MÁSTER EN DATA SCIENCE Y FINANZAS

Práctica FINAL, Hoteles: EDA y Selección Variables

Fecha: 23 de enero de 2020

Autor: Lucía Saiz Lapique

Profesor: Ricardo Queralt

Índice:

ABSTRACT

El sector del turismo es uno de los más grandes de hoy en día, afectando a miles de empresas y personas, que pertenecen a diversos tipos de industrias. Una de las industrias más afectadas por el flujo del turismo y la demanda de clientes por viajar es la industria hotelera, que, anual y temporalmente, reciben personas de todas partes del mundo con requisitos y experiencias totalmente diferentes unas de otras. La intención, con el siguiente informe, es averiguar una forma de, con todas esas experiencias personales de los clientes de un hotel en Portugal, entre 2015 y 2017, identificar qué factores determinan que una persona decida reservar en un hotel antes que en otro y si decidirá cancelar la reserva antes de llegar o no.

INTRODUCCIÓN

En el sector hotelero, una gran causa de pérdidas y necesidad de ajuste de stock son las cancelaciones que los clientes realizan previas a su visita, tanto con mucha antelación como de última hora. En este informe se verá cómo, con una serie de datos recopilados de las reservas de un hotel en Portugal, se relacionan entre sí los clientes y cuáles son las expectativas que puede tener el hotel para las próximas temporadas, con el objetivo de poder analizar las probabilidades de cancelación de cada tipo de cliente y segmentarlos para evitar pérdidas. En primer lugar, se hará un análisis exploratorio de los datos (EDA), donde se podrá ver la relación y dependencia de una variables con otras, al igual que una breve descripción de los estadísticos básicos de los datos que estudiaremos. Después se estudiarán las variables de la base de datos, en primer lugar, trabajando y seleccionando las variables relevantes para ambos estudios. Una vez estén seleccionadas y tratadas esas variables comunes, se estudiarán por separado las variables más específicas de la parte de segmentación y las de la parte de predicción que se analizarán en trabajos más adelante. Por último, se presentará una conclusión de lo trabajado hasta el momento, al igual que una breve introducción a lo que serán los próximos estudios.

OBJETIVO

El objetivo de este estudio, previo a los estudios de segmentación y de predicción que se quieren realizar, es conseguir simplificar, limpiar y analizar la base de datos con la que se trabajará de ahora en adelante, para así facilitar el uso y aplicación de los datos a los modelos de agrupación

y predicción que se aplicarán más adelante. Se intentará eliminar el mayor número posible de variables, manteniendo la mayor cantidad de información posible, y agrupar o sintetizar la información de estos datos de forma que nos aporten, única y exclusivamente, información relevante para los estudios.

ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

La base de datos original con la que se va a trabajar a lo largo de estos estudios es una de dos bases de datos que tratan la demanda de hoteles en Portugal. H1 (la base de datos utilizada en este trabajo) trata los datos de demanda de un hotel de tipo resort, que contiene 31 variables y 40060 observaciones en total, cada cual representando una reserva en el hotel. Estas reservas rondan entre julio de 2015 y agosto de 2017, por lo que existen un poco más de dos años de datos, teniendo en cuenta que el único año completo (con todas las temporadas) sobre el que existen datos, es 2016. (Antonio, de Almeida y Nunes, 2019)

A la hora de comenzar el estudio de los datos, nada más empezar, se decidió hacer un pequeño resumen que muestre las características básicas sobre ellos, como su estructura (si son de tipo numérico o categórico), sus histogramas, sus estadísticos generales, si contienen valores nulos, etc. Se puede ver que, de las 31 variables, 12 de ellas son de tipo carácter, es decir, que se tendrán que tratar como variables categóricas; una variable de tipo fecha, con lo cual se deberá tener en cuenta aparte y con otro formato; y el resto (18 variables) son de tipo numérico. Estas son las más interesantes durante el EDA, ya que son de las que se puede hacer el análisis de correlaciones y calcular los estadísticos básicos. Se puede ver también que no hay ningún valor nulo en toda la base de datos, lo cual facilitará el trabajo a la hora de tratar y seleccionar las variables, además de que hay varias de las variables numéricas que son dicotómicas (ceros y unos), así que al final esas tendrán que ser tratadas como categóricas.

