Capstone Project

119CS0013-KAMMILA SWATHI

Problem Statement

Hotel Reservation Cancellation Prediction

Given a Dataset containing data of reservations made by customers in different hotels, train the machine learning model to predict whether the customer cancels hotel reservation or not.

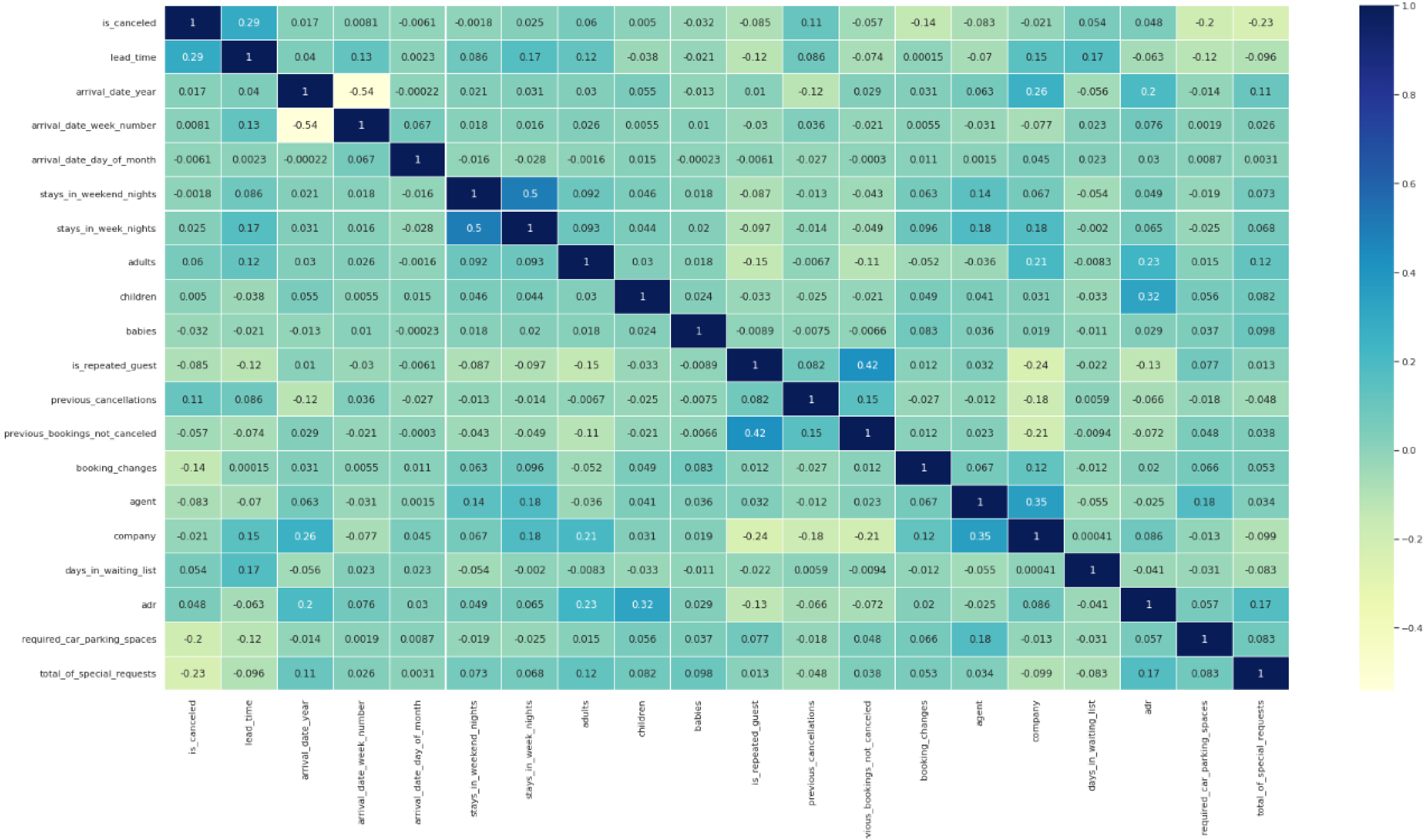
Dataset Details

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

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
- adr
- 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
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
- reservation_status_date

