HOTEL RESERVATION CANCELLATION PREDICTION

119cs0025
T. Swetha Reddy

Problem Statement

- Given a dataset containing data of reservations made by customers in different hotels, build a machine learning model to predict whether the customer cancels his/her hotel reservation or not.

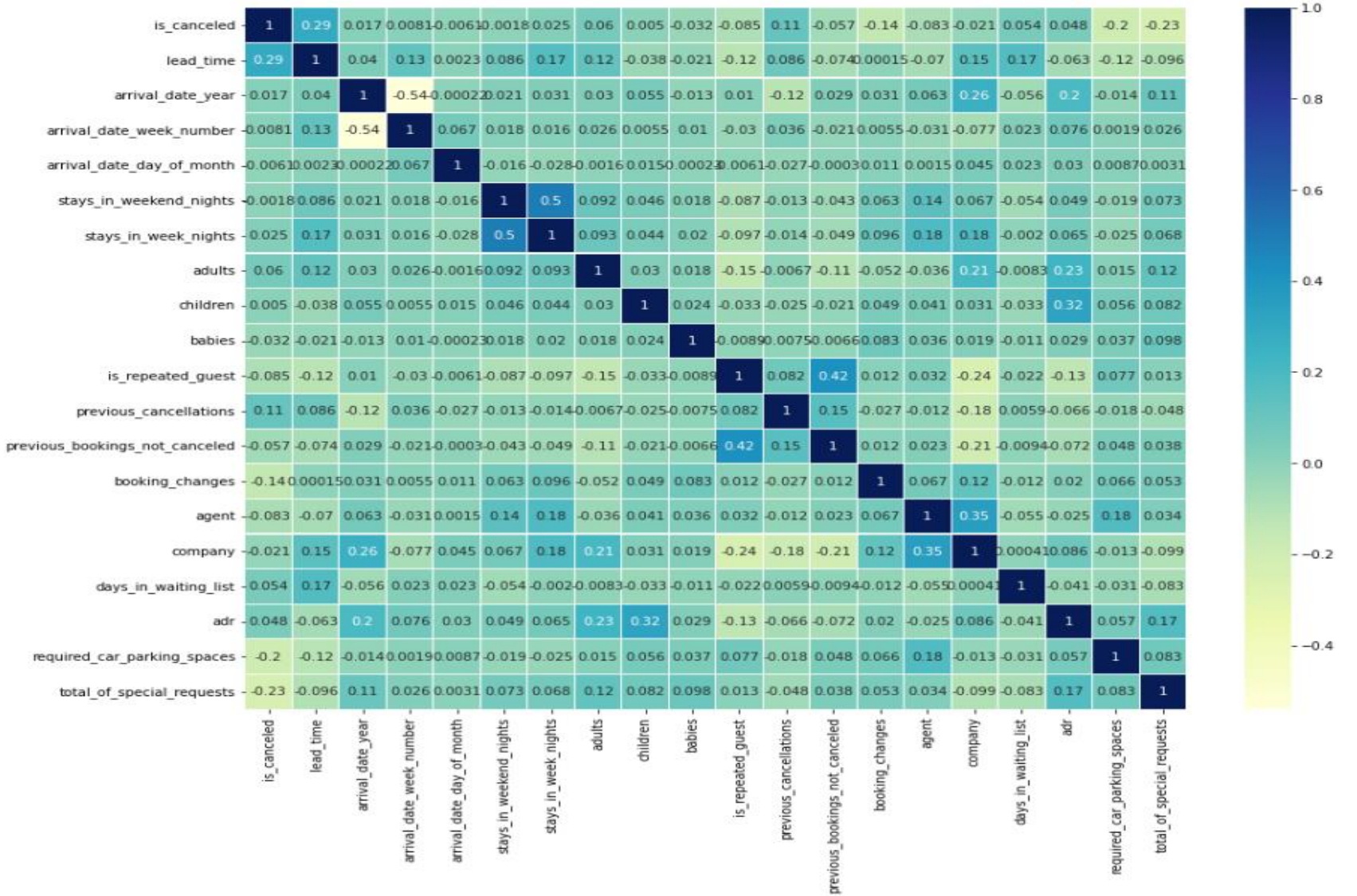
About the Dataset

- No. of rows = 119390
- No. of attributes = 32
- **Target variable = is_cancelled**
- No. of independent variables = 31
- No. of numeric variables = 12
- No. of object variables = 19

