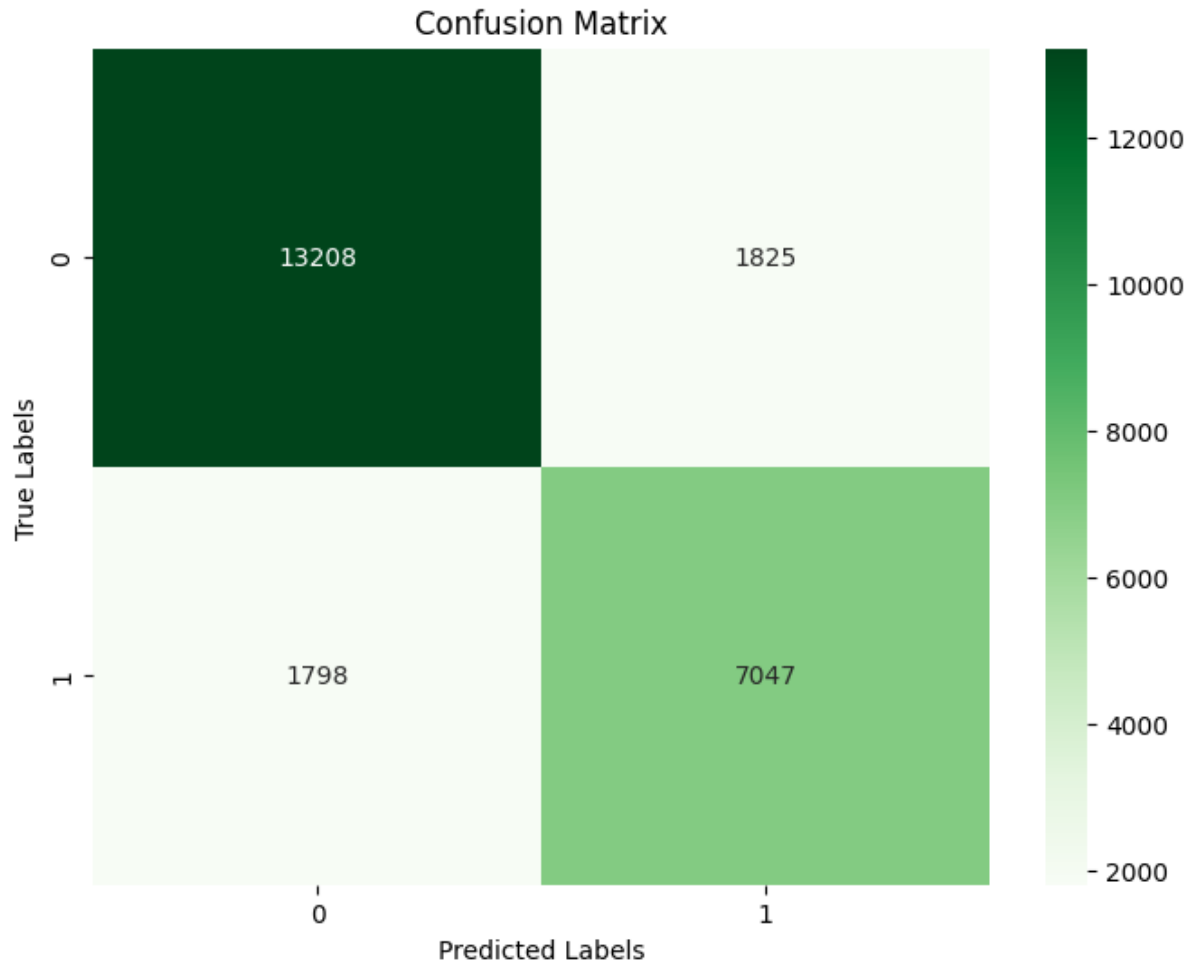


Figure 6: Confusion Matrix of XGBoost Model

Inference:

True Positives (TP): 7047

- The model correctly predicted 7,047 instances as positive (class 1).

False Positives (FP): 1825

- The model incorrectly predicted 1,825 instances as positive (class 1) when they were actually negative (class 0).