Correlation Matrix



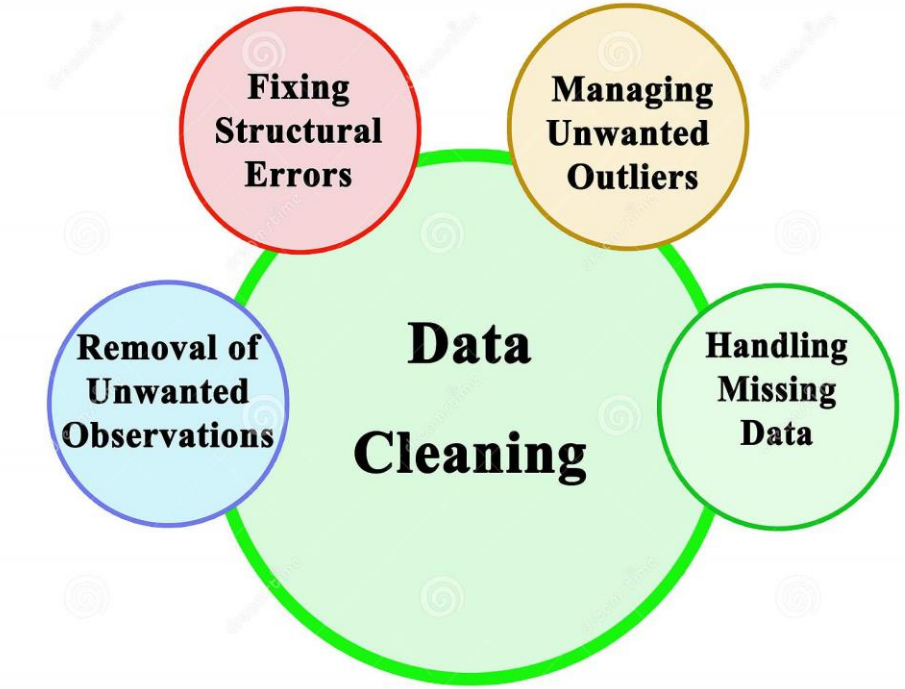
	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	vious_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests
is_canceled	1	0.29	0.017	0.0081	-0.0061	-0.0018	0.025	0.06	0.005	-0.032	-0.085	0.11	-0.057	-0.14	-0.083	-0.021	0.054	0.048	-0.2	-0.23
lead_time	0.29	1	0.04	0.13	0.0023	0.086	0.17	0.12	-0.038	-0.021	-0.12	0.086	-0.074	0.00015	-0.07	0.15	0.17	-0.063	-0.12	-0.096
arrival_date_year	0.017	0.04	1	-0.54	-0.00022	0.021	0.031	0.03	0.055	-0.013	0.01	-0.12	0.029	0.031	0.063	0.26	-0.056	0.2	-0.014	0.11
arrival_date_week_number	0.0081	0.13	-0.54	1	0.067	0.018	0.016	0.026	0.0055	0.01	-0.03	0.036	-0.021	0.0055	-0.031	-0.077	0.023	0.076	0.0019	0.026
arrival_date_day_of_month	-0.0061	0.0023	-0.00022	0.067	1	-0.016	-0.028	-0.0016	0.015	-0.00023	-0.0061	-0.027	-0.0003	0.011	0.0015	0.045	0.023	0.03	0.0087	0.0031
stays_in_weekend_nights	-0.0018	0.086	0.021	0.018	-0.016	1	0.5	0.092	0.046	0.018	-0.087	-0.013	-0.043	0.063	0.14	0.067	-0.054	0.049	-0.019	0.073
stays_in_week_nights	0.025	0.17	0.031	0.016	-0.028	0.5	1	0.093	0.044	0.02	-0.097	-0.014	-0.049	0.096	0.18	0.18	-0.002	0.065	-0.025	0.068
adults	0.06	0.12	0.03	0.026	-0.0016	0.092	0.093	1	0.03	0.018	-0.15	-0.0067	-0.11	-0.052	-0.036	0.21	-0.0083	0.23	0.015	0.12
children	0.005	-0.038	0.055	0.0055	0.015	0.046	0.044	0.03	1	0.024	-0.033	-0.025	-0.021	0.049	0.041	0.031	-0.033	0.32	0.056	0.082