Independent variables in the dataset

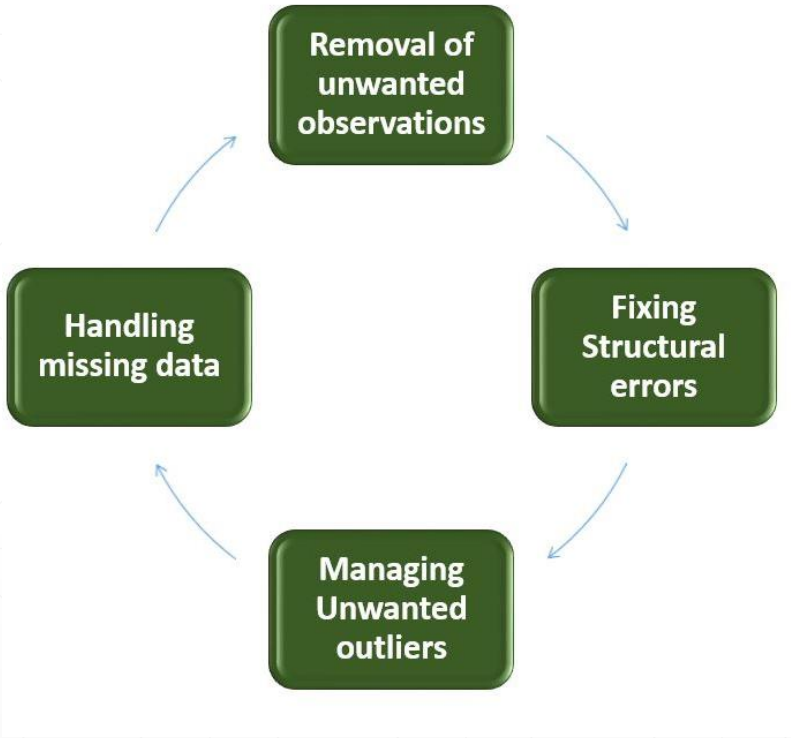
- Hotel
- Lead_time
- Arrival_date_year
- Arrival_date_month
- Arrival_date_week_number
- Arrival_date_day_of_month
- Stays_in_weekend_nights
- Stays_in_week_nights
- Adults
- Children
- Babies
- Meal
- Country
- Market_segment
- distribution_channel
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_cancelled
- reserved_room_type
- assigned_room_type
- booking_changes
- deposit_type
- agent
- company
- days_in_waiting_list
- customer_type
- adr
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
- reservation_status_date

Correlation Matrix



Data Cleaning

Data Cleaning is the process of identifying the incorrect, incomplete, inaccurate, duplicated, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity



- **Replacing NULL/MISSING Values**
  - Replacing numerical missing values with **MEAN/MEDIAN**
    - No. of null values in [children] = 4
    - No. of null values in [agent] = 16340
    - No. of null values in [company] = 112593

Data Cleaning

- **Replacing NULL/MISSING Values**
  - Replacing categorical missing values with **MODE**
    - No. of null values in [children] = 4
    - No. of null values in [agent] = 16340
    - No. of null values in [company] = 112593
- **Removing the Duplicate Values**
  - No. of duplicate values in the data set = 32013
  - Since we have 32013 duplicate records in the data, we will remove this from the data set so that we get only distinct records. Post removing the duplicate, we will check whether the duplicates have been removed from the data set or not.
  - No. of rows in the dataset after removing duplicates = 87377

Encoding Categorical Data

- Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.
- An approach to encoding categorical values is to use a technique called label encoding. Label encoding is simply converting each value in a column to a number.

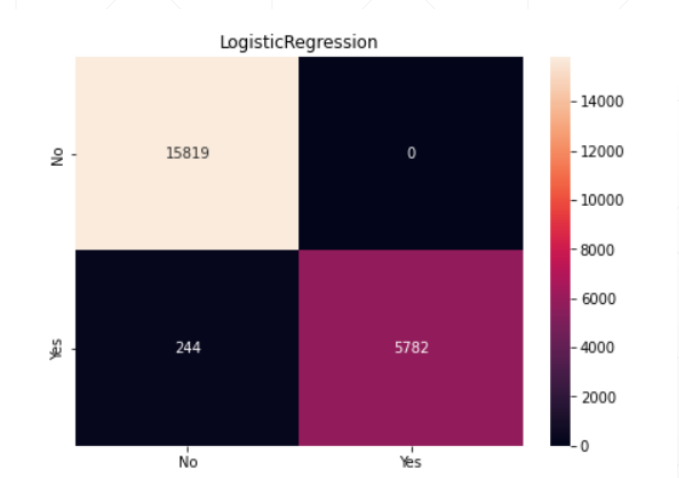
Categorical variables in our data set:

