Instrucciones Tarea 3.

Para esta tarea se utilizo el dataset llamado hotel_booking.csv disponible en Kaggle en https://www.kaggle.com/datasets/mojtaba142/hotel-booking.

Se define como la variable objetivo 'hotel' la cual es una variable binaria que puede tomar dos valores: 'Resort Hotel' o 'City Hotel'.

El objetivo de los modelos desarrollados es predecir a que tipo de hotel pertenece la reservación ingresada. La distribución de dicha variable es 66% para 'City Hotel' y 33% para 'Resort Hotel' la cual se considera aceptable.

El dataset contiene 119.390 observaciones y 36 columnas con información recabada entre 2015 y 2017. La información personal disponible en el dataset se construyó artificialmente por lo cual para este estudio se decide no utilizar. Así mismo se decide no utilizar las observ aciones que tienen ausente el país de procedencia de la reservación (488) y todos los demá s valores nulos se decide inputarlos con 0 ya que dada la naturaleza de las variables ('childr en', 'agent', 'company') el 0 representa la ausencia de uno de estos elementos lo cual se int erpreta como adecuado. Esto deja el numero total de observaciones del dataset en 118.902.

Se presenta a continuación un resumen de las variables del dataset:

summary	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
count	119390	119390	119390	119390	119390	119390
mean	null	0.37041628277075134	104.01141636652986	2016.156554150264	null	27.16517296255968
stddev	null	0.4829182265925997	106.86309704798795	0.7074759445202426	null	13.605138355497651
min	City Hotel	0	0	2015	April	1
25%	null	0.0	18	2016.0	null	16
50%	null	0.0	69	2016.0	null	28
75%	null	1.0	160	2017.0	null	38
max	Resort Hotel	1	737	2017	September	53

summary	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children
count	119390	119390	119390	119390	119386
mean	15.798241058715135	0.9275986263506156	2.500301532791691	1.8564033838679956	0.10388990333874994
stddev	8.780829470578345	0.9986134945978777	1.908285615047907	0.5792609988327554	0.39856144478644145
min	1	0	0	0	0.0
25%	8	0	1	2	0.0
50%	16	1	2	2	0.0
75%	23	2	3	2	0.0
max	31	19	50	55	10.0

summary	babies	meal	country	market_segment	distribution_channel	is_repeated_guest	previous_cancellations
count	119390	119390	118902	119390	119390	119390	119390
mean	0.007948739425412514	null	null	null	null	0.03191222045397437	0.08711784906608594
stddev	0.09743619130126 43	null	null	null	null	0.1756714541065 66	0.8443363841545122
min	0.0	BB	ABW	Aviation	Corporate	0	0
25%	0.0	null	null	null	null	0.0	0
50%	0.0	null	null	null	null	0.0	0
75%	0.0	null	null	null	null	0.0	0
max	10.0	Undefined	ZWE	Undefined	Undefined	1	26

summary	previous_bookings_not_canceled	reserved_room_type	assigned_room_type	booking_changes	deposit_type
count	119390	119390	119390	119390	119390
mean	0.13709690928888515	null	null	0.22112404724013737	null
stddev	1.497436847707677	null	null	0.6523055726747712	null
min	0	A	A	0	No Deposit
25%	0	null	null	0	null
50%	0	null	null	0	null
75%	0	null	null	0	null
max	72	P	P	21	Refundable

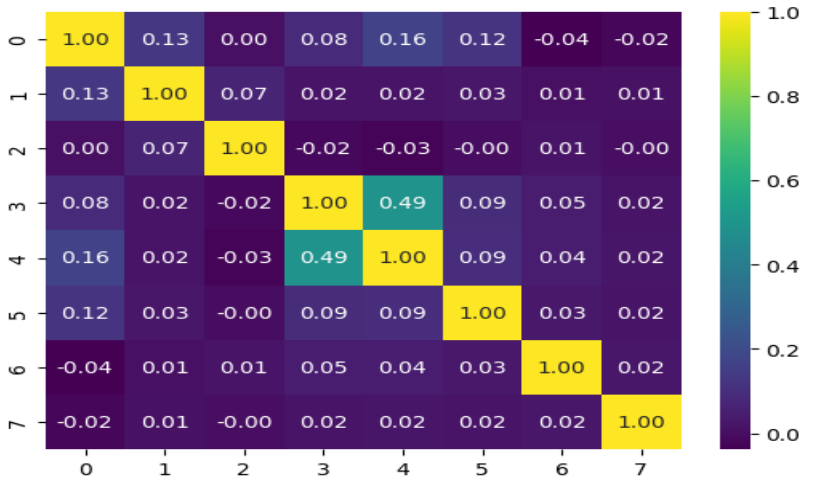
summary	days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_requests
count	119390	119390	119390	119390	119390
mean	2.321149174972778	null	101.83112154663662	0.06251779881062065	0.5713627607002262
stddev	17.594720878776197	null	50.53579031886051	0.2452911474674937	0.7927984228094128
min	0	Contract	-6.38	0	0
25%	0	null	69.22	0	0
50%	0	null	94.5	0	0
75%	0	null	126.0	0	1
max	391	Transient-Party	5400.0	8	5

summary	reservation_status
count	119390
mean	null
stddev	null
min	Canceled
25%	null
50%	null
75%	null
max	No-Show

Distribución de la variable objetivo ‘Hotel’:

hotel	count	porcentage
Resort Hotel	40060	33.55389898651479
City Hotel	79330	66.44610101348522

Se crea la matriz de correlación para las variables numéricas:



Se crean dos modelos de clasificación binaria (regresión logística y árbol de decisión) los cuales generan los siguientes resultados para el área bajo la curva ROC.

```
Area bajo la curva ROC para el modelo de Regresion Logistica: 0.8796280609020212
Area bajo la curva ROC para el modelo de Arbol de Decision: 0.623795902419058
```

Como puede observarse el área bajo la curva ROC (AUC) es más alta para el modelo de regresión logística que para el árbol de decisión lo cual implica que este modelo hace un mejor trabajo discriminando entre positivos verdaderos y falsos positivos para cada umbral de decisión. Por tanto el modelo de regresión logística es mejor.