babies	-0.032	-0.021	-0.013	0.01	-0.00023	0.018	0.02	0.018	0.024	1	-0.0089	-0.0075	-0.0066	0.083	0.036	0.019	-0.011	0.029	0.037	0.098
is_repeated_guest	-0.085	-0.12	0.01	-0.03	-0.0061	-0.087	-0.097	-0.15	-0.033	-0.0089	1	0.082	0.42	0.012	0.032	-0.24	-0.022	-0.13	0.077	0.013
previous_cancellations	0.11	0.086	-0.12	0.036	-0.027	-0.013	-0.014	-0.0067	-0.025	-0.0075	0.082	1	0.15	-0.027	-0.012	-0.18	0.0059	-0.066	-0.018	-0.048
previous_bookings_not_canceled	-0.057	-0.074	0.029	-0.021	-0.0003	-0.043	-0.049	-0.11	-0.021	-0.0066	0.42	0.15	1	0.012	0.023	-0.21	-0.0094	-0.072	0.048	0.038
booking_changes	-0.14	0.00015	0.031	0.0055	0.011	0.063	0.096	-0.052	0.049	0.083	0.012	-0.027	0.012	1	0.067	0.12	-0.012	0.02	0.066	0.053
agent	-0.083	-0.07	0.063	-0.031	0.0015	0.14	0.18	-0.036	0.041	0.036	0.032	-0.012	0.023	0.067	1	0.35	-0.055	-0.025	0.18	0.034
company	-0.021	0.15	0.26	-0.077	0.045	0.067	0.18	0.21	0.031	0.019	-0.24	-0.18	-0.21	0.12	0.35	1	0.00041	0.086	-0.013	-0.099
days_in_waiting_list	0.054	0.17	-0.056	0.023	0.023	-0.054	-0.002	-0.0083	-0.033	-0.011	-0.022	0.0059	-0.0094	-0.012	-0.055	0.00041	1	-0.041	-0.031	-0.083
adr	0.048	-0.063	0.2	0.076	0.03	0.049	0.065	0.23	0.32	0.029	-0.13	-0.066	-0.072	0.02	-0.025	0.086	-0.041	1	0.057	0.17
required_car_parking_spaces	-0.2	-0.12	-0.014	0.0019	0.0087	-0.019	-0.025	0.015	0.056	0.037	0.077	-0.018	0.048	0.066	0.18	-0.013	-0.031	0.057	1	0.083
total_of_special_requests	-0.23	-0.096	0.11	0.026	0.0031	0.073	0.068	0.12	0.082	0.098	0.013	-0.048	0.038	0.053	0.034	-0.099	-0.083	0.17	0.083	1

