**HOTEL BOOKING CANCELLATION**

**MACHINE LEARNING**

A pool with a gazebo and chairs

Description automatically generated

**I.SCENARIO**

**II.DATASET OVERVIEW**

**III. PREPROCESSING DATA**

**IV.ANALYSIS AND FINDINGS**

**V.MACHINE LEARNING**

**IV. SUGGESTIONS**

# **I.SCENARIO**

**1. Business Problem**

You are a Data Analyst at the famous city hotel named The Continental. The company is presently engaged in the business of operating City Hotel and Resort Hotel. But in the recent year, The hotel has witnessed high cancellation rates. Each of the branch hotels is now dealing with a number of issues as a result, including a significant decline in revenues and less than ideal hotel room use. As a result, lowering the cancellation rates is the highest priority of the branch hotels to increase their efficiency in generating revenue and for us to offer thorough business advice to address this problem.



**2. Assumptions**

1. No unusual occurrences between 2015 and 2017 will substantially impact the data used.
2. The information is still current and can be used to analyze a hotel's possible plans.
3. There are no unanticipated negatives to the hotel.
4. The most significant factor affecting the effectiveness of earning income is booking cancellations.
5. The hotels are not currently using any of the suggested solutions.

**3. Research Question**

1. What are the factors that affect hotel reservation cancellations?
2. How can we make hotel reservation cancellations better?
3. How will hotels be assisted in making pricing and promotional decisions?

# **II.DATASET OVERVIEW**

1. **Introduction**

The data set source :*<https://www.kaggle.com/jessemostipak/hotel-booking-demand>*

This data set consists of ***119,390*** observations and holds booking data for a city hotel and a resort hotel from 2015 to 2017. It has 32 variables which include *reservation and arrival date, length of stay, canceled or not, the number of adults, children, or babies, the number of available parking spaces, how many special guests, companies, and agents pushed the reservation*, etc.

1. **Details of Data**

|  |  |
| --- | --- |
| **hotel** :(H1 = Resort Hotel or H2 = City Hotel). | **reserved\_room\_type**: Code of room type reserved. Code is presented instead of designation for anonymity reasons. |
| **is\_canceled Value**: showing if the booking had been cancelled (1) or not (0). | **assigned\_room\_typ**e: Code for the type of room assigned to the booking. Code is presented instead of designation for anonymity reasons. |
| **lead\_time**: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date. | **booking\_changes**: How many times did booking changes happen |
| **arrival\_date\_year:** Year of arrival date. | **deposit\_type**: Indication on if the customer deposited something to confirm the booking. |
| **arrival\_date\_month**: The months in which guests are coming. | **agent:** If the booking happens through agents or not. |
| **arrival\_date\_week\_number**: Week number of year for arrival date. | **company**: If the booking happens through companies, the company ID that made the booking or responsible for paying the booking. |
| **arrival\_date\_day\_of\_month**: Which day of the months guest is arriving. | **days\_in\_waiting\_list**: Number of days the booking was on the waiting list before the confirmation to the customer. |
| **stays\_in\_weekend\_nights**: Number of weekend stay at night (Saturday or Sunday) the guest stayed or booked to stay at the hotel. | **customer\_type**: Booking type like Transient – Transient-Party – Contract – Group. |
| **stays\_in\_week\_nights:** Number of weekdays stay at night (Monday to Friday) in the hotel. | **adr**: Average Daily Rates that described via way of means of dividing the sum of all accommodations transactions using entire numbers of staying nights. |
| **adults**: Number of adults. | **required\_car\_parking\_spaces**: How many parking areas are necessary for the customers. |
| **children**: Number of children. | **total\_of\_special\_requests**: Total unique requests from consumers. |
| **babies**: Number of babies. | **reservation\_status**: The last status of reservation, assuming one of three categories: Canceled – booking was cancelled by the customer; Check-Out |
| **meal:** Type of meal booked. | **reservation\_status\_date**: The last status date. |
| **country:** Country of origin. | **is\_repeated\_guest**: The values indicating if the booking name was from a repeated guest (1) or not (0) |
| **market\_segmen**t: Through which channel hotels were booked. | **previous\_cancellations:** Show if the repeated guest has cancelled the booking before. |
| **distribution\_channel:** Booking distribution channel. | **previous\_bookings\_not\_canceled**: Show if the repeated guest has not cancelled the booking before. |

**III PREPROCESSING DATA**

Data preprocessing is a data mining technique used to transform raw data into a useful and efficient format. The team primarily performs data preprocessing tasks such as filling in missing values, removing duplicate data, handling outliers, data transformation, data splitting, and feature scaling.