False Negatives (FN): 1798

- The model incorrectly predicted 1,798 instances as negative (class 0) when they were actually positive (class 1).

True Negatives (TN): 13208

- The model correctly predicted 13,208 instances as negative (class 0).

Classification report:

	precision	recall	f1-score	support
0	0.88	0.88	0.88	15033
1	0.79	0.80	0.80	8845
accuracy			0.85	23878
macro avg	0.84	0.84	0.84	23878
weighted avg	0.85	0.85	0.85	23878

Inference:

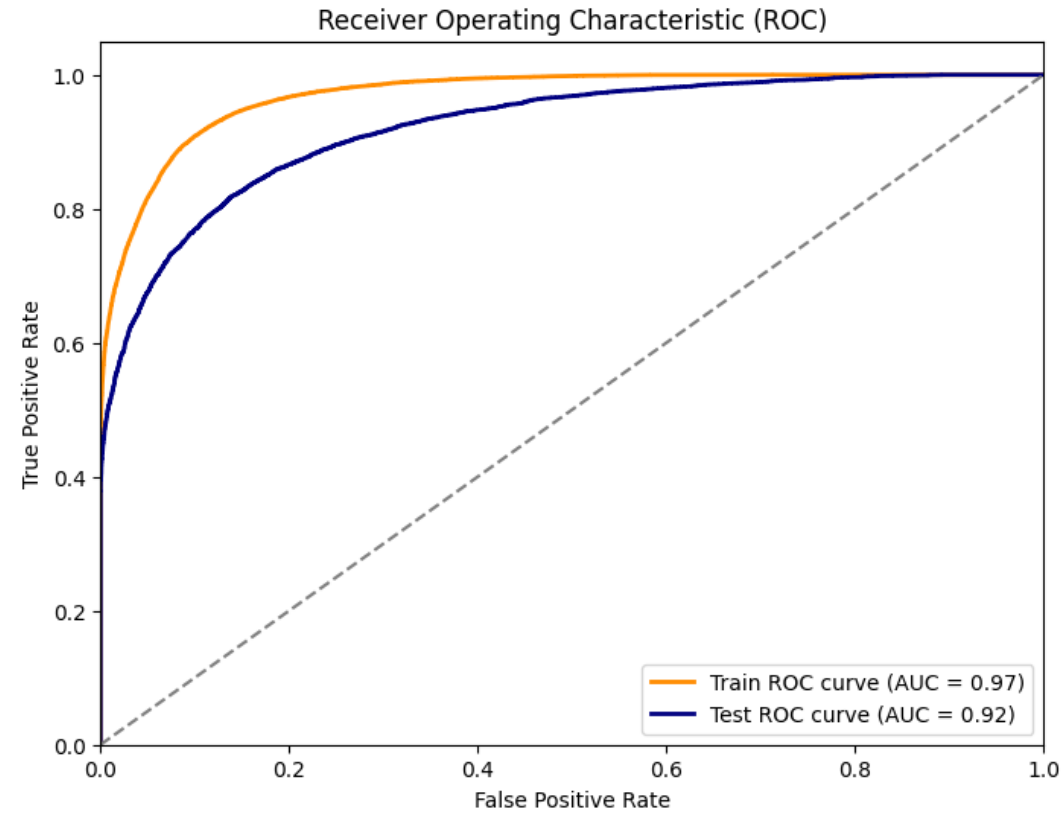
• The model demonstrates strong overall performance with an accuracy of 85%.

• It shows balanced precision and recall for both classes: 88% precision and 88% recall for non-cancellations (class 0), and 79% precision and 80% recall for cancellations (class 1).

• The F1-score, which balances precision and recall, is higher for non-cancellations (88%) than for cancellations (80%), indicating better performance in predicting non-cancellations.

• The model effectively identifies most actual non-cancellations but misses a notable number of actual cancellations, suggesting potential for improvement in predicting cancellations.