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. It involves

➢ **Replacing NULL/MISSING Values**

- Replacing categorical 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
- Replacing numerical missing values with **Mode**
  - No. of null values in [country] = 488



Data Cleaning

➢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.

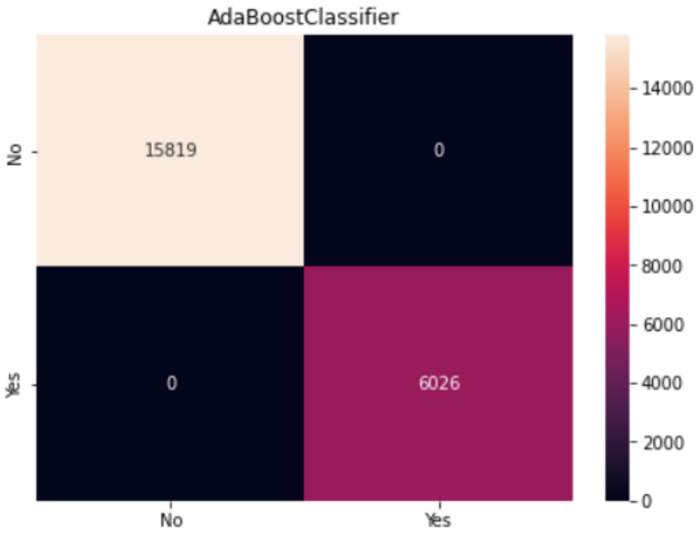
Categorical variables in our data set:

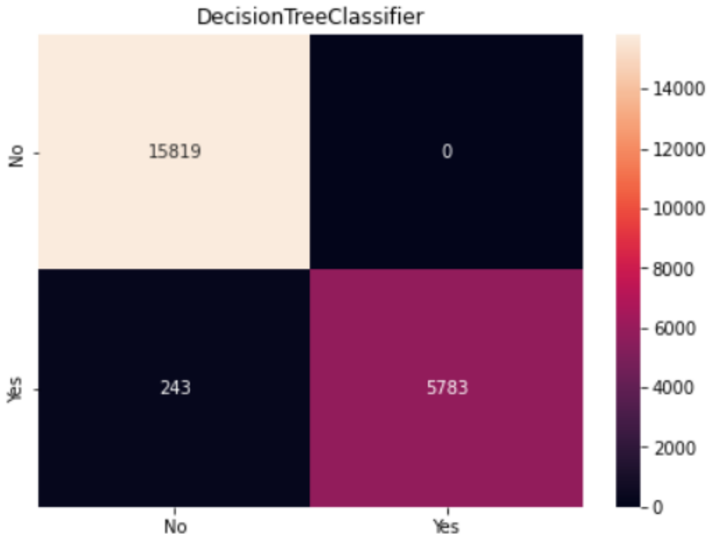
Hotel, arrival_date_month, mean, country, Market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, customer_type, reservation_status, reservation_status_date.

Training the Model

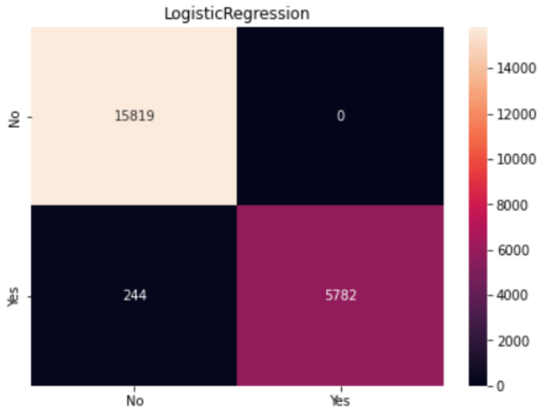
- Size of training data = 75%
- Size of testing data = 25%