* + - 1. **Import Libraries**

After collecting data we have to import the necessary libraries to build machine learning models. Numpy, Pandas, Matploilib, and Seaborn are the known libraries used in the machine learning model.

+ **Numpy:** NumPy is a Numerical Python Library that helps perform mathematical operations.

**+ Pandas:** Panda is an open-source library that helps understand relational or labelled data.

**A black and grey rectangular object

Description automatically generated**

The data shown below represents the Hotel booking cancellation dataset that is available on Kaggle

A screenshot of a computer

Description automatically generated A black and white text on a grey background

Description automatically generated

**2. Data Cleaning**

Real-world raw data is often incomplete, inconsistent, and lacking in certain behaviors or trends. The raw data contain many errors. So, once collected, the next step machine learning pipeline is to clean data which refers to removing unwanted data.

Some steps which are used to clean data are:

* Remove missing values, outliers, and unnecessary rows/ columns.
* Check and impute null values.
* Check Imbalanced data.
* Re-indexing and reformatting our data.

**First**, have to check if our data holds duplicate values.

A screen shot of a computer

Description automatically generatedPandas **duplicated()** method helps to return duplicate values only, and any() method returns a boolean series that is True only if the conditions are true.

The dataset holds the duplicate entries.

A black background with blue and yellow text

Description automatically generatedPandas **drop\_duplicate()**method helps to remove duplicate entries.  
The keep parameter is set to False so that only Unique values are taken and the duplicate values are removed from the data.

**A screenshot of a computer

Description automatically generated**

**A black rectangular object with white text

Description automatically generatedSecondly,** we should check the number of missing values in each column and the percentage of missing values in the columns. Pandas **isna()** method helps to find out null values and also convert the result to the percentage for review purpose

**A computer code with text

Description automatically generated with medium confidence**

**A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated**

Column company holds 94% missing data, so we can drop that column.

A close-up of a computer screen

Description automatically generated

Now we check the result after dropped

A screenshot of a computer screen

Description automatically generated

**Finally**, before analysts make a prediction, the data should have been encoded and categorized.

 Dropping columns that are not useful :

Conducting categorization by following the codes :

A screenshot of a computer

Description automatically generated A screenshot of a computer code

Description automatically generated

City Hotel and Resort Hotel are string types, so we need to encode categorical variables

A screenshot of a computer

Description automatically generatedThe results after encoding:

A screen shot of a computer code

Description automatically generated

The Dataset is ready to Analyst and Predict.

**IV ANALYSIS AND FINDINGS**

**A blue rectangular bars with white text

Description automatically generatedReservation status**



0 : not canceled

1 : canceled

The accompanying bar graph shows the percentage of reservations that are canceled

and those that are not. It is obvious that there are still a significant number of

reservations that have not been canceled. There are still **37% of clients who canceled**

their reservations, which has a significant impact on the hotels' earnings.

**2. Reservation status in different hotels**

**A graph of blue squares

Description automatically generated**

Compared to the Resort Hotel, it's evident that the City Hotel has a higher volume of bookings, which consequently leads to a larger number of cancellations. Consequently, there has been a substantial decrease in booking revenue, particularly from the City Hotel, where the number of cancellations has exceeded 30,000 times.

A graph of blue and orange lines

Description automatically generated**3. Average Daily Rate in City and Resort Hotel**

**adr**: Average Daily Rates that described via way of means of dividing the sum of all accommodations transactions using entire numbers of staying nights

*To calculate the ADR, you would add up the total revenue generated from all the room bookings and then divide it by the total number of nights that guests stayed in those rooms. This gives you an average cost per room per day. It's a useful metric because it helps hotel managers and analysts understand the* ***average pricing of their rooms and track trends in pricing over time.***

*By calculating the ADR, hotels can assess their pricing strategies, monitor changes in customer demand, and make informed decisions about their room rates to maximize revenue and profitability.*

According to the line chart above that illustrates the daily rate for a City hotel is occasionally lower than Resort Hotel. Furthermore, there are circumstances where the City Hotel’s even lower nearly reached **0%** Average Daily Rate around in March and April 2015. It is evident that weekends and holidays could witness an increase in the rates for Resort Hotel.