A continuación, se crea una tabla en donde se comparen una serie de estadísticos de las variables numéricas (media, mediana, valores máximos y mínimos y cuantiles), que se pueden ver a continuación en la figura 1:

Como se presenta en la tabla más claramente, algunas de las variables son de tipo dicotómico (*ISCanceled*, *IsRepeatedGuest*), y se tendrán que considerarlas categóricas a la hora de usarlas en el estudio de segmentación y el de predicción. Hay variables que tienen muchos ceros, como

el número de bebés o niños por reserva, el número de plazas de parking reservadas, o la cantidad de cambios que se hicieron a la reserva previos a la llegada del cliente.

Figura 1: Tabla de estadísticos generales de las variables

Variable	Min	Q1	Med	Mean	Q3	Max
IsCanceled	0.0	0,0	0,0	0.3	1,0	1,0
LeadTime	0.0	10,0	57,0	92.7	155,0	737,0
ArrivalDateYear	2015.0	2.016	2.016	2016.1	2.017	2.017
ArrivalDateWeekNumber	1.0	16,0	28,0	27.1	38,0	53,0
ArrivalDateDayOfMonth	1.0	8,0	16,0	15.8	24,0	31,0
StaysInWeekendNights	0.0	0,0	1,0	1.2	2,0	19,0
StaysInWeekNights	0.0	1,0	3,0	3.1	5,0	50,0
Adults	0.0	2,0	2,0	1.9	2,0	55,0
Children	0.0	0,0	0,0	0.1	0,0	10,0
Babies	0.0	0,0	0,0	0.0	0,0	2,0
IsRepeatedGuest	0.0	0,0	0,0	0.0	0,0	1,0
PreviousCancellations	0.0	0,0	0,0	0.1	0,0	26,0
PreviousBookingsNotCanceled	0.0	0,0	0,0	0.1	0,0	30,0
BookingChanges	0.0	0,0	0,0	0.3	0,0	17,0
DaysInWaitingList	0.0	0,0	0,0	0.5	0,0	185,0
ADR	-6,4	50,0	75,0	95.0	125,0	508,0
RequiredCarParkingSpaces	0.0	0,0	0,0	0.1	0,0	8,0
TotalOfSpecialRequests	0.0	0,0	0,0	0.6	1,0	5,0

El único valor negativo que hay en la base de datos a estudiar se encuentra en la variable ADR, que señaliza el precio diario de cada reserva, lo cual significa que, o es un outlier, o en su momento se introdujo mal en la base de datos. De cualquier forma, como no existen más datos negativos, se seguirá tratando esa variable con normalidad. A priori, se podría mencionar, sin indagar más, que las variables no tienen mucho que ver unas con otras, ya que todos los datos de cada una son totalmente distintos, algunos significando precio, cantidades, fechas, etc. Habrá que estudiar cómo tratarlas más adelante.

Mientras tanto, para ver si es posible tener en cuenta o descartar alguna variable según lo dependiente o relacionada que esté con las demás, el siguiente paso es estudiar la correlación de todas las variables numéricas. El resultado es que, según lo predicho, no existe ningún tipo de correlación entre las variables. La única correlación considerante (y es solo del 0.4) existe entre la variable que indica la semana del año que se llegó al hotel y el año, lo cual no aporta ningún tipo de información adicional, y luego también una relación del 0.4 entre el número de noches que se queda el cliente entre semana y las noches que se queda en fin de semana.

Tras este análisis exploratorio de los datos, se decide tratar las variables en base al conocimiento del negocio que poseemos, ya que no hay correlaciones significantes entre ninguna

variable y el estudio superficial que hemos realizado de ellas no nos ha aportado información suficiente para seleccionar o eliminar ninguna de ellas.

OPERACIONES COMUNES A LOS DOS ESTUDIOS

Tras un previo estudio del negocio utilizando diversos informes y artículos, decidimos comenzar a analizar las variables de la base de datos una a una, para decidir cuales son las que serían eliminadas, cuales se conservarían, si se hiciera alguna modificación o combinación de variables, y si hay alguna variable que deba ser específica de la parte del estudio que corresponde a segmentación o a agrupación.

En primer lugar, se considera la variable *MarketSegment*. Sabemos que es categórica, y tras estudiar cuales son los valores únicos que la componen, se decidió que era suficientemente relevante para dejarla tal cual estaba y utilizarla en ambos estudios. A su vez se analizó también la variable *DistributionChannel*, que vimos que nos aportaba información muy parecida a *MarketSegment,* pero menos completa; tiene menos valores únicos, con lo cual especifica menos, y los que sí que tiene, coinciden en todos los casos con *MarketSegment*, así que se decidió que lo debido era eliminarla de la base de datos directamente.

