Análisis de Predicción Reservación/Cancelación Hoteles

Universidad Cenfotec

Data Analytics & Big Data

Grettel Rojas Conejo

Contenido

Justificación del Análisis

La industria hotelera es una de las más importantes en el mundo, y en la actualidad, con la situación de la pandemia ocasionada por virus Covid19 la hotelería ha venido en picada, a tal punto que muchas compañías han cerrado, generando a su paso desempleo e incertidumbre en aquellas ciudades dedicadas a la atención de turistas la mayor parte del año.

Con esta situación, se considera por lo tanto de gran valor para la industria generar un modelo que permita mediante el análisis de un conjunto de variables propias de una reserva de hotel, predecir si las personas cancelarán o no dicha reserva, con el fin de que el negocio pueda preparase y tomar acciones proactivas, en caso de existir alto riesgo de cancelación según las variables que se determinen.

Descripción de Data Set

El Data set que fue seleccionado para este análisis es parte del repositorio de Kaggle, el mismo describe dos conjuntos de datos con registros de demanda dos hoteles.

Uno de los hoteles (H1) es un complejo hotelero y el otro es un hotel urbano (H2). Ambos conjuntos de datos comparten la misma estructura, los datos del Hotel H1 constan de con 31 variables que describen las 40,060 observaciones y 79,330 observaciones para el caso del Hotel H2. Cada observación representa una reserva de hotel.

Ambos conjuntos de datos comprenden las reservas que llegarán entre el 1 de julio de 2015 y el 31 de agosto de 2017, incluidas las reservas que efectivamente llegaron y las que se cancelaron. Como se trata de datos reales del hotel, se eliminaron todos los elementos de datos relacionados con la identificación del hotel o del cliente.

Características del Data Set

En la siguiente tabla se muestran las características de las diferentes variables con las que se cuenta para la realización de este análisis:

Variable	Type	Description
ADR	Numeric	Tarifa diaria promedio
Adults	Integer	Número de Adultos
Agent	Categorical	Identificación de la agencia de viajes que realizó la reserva
ArrivalDateDayOfMonth	Integer	Día del mes de la fecha de llegada
ArrivalDateMonth	Categorical	Mes de la fecha de llegada con 12 categorías
ArrivalDateWeekNumber	Integer	Número de semana de la fecha de llegada.
ArrivalDateYear	Integer	Año de llegada
AssignedRoomType	Categorical	Código para el tipo de habitación asignada a la reserva. A veces, el tipo de habitación asignada difiere del tipo de habitación reservada debido a razones de operación del hotel
Babies	Integer	Número de bebés

BookingChanges	Integer	Número de cambios / modificaciones realizados en la reserva desde el momento en que la reserva se ingresó hasta el momento del check-in o cancelación
Children	Integer	Número de niños
Company	Categorical	Identificación de la empresa / entidad que realizó la reserva o responsable de pagar la reserva.
Country	Categorical	País de origen. Las categorías se representan en el formato ISO 3155–3: 2013
CustomerType	Categorical	Tipo de reserva, asumiendo una de cuatro categorías: **Contrato**: cuando la reserva tiene un contrato asociado. **Grupo**: cuando la reserva está asociada a un grupo. **Transitoria**: cuando la reserva no forma parte de un grupo o contrato y no está asociada a otra reserva transitoria. **Parte transitoria**: cuando la reserva es transitoria, pero está asociada al menos a otra reserva transitoria.
DaysInWaitingList	Integer	Número de días que la reserva estuvo en la lista de espera antes de ser confirmada al cliente
DepositType	Categorical	Indicación de si el cliente realizó un depósito para garantizar la reserva. Esta variable puede asumir tres categorías: **Sin depósito**: no se realizó ningún depósito **Sin reembolso**: se realizó un depósito por el valor del costo total de la estadía. **Reembolsable**: se realizó un depósito con un valor por debajo del costo total de la estadía.
DistributionChannel	Categorical	Canal de distribución de reservas. El término "TA" significa "Agentes de viajes" y "TO" significa "Operadores turísticos"
IsCanceled	Categorical	Valor que indica si la reserva se canceló (1) o no (0)
IsRepeatedGuest	Categorical	Valor que indica si el nombre de la reserva era de un huésped repetido (1) o no (0)
LeadTime	Integer	Número de días transcurridos entre la fecha de entrada de la reserva y la fecha de llegada
MarketSegment	Categorical	Designación del segmento de mercado. En categorías, el término "TA" significa "Agentes de viajes" y "TO" significa "Operadores turísticos"
Meal	Categorical	Tipo de comida reservada: Indefinido / SC: sin paquete de comida. BB - Alojamiento y desayuno. HB - Media pensión (desayuno y otra comida, generalmente cena). FB - Pensión completa (desayuno, almuerzo y cena)
PreviousBookingsNotCanceled	Integer	Número de reservas anteriores no canceladas por el cliente antes de la reserva actual
PreviousCancellations	Integer	Número de reservas anteriores que el cliente canceló antes de la reserva actual
RequiredCardParkingSpaces	Integer	Número de plazas de aparcamiento requeridas por el cliente.
ReservationStatus	Categorical	Último estado de la reserva, asumiendo una de tres categorías: Cancelado: la reserva fue cancelada por el cliente; Salida: el cliente se ha registrado pero ya se fue; No-Show: el cliente no se registró e informó al hotel del motivo
ReservationStatusDate	Date	Fecha en la que se estableció el último estado. Esta variable se puede usar