AUC-ROC Curve:




```
AUC for Training Set: 0.97
AUC for Testing Set: 0.92
```

Figure 7:AUC ROC Curve XGB Model

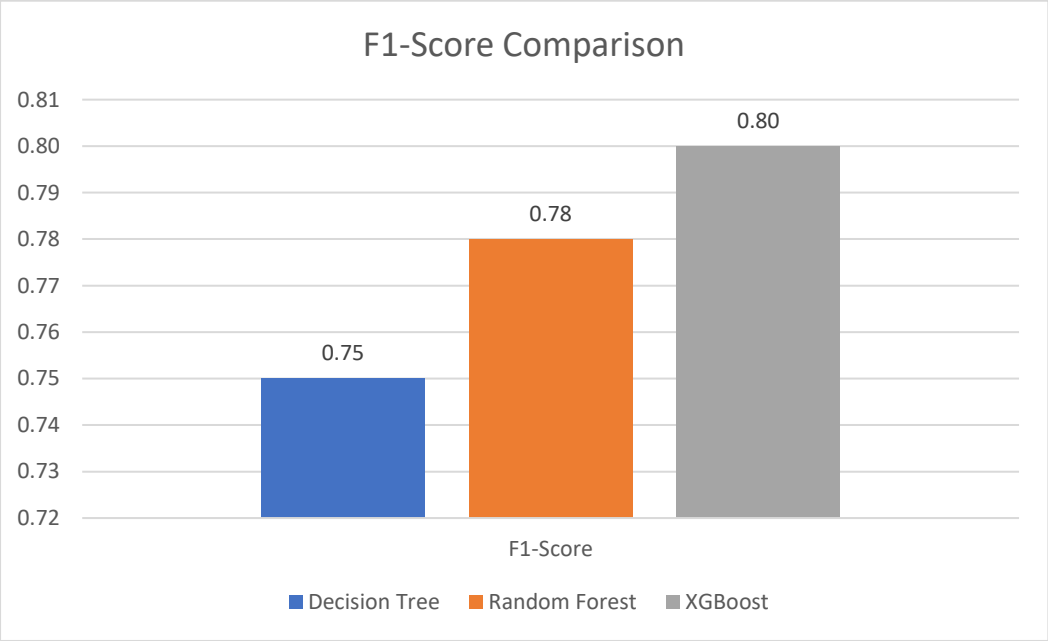
Inference:

- The XGBoost model shows excellent discrimination with AUC-ROC scores of 0.97 on the training set and 0.92 on the test set, indicating strong performance in distinguishing between classes.
- It demonstrates robust training fit and good generalization to unseen data.

Model Comparison:

Model / (Class 1)	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree	0.83	0.83	0.68	0.75	0.89
Random Forest	0.84	0.87	0.67	0.78	0.91
XGBoost	0.85	0.79	0.80	0.80	0.92

As previously mentioned, the critical metric for this project is the F1-score for class '1' (canceled). A high F1-score indicates a balanced approach to minimizing the costs associated with false negatives (overbooking rooms expecting cancellations that do not occur) and false positives (predicting cancellations when guests actually arrive). This balance is crucial for optimizing room availability while reducing overbooking and ensuring customer satisfaction.



Among the trained and tested models, XGBoost classifier performs the best.

We further create a feature importance table to understand the contribution of each features in classification and sort the top 5 features contributing the most towards the outcome of predicition

Ranking	Features contribution in classification	percentage
1	deposit_type_Non Refund	66.0%
2	required_car_parking_spaces	10.1%
3	previous_cancellations	4.9%
4	market_segment_Online TA	2.0%
5	customer_type_Transient	1.1%

Recommendation:

- From the XGBoost Classification model, we can understand deposit type plays a major factor in deciding if the customer is going to cancel the booking or not, followed up by required car parking spaces. Other minor factors are previous cancellations, online market segment, customer type transients and 44 others columns.
- This suggest financial and logistic aspects are determining the cancellation of bookings for this hotel.
- Further understanding and collecting data from customers regarding financials and strategizing their pricing accordingly could help improve the booking and inversely reduce cancellation rate.