A continuación, nos centramos en que las variables *Agent* y *Company* también parecen tener algo que ver, no solo entre ellas, sino con *MarketSegment* también. Se hizo un pequeño resumen de las tres donde era evidente que entre ellas si que se parecen, pero que, con relación a *MarketSegment*, no tenían nada que ver, así que no se tendría más en cuenta en este estudio. Al analizarlas pudimos ver que ambas variables aportan el código de la agencia o empresa que hizo la reserva por la persona que quiere ir al hotel, de lo cual sacamos una conclusión: no es relevante el código de la agencia o compañía, pues todos son distintos y no van a crear ningún tipo de dependencia o relación a los clientes. Sin embargo, sí que interesa el hecho de que una persona haya decidido, o no, hacer una reserva de forma individual o a través de un "intermediario". Por ello, se decidió fusionar estas dos variables en una sola (*MotivoViaje*) donde, en vez de incluir el código, se incluyen cuatro valores únicos: que se haga la reserva a través de una agencia, a través de una empresa, con ambos, o de forma independiente. De esta forma, se eliminan dos variables y se crea una nueva con información más relevante a los estudios.

La siguiente decisión que se tomó con respecto a las variables es sobre las cuatro variables que explican la fecha en la que llega el cliente al hotel (*ArrivalDateYear*, *ArrivalDateMonth*, *ArrivalDateDayOfMonth* y *ArrivalDateWeekNumber*). Se decidió que, combinando las cuatro variables en una sola, es posible definir al fecha de llegada directamente y sin tener que usar tantos datos en el estudio. Se combinaron las cuatro, guardándolas en formato de fecha (año-mes-día) y se almacenaron en una variable general llamada *ArrivalDate*, eliminando las cuatro originales. La base de datos ya ha sido reducida a 26 variables.

La siguiente variable que fue analizada es la variable *BookingChanges*, para la cual, tras ver que los datos que aportaba eran cantidades de cambios en la reserva por cliente, se decidió hacer una serie de cálculos para ver si compensaba quedárnosla. Primero, se calculó cuántos clientes verdaderamente hacían algún cambio en su reserva y se descubrió que estos casos solo formaban el 20% de la base de datos. Fue considerado, además, cuántos de esos casos habían cancelado su reserva a lo largo de los tres años (ya que es una de las variables más importantes) y vimos que no llegaba ni al 15% de todos ellos; es decir, que apenas el 3% de todos los casos que hacían una reserva en el hotel y luego hacían algún cambio, acababan cancelando su reserva. Finalmente, decidimos que los datos de esta variable no aportaban ninguna información adicional o relevante y se decidió eliminarla.

Las siguientes cinco variables que se estudiaron decidimos quedárnoslas tal cual vienen en la base de datos. Estas variables son *ADR, Country, DepositType, CustomerType* y *IsRepeatedGuest*, ya que vimos que cada una aportaba información importante sobre los clientes que no viene dada por ninguna otra variable y que no puede ser alterada de ninguna forma, ya que es relevante en el formato original en el que viene indicada. Se generaron una serie de dudas con la variable Country, la nacionalidad del cliente, pues descubrimos con el estudio de negocio que, en el caso de esta base de datos, si un cliente no llegaba al hotel y no había indicado su nacionalidad, se le adjudicaba que era portugués en los datos. Finalmente, se comprobó que los portugueses no llegaban a componer el 50% de las cancelaciones totales, así que se decidió mantener la variable original.

Parecida a la variable de si el cliente repite experiencia o no (*IsRepeatedGuest*), se vio que existían dos variables (*PreviousBookingsNotCanceled* y *PreviousCancellations*) que podían estar relacionadas entre ellas y con la anterior. Se estudiaron juntas y se vio que, efectivamente, estaban relacionadas, pero que esas dos y *IsRepeatedGuest* aportaban información distinta, así

que esta se dejó de lado. Al comparar las dos variables con una serie de cálculos, se sacaron las siguientes conclusiones: que existen 2031 casos en los que el cliente repite experiencia en el hotel y nunca cancela; que hay 1095 casos en los que el cliente repite experiencia y sí que cancela; que hay 212 casos que se solapan y dicen que el cliente repite y cancela o no cancela. Esto puede ser por un error de la base de datos o porque en alguna ocasión haya cancelado y en otra no; en cualquier caso, son casos conflictivos, así que se decidió combinar las dos variables en una llamada *Repetir* y eliminar los 212 casos que se solapaban. De paso, se decidió eliminar también la variable *ReservationStatusDate*, ya que se decidió que no aportaba ninguna información adicional.