		junto con el Estado de reserva para comprender cuándo se canceló la reserva o cuándo el cliente realizó el check-out del hotel
ReservedRoomType	Categorical	Código de tipo de habitación reservada
StaysInWeekendNights	Integer	Número de noches de fin de semana (sábado o domingo) que el huésped se hospedó o reservó para quedarse en el hotel
StaysInWeekNights	Integer	Número de noches semanales (de lunes a viernes) que el huésped se hospedó o reservó para hospedarse en el hotel
TotalOfSpecialRequests	Integer	Número de solicitudes especiales realizadas por el cliente (por ejemplo, cama doble o piso alto)

Preprocesamiento de los Datos

Con el fin disminuir los tiempos de procesamiento de entrenamiento de los modelos y optimizar los resultados fue requerido realizar varias tareas de preprocesamiento:

• Eliminación de Variables

Se analizan las variables contenidas en el data set con el fin de eliminar aquellas que por criterio de experto, se sabe que no agregarán valor a los resultados de predicción. Se eliminan las siguientes:

- arrival_date_year
- arrival_date_week_number
- arrival_date_day_of_month
- reservation_status_date
- reservation_status

• Sustitución de NA por Cero

En las variables “Country” y “Children” se sustituyen los valores definidos como NA por el valor cero.

• Sustitución de Null por Cero

En las variables “Agent” y “Company” se sustituyen los valores definidos como Null por el valor cero.

• Cambios en Tipos de Variables

Para poder continuar con el análisis, se analizaron los tipos de variable que debían cambiarse. Con esto, las siguientes variables son convertidas a tipo “Category”:

- hotel
- is_canceled
- arrival_date_month
- meal
- country
- market_segment

- distribution_channel
- is_repeated_guest
- reserved_room_type
- assigned_room_type
- deposit_type
- agent
- company
- customer_type

Análisis de correlación

Se realizó un análisis de correlación para determinar si podría depurarse aún más el set de datos, sin embargo se llegó a la conclusión que **no existe una correlación significativa** entre ninguna de las variables, siendo el valor mayor de correlación de 0,17. En la siguiente imagen se muestran algunas de las interacciones encontradas:

	lead_time	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	previous_cancellations
lead_time	1.000000	0.085671	0.165799	0.119519	-0.037613	-0.020915	0.086042
stays_in_weekend_nights	0.085671	1.000000	0.498969	0.091871	0.045794	0.018483	-0.012775
stays_in_week_nights	0.165799	0.498969	1.000000	0.092976	0.044203	0.020191	-0.013992
adults	0.119519	0.091871	0.092976	1.000000	0.030440	0.018146	-0.006738
children	-0.037613	0.045794	0.044203	0.030440	1.000000	0.024030	-0.024729
babies	-0.020915	0.018483	0.020191	0.018146	0.024030	1.000000	-0.007501
previous_cancellations	0.086042	-0.012775	-0.013992	-0.006738	-0.024729	-0.007501	1.000000
previous_bookings_not_canceled	-0.073548	-0.042715	-0.048743	-0.107983	-0.021072	-0.006550	0.152728
booking_changes	0.000149	0.063281	0.096209	-0.051673	0.048952	0.083440	-0.026993
days_in_waiting_list	0.170084	-0.054151	-0.002020	-0.008283	-0.033271	-0.010621	0.005929
adr	-0.063077	0.049342	0.065237	0.230641	0.324853	0.029186	-0.065646
required_car_parking_spaces	-0.116451	-0.018554	-0.024859	0.014785	0.056255	0.037383	-0.018492
total_of_special_requests	-0.095712	0.072671	0.068192	0.122884	0.081736	0.097889	-0.048384