**4. Reservation per month**

**A graph of blue and orange bars

Description automatically generated**

The grouped bar graph examines the greatest and lowest reservation levels by status of reservations for each month. The data illustrated that the month of August exhibits the highest figures, reached at the top with **8000** for confirmed bookings and the lowest for cancellations. Conversely, it is notable that the month of January stands out due to having the highest count of cancellations of bookings with nearly **6000 cancellations.**

**A graph of different colored bars

Description automatically generated5. ADR for each month**

The depicted bar graph provides clear evidence that instances of cancellations are observed when **accommodation prices reach their highest points**, **while occurrences of cancellations diminish notably when prices are at their lowest**. Hence, it can be deduced that the sole determinant for these cancellations is the **cost** **associated** with the accommodations.

**A pie chart with numbers and a triangle in the center

Description automatically generated6. Top 10 countries with reservations cancelled**

Based on the information presented by the provided pie chart, it is discernible that Portugal emerges as the leading country, contributing to **70%** of the total canceled reservations.

**7. Market segment**

**A screenshot of a computer

Description automatically generated**

The area from where guests visit the hotels and make reservations.

Is it coming from Direct or Groups, Online or Offline Travel Agents? The given dataset shows that **20%** of customers come from traditional travel agents and **47%** from online travel agencies respectively come from **offline travel agencies**. whereas 33% of customers are from untapped markets.

**V MACHINE LEARNING**

**1. Model Building**

The 4 models were applied in my research to forecast and assess the cancellation.

+ **KNN**

+ **Decision Tree**

+ **Random Forest**

+ **Gradient Boosting**

A computer screen shot of a computer code

Description automatically generatedA screenshot of a computer program

Description automatically generatedImport necessary libraries:

**A screenshot of a computer code

Description automatically generateda) KNeighbors regression**

**A screenshot of a computer

Description automatically generated**

**b) Decision Tree**

**A screenshot of a computer program

Description automatically generated**

**c) Random Forest**

**A screenshot of a computer program

Description automatically generated**

**d) Gradient Boosting**

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer program

Description automatically generated

**2. Visualizing the results**

**A screenshot of a computer code

Description automatically generated** **a) Accuracy score**

A chart of different models

Description automatically generated

**b) ROC Curve**

**A screenshot of a computer program

Description automatically generated**

A graph of a curve

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated**c) Confusion matrix**

A comparison of a graph

Description automatically generated with medium confidenceA group of blue and green squares

Description automatically generated

**VI.SUGGESTION**

**1. Business problems**

**a) Dynamic Pricing and Discounts**

**+** **Seasonal Adjustments**: Noticing that cancellations are highest in January due to price increases, the hotels should reconsider their pricing strategy for that month. Given that ADR is lowest in January, it might be beneficial to offer special discounts or promotions during this period to encourage bookings

**+ Peak Pricing:** August has the highest ADR, suggesting high demand. The hotel could employ a dynamic pricing strategy where prices slightly increase during peak times but offer non-refundable rates at discounted prices to reduce cancellations**.**

**b) Partnerships with Travel Agencies**

**+**With **47%** bookings coming from online travel agencies and **20%** from offline agencies, strengthening partnerships with these agencies can be beneficial. Consider offering special rates for agency customers, and in return, ask agencies to push for non-refundable bookings.

**+Collaborate with** these agencies for joint **marketing campaigns** or **bundle offers** (like combining hotel stay with sightseeing tours) to attract more customers

**c) Flexible Booking Policies with Incentives:**

+ While non-refundable rates can deter cancellations, they might not be appealing to all. Therefore, provide a flexible booking option but with incentives for those who don't cancel. For instance, if a customer doesn't cancel, they might get a free meal, a discount for a room upgrade, or a discount on their next stay.