Las siguientes variables que se estudiaron son, a la vez, *StaysInWeekendNights* y *StaysInWeekNights*, ya que indicaban prácticamente el mismo tipo de datos y podían estar relacionadas de alguna forma. Se realizó un recuento de la cantidad de clientes que se quedaban en el hotel entre semana (13658 reservas), la cantidad de reservas que había solo los fines de semana (2276) y la cantidad que había entre semana y en fin de semana en la misma reserva (23531). Como vimos que las tres cantidades eran bastante influyentes y que importa por igual tanto el hecho de quedarse entre semana o en fin de semana como el número de noches que se quede cada cliente, se decidió, a partir de estas dos variables, eliminarlas y crear dos nuevas. En una de las variables, se tendría en cuenta el número de noches total (tanto en fin de semana como entre semana) de cada reserva (*TotalDias*). En la otra variable, se tendría en cuenta si era fin de semana, entre semana o ambas (*Findes*), para ver si eso también pudiera tener algún tipo de influencia en la probabilidad de cancelación de los clientes y cómo se podrían agrupar.

Una variable que nos pareció importante tener en cuenta a la hora de realizar nuestros estudios más adelante era la variable edad. Sin embargo, en la base de datos no se incluye ninguna variable que hable de la edad específica de los clientes; lo único con lo que contaba es el número de adultos, niños y bebés que hay en cada reserva. Tras varios intentos por crear una variable que pudiera relacionar las reservas a la edad del cliente, se llegó a la conclusión de que no era posible, pues había casos en los que había varios adultos, niños y bebés en la misma reserva, y no había forma de calcular la edad de todos en un mismo campo. Sin embargo, estábamos de acuerdo con que las tres variables por separado nos aportaban información poco relevante, así que se concluyó que lo mejor sería juntar las tres en una variable que nos calculase el número de personas total por reserva (*TotalPersonas*).

Por último, durante el estudio de negocio realizado previo al análisis de las variables, aprendimos que una variable muy importante a la hora de analizar la demanda de un hotel era la anticipación con la que los clientes hacen la reserva. Para poder calcular esto, vimos que había dos variables (*LeadTime* y *DaysInWaitingList*) que, al sumarlas, proporcionaban el número de días en total con los que el cliente había hecho la reserva. De esta forma, se eliminaron esas dos variables y se creó una nueva variable llamada *Anticipación* donde se introdujo la suma de esas dos. Finalmente, y previo al análisis de las variables específicas del estudio de segmentación y el de predicción, contamos con un total de 13 variables.

OPERACIONES ESPECÍFICAS DE SEGMENTACIÓN

Para el estudio que se realizará de segmentación, hubo una serie de variables que consideramos relevantes específicamente para él, mientras que otras eran innecesarias. En el caso de las variables que fueron seleccionadas para el estudio de agrupación, tuvimos en cuenta el problema de negocio planteado, que será expuesto en otro trabajo más adelante. Se decidió seleccionar cuatro variables más. Una de ellas era la variable *Meal*, a la cual, el único cambio que se le hizo fue combinar los campos que contuviesen SC o Undefined como uno mismo que significase que no habían solicitado ningún servicio de comida.

La segunda variable que se consideró para este estudio fue *RequiredCarParkingSpaces*, pero en vez de mantener la variable como venía originalmente en la base de datos (número de plazas solicitadas en la reserva), que contenía muchos outliers y la información era poco relevante, decidimos considerar el hecho de si el cliente había solicitado plaza de garaje o no, con lo cual convertimos la variable en dicotómica, llenando el 14% de los campos con "Sí" y el resto con "No". Hicimos lo mismo con la variable *TotalOfSpecialRequests*, ya que no era relevante el número de extras que pidiesen los clientes, pero sí el hecho de haberlos pedido.