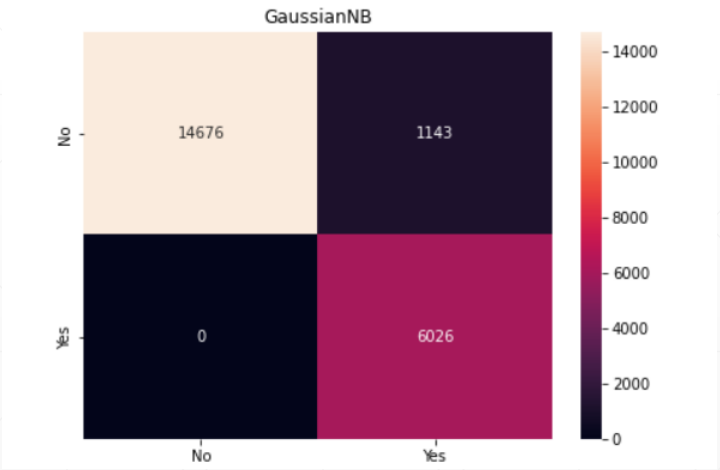
- Hotel
- arrival_date_month
- meal
- country
- Market_segment
- distribution_channel
- reserved_room_type
- assigned_room_type
- deposit_type
- customer_type
- reservation_status
- reservation_status_date