+ Consider implementing a policy where customers can reschedule (instead of canceling) their stay to a future date without penalty. This ensures that the hotel retains the revenue even if the customer can't make their original dates

**2. Model machine learning**

+ **Accuracy:**  
KNN: 84.85%  
Decision Tree: 91.37%

Random Forest: 95.99%  
Gradient Boosting: 86.54%

* **The highest accuracy is obtained by the Random Forest model.**

**+ Precision**:

KNN: 78%

Decision Tree: 84%

Random Forest: 99%

Gradient Boosting: 89% for

* **The highest precision for class 1 is obtained by the Random Forest model.**

**+ Recall:**

KNN: 62%

Decision Tree: 85%

Random Forest: 87%

Gradient Boosting: 58%   
=> **The highest recall for class 1 is obtained by the Random Forest model.**

+ **F1-Score**:

KNN: 69%

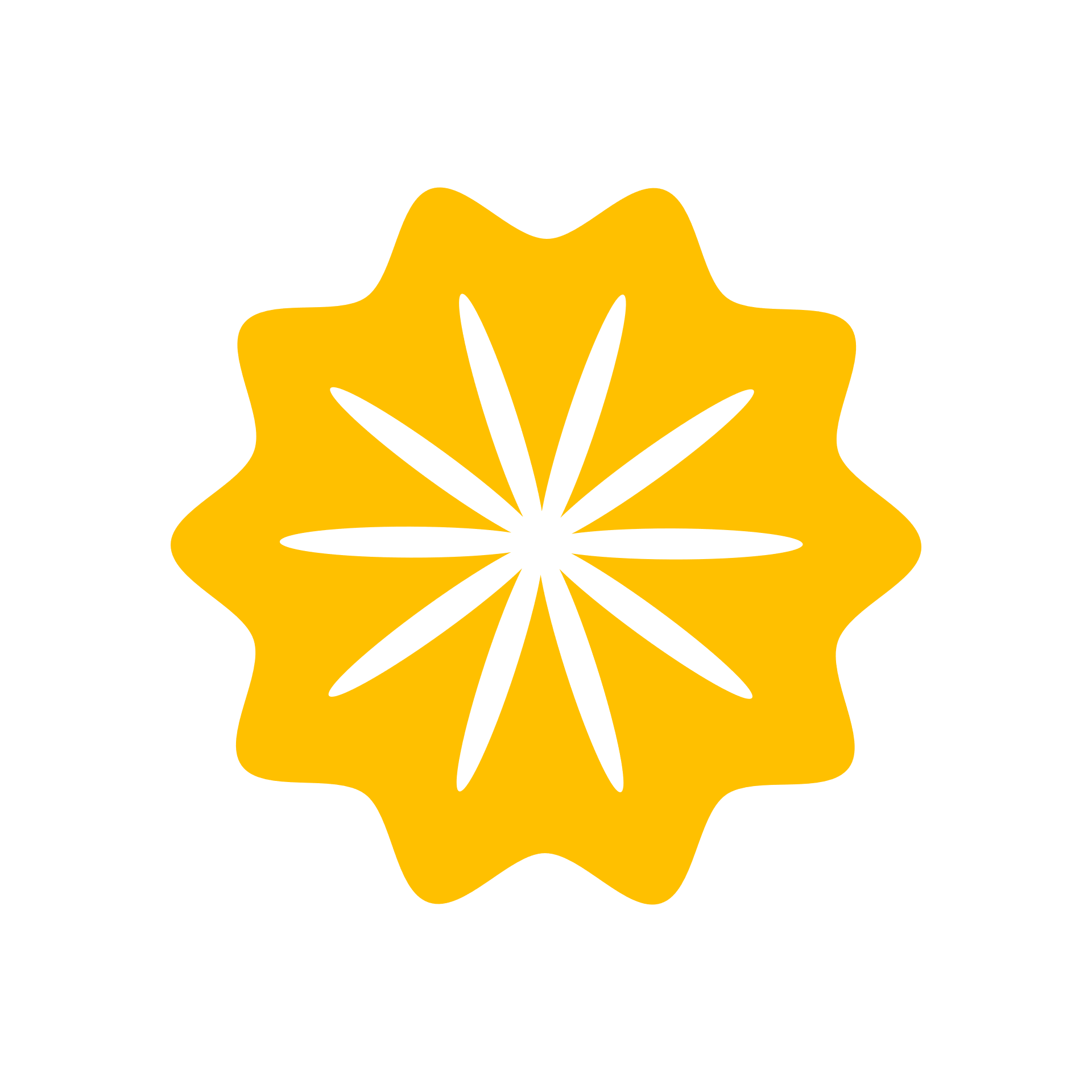
Decision Tree: 84%

Random Forest: 92%

Gradient Boosting: 70%

* **The highest F1-Score for class 1 is obtained by the Random Forest** model.

**Conclusion**: Based on the results, the Random Forest model seems to be the best performer for predicting cancellations. It has the highest accuracy, precision, recall, and F1 score among all the models.

**** The End