Por último, la cuarta variable que seleccionamos para el estudio de segmentación de los datos es *ReservationStatus*. Se generaron ciertas dudas con esta variable, ya que aporta prácticamente la misma información que la variable *IsCanceled*, (variable dependiente del estudio de predicción), que es muy relevante en el estudio y no podíamos eliminar; con una excepción. Esta variable es más específica que *IsCanceled*, ya que además de informar sobre el número de clientes que cancelan la reserva, informa también sobre quiénes no cancelaron, pero tampoco

aparecieron el día de llegada. Al no poder eliminar la variable anterior, pero por miedo a perder información relevante que aporta *ReservationStatus*, se decidió dejar una variable para un estudio y la otra para el otro.

OPERACIONES ESPECÍFICAS DE SEGMENTACIÓN

En la selección de variables para el trabajo que se realizará más adelante en predicción, se tuvieron en cuenta las características de los modelos que se deben ejecutar para el objetivo de cada trabajo, que serán expuestos más adelante. Como ha sido explicado en el apartado anterior, una de las dos variables que fueron seleccionadas para el estudio de predicción fue la variable de cancelaciones de reservas a lo largo de los tres años. Esta es la más importante del estudio que se realizará, pues es la variable dependiente sobre la cual se ejecutarán los modelos predictivos de las principales causas de cancelación de reserva en el hotel.

La segunda variable seleccionada fue combinación de dos variables: *ReservedRoomType* y *AssignedRoomType*. Se consideró que una razón de cancelación en un futuro podría ser que, un cliente, al llegar al hotel, reciba una habitación distinta a la que ha pedido. Esto podría generar descontento en el cliente y que, o no vuelva nunca (que también es negativo para el hotel) o que tenga más probabilidades de cancelar una reserva en el futuro. Por ello, se decidió, a partir de estas dos, crear una variable dicotómica donde incluyésemos dos tipos de campos: si el cliente ha recibido la habitación que había solicitado (un 81% de los clientes) o no.

CONCLUSIÓN:

Finalmente, nuestra base de datos incluye 19 variables: "IsCanceled", "Meal", "Country", "MarketSegment", "IsRepeatedGuest", "DepositType", "CustomerType", "ADR", "RequiredCarParkingSpaces", "TotalOfSpecialRequests", "ReservationStatus", "MotivoViaje", "ArrivalDate", "DiferenciasReserva", "Repetir", "Findes", "TotalDias", "TotalPersonas" y "Anticipacion_reserva". Estas son las variables que se han considerado relevantes para estudiar los siguientes dos temas.

A continuación, realizaremos tres estudios distintos; uno de ellos sobre la segmentación de clientes de acuerdo con una serie de características comunes; dos sobre modelos predictivos.

Uno de los modelos predictivos estará enfocado a la probabilidad de cancelación de un cliente en base a una serie de variables influyentes que se estudiarán. El otro estudio que se realizará estará enfocado en una serie temporal que analice los datos a lo largo del tiempo, para predecir como cambiará la demanda en los próximos años. El objetivo en los tres estudios es obtener los resultados óptimos, gracias al análisis y selección de variables realizado en este informe.

BIBLIOGRAFÍA:

R Core Team, (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Vitta, A. (2020). *Segmentación de la demanda hotelera - Les Hoteliers*. Les Hoteliers: Revenue Management y Marketing Hotelero. URL: https://www.leshoteliers.com/segmentacion-demanda-hotelera/

Vitta, A. (2020). *Cómo optimizar los precios de un hotel - Les Hoteliers*. Les Hoteliers: Revenue Management y Marketing Hotelero. URL: https://www.leshoteliers.com/optimizar-estrategia-precios-hotel/

Correia, A., Moital, M., Oliveira, N., Ferreira Da Costa, C. (2009). *Multidimensional segmentation of gastronomic tourists based on motivation and satisfaction.* Int. J. Tourism Policy, Vol. 2, Nos. 1/2, 2009.

López, E.D., Guzmán-Sala, A. (2018). *ANÁLISIS DE LA OFERTA Y PROMOCIONES EN EL SECTOR HOTELERO: EL CASO TABASCO EN MÉXICO*. International Journal of Scientific Management and Tourism (2018) 4-2: 367-389,

Antonio, N., de Almeida, A. Nunes, L. (2017). *Predicting Hotel Bookings Cancellation With a Machine Learning Classification Model*. 16th IEEE International Conference on Machine Learning and Applications. Lisboa, Portugal.

Antonio, N., de Almeida, A. Nunes, L. (2017). *Data Article - Hotel booking demand datasets*. Published by: ELSEVIER. journal homepage: www.elsevier.com/locate/dib. Pp. 41-49.