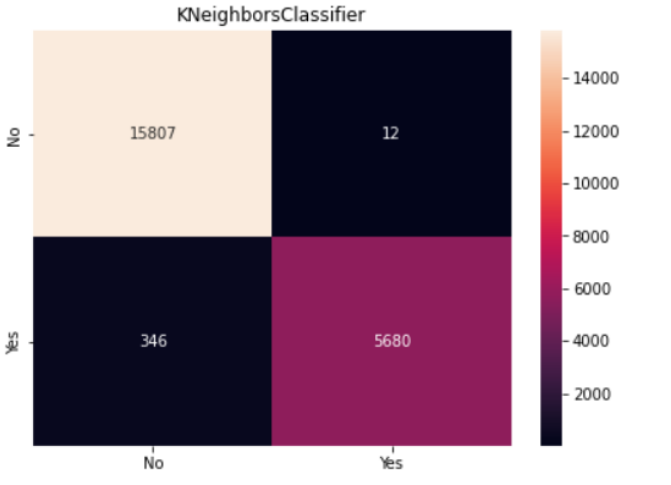
Training the Model

- Size of training data = 75%
- Size of testing data = 25%

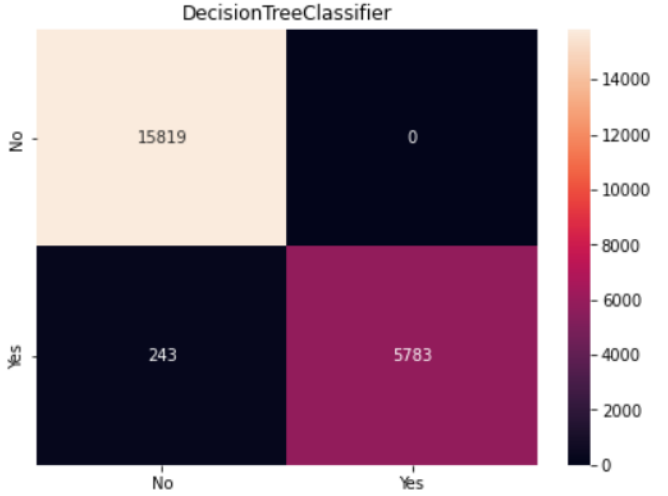
Confusion Matrix

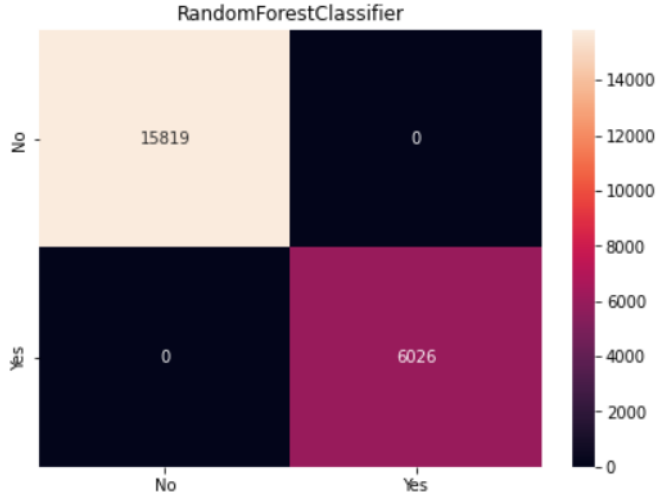


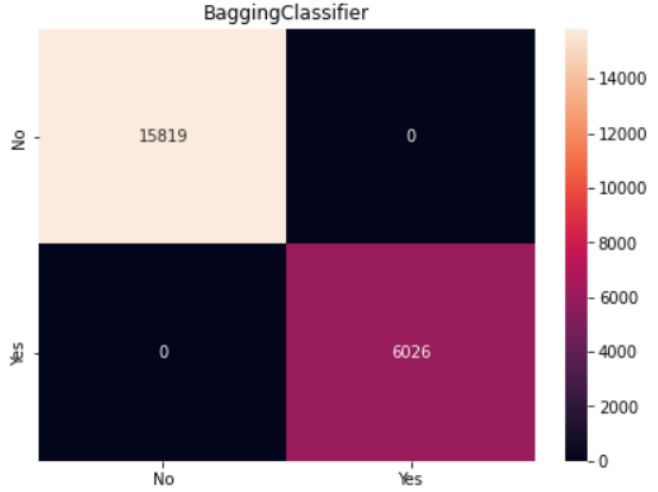


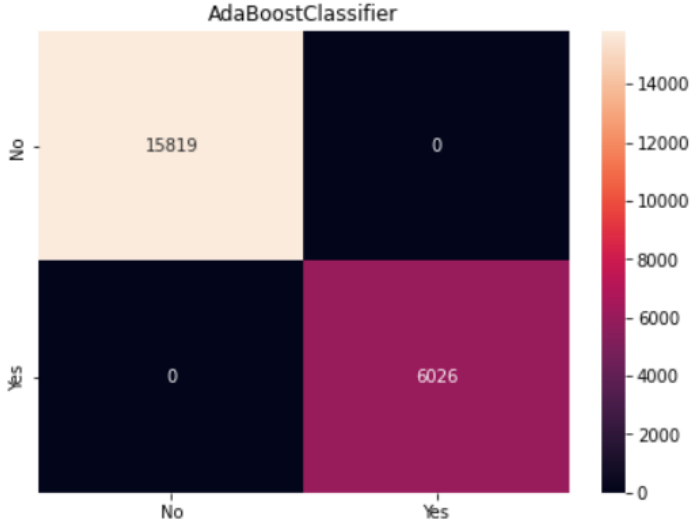


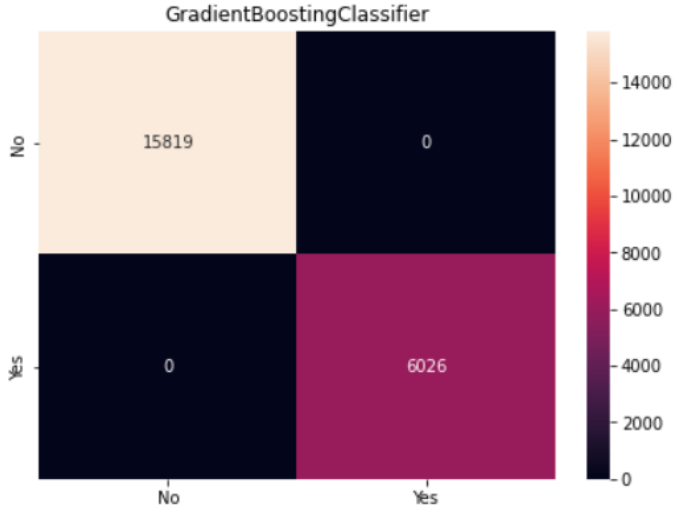
Confusion Matrix











Accuracy using Various Classifiers

- **Logistic Regression**
  - Training Accuracy : `0.9882194958188366`
  - Testing Accuracy : `0.9888303959716183`
- **Gaussian Naive Bayes**
  - Training Accuracy : `0.946056888237807 5`
  - Testing Accuracy : `0.9476768139162279`
- **K Neighbors Classifier**
  - Training Accuracy : `0.9882500152597204`
  - Testing Accuracy : `0.983611810482948`
- **Decision Tree Classifier**
  - Training Accuracy : `0.9882347555392785`
  - Testing Accuracy : `0.9888761730373083`
- **Random Forest Classifier**
  - Training Accuracy : `1.0`
  - Testing Accuracy : `1.0`

Conclusion

- The highest accuracy in this problem is obtained using the **Random Forest Classifier.**
- Highest accuracy **= 100%**