Tareas de Entrenamiento

Se utiliza un valor de aleatorización 42. Para la predicción de la variable **is_canceled**, se procede a entrenar los algoritmos de clasificación, con el fin de evaluar los resultados de calidad de la predicción.

El dataset a entrenar se dividió en datos de Train y de Test. A continuación se detallan los porcentajes y cantidad de registros para cada subset. Considerando un 80% para training y un 30% para testing.

Una vez realizado el preprocesamiento y la partición de los datos, se procede a entrenar el set de datos de train, con cada uno de los algoritmos que fueron seleccionados. En el siguiente cuadro se muestran los resultados obtenidos de los parámetros de calidad en el set de datos de test:

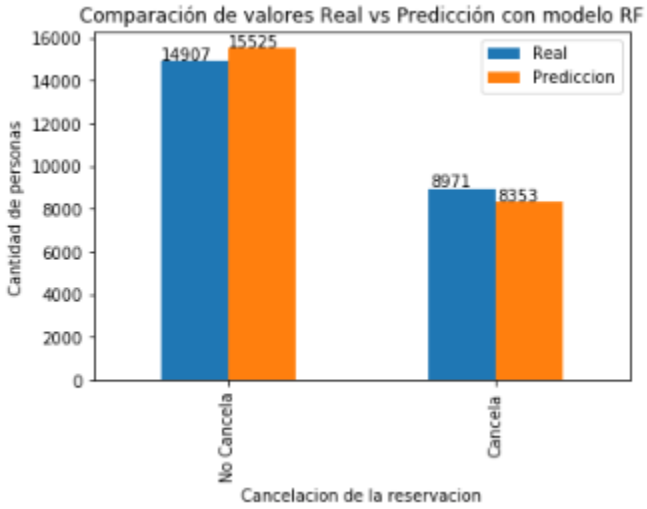
Modelo	Accuracy	F1
Logistic Regression	0,7675	0,6533
Random Forest Classifier	**0,8957**	**0,8563**
Gradient Boosting Classifier	0,8468	0,7825
K Neighbors Classifier	0,8046	0,7310
Gaussian Naïve Bayes	0,5567	0,6099

Escogencia del modelo

En la tabla anterior, con respecto a los parámetros de calidad obtenidos con los datos de test, se determina que el modelo que obtuvo mejores resultados de Accuracy y F1 es el generado con el algoritmo RFC (Random Forest Classifier), por lo que se considera como el más óptimo a utilizar para predecir la cantidad de reservaciones que serán canceladas.

Resultados

En la siguiente tabla se muestra una comparación entre los datos reales del set de datos seleccionado y los resultados de la predicción.



Como se observa los valores son bastante cercanos, ya que los parámetros de calidad antes indicados del modelo son altos. Se obtiene el resultado de que la mayoría de los clientes no cancelará la reserva, en total un 65% de los casos, y un 35% reservará pero cancelará la reserva.

Este resultado es de gran valor para la compañía hotelera, ya que se podrían tomar acciones proactivas para el ahorro de costos, considerando que en un 35% de las veces el cliente no llegará al hotel, por ejemplo acciones en el ahorro de inversión en parqueos, reservas de alimentación y en general planificación de los hospedajes.

También es importante tomar en cuenta que, las variables que se presentan en este análisis son factores “comunes” que se pueden consideran influyan en temas de demanda de la industria

hotelera, podrían existir factores ocasionales o de desastres (por ejemplo la aparición de una nueva enfermedad como es el caso de la situación actual) que no se encuentren considerados y que por lo tanto las predicciones futuras.