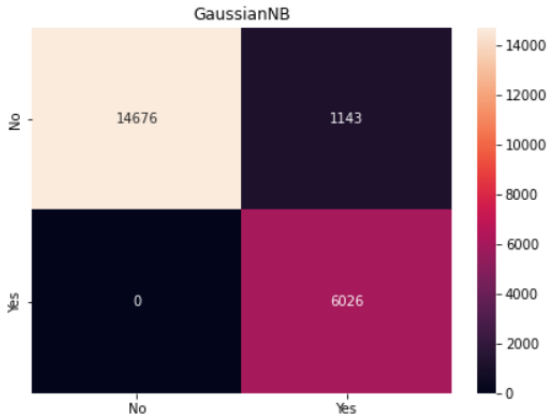
Confusion Matrix

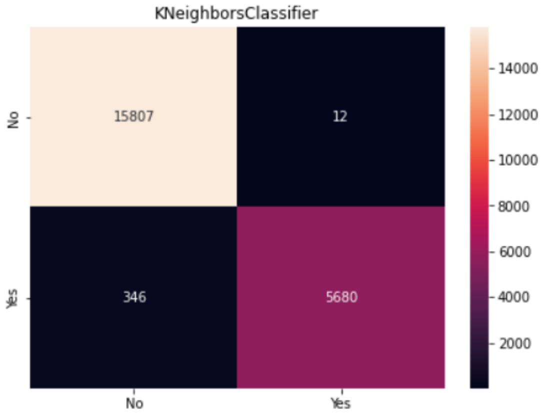


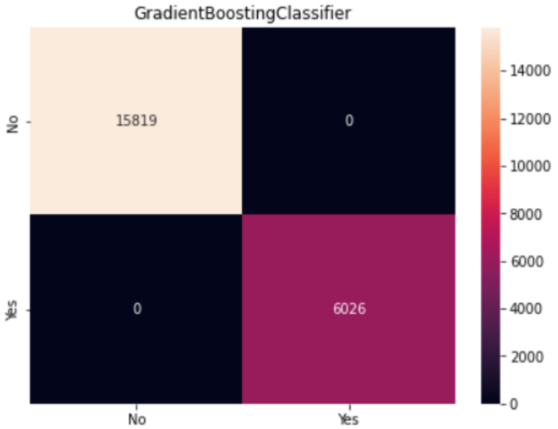


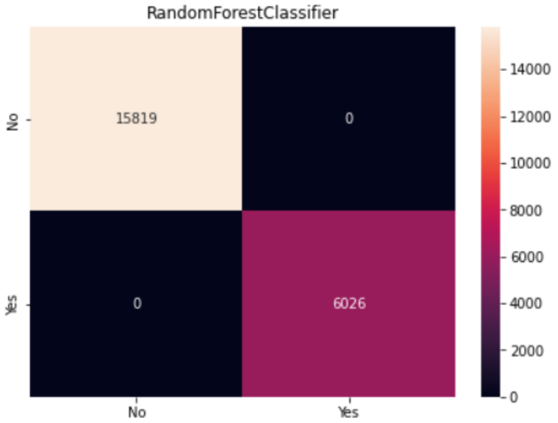
Confusion Matrix

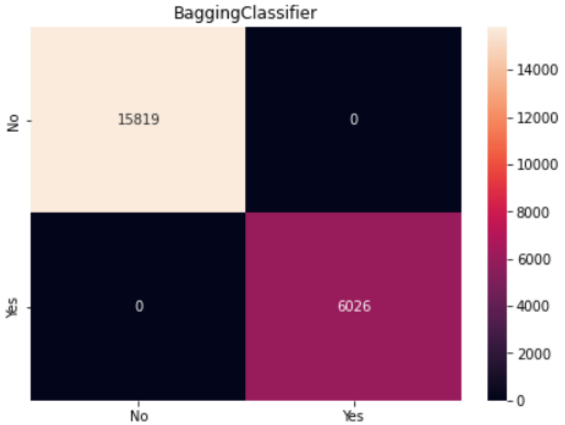












Accuracy using Various Classifiers

- LogisticRegression
  - Training Accuracy : 0.9882194958188366
  - Testing Accuracy : 0.9888303959716183
- Gaussian Naive Bayes
  - Training Accuracy : 0.9468656534212293
  - Testing Accuracy : 0.9487296864271
- K Neighbors Classifier
  - Training Accuracy : 0.9882957944210462
  - Testing Accuracy : 0.9831998168917372
- DecisionTreeClassifier
  - Training Accuracy : 0.9882347555392785
  - Testing Accuracy : 0.9888761730373083
- RandomForestClassifier
  - 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%**