MÁSTER EN DATA SCIENCE Y FINANZAS

AGRUPACIÓN: Práctica FINAL, Hoteles

Fecha: 22 de enero de 2020

Autor: Lucía Saiz Lapique

Profesor: Juan Manuel López Zafra

Índice:

ABSTRACT

¿Es posible seleccionar a los clientes del sector hotelero que supongan mayor coste a la hora de cancelar una reserva y así evitar pérdidas? Con el estudio a continuación, se verá como 17 variables que explican la demanda de un hotel en Portugal durante dos años, pueden agrupar los clientes de dicho hotel en esas fechas, teniendo en cuenta los gastos de cada cliente y las probabilidades que tienen de cancelar una reserva. Si es posible averiguar qué clientes, que suponen un alto beneficio para el hotel, van a cancelar su reserva antes de llegar, se podría hacer un esfuerzo por reducir la cantidad de pérdidas asociadas a esas cancelaciones.

INTRODUCCIÓN

Partiendo de la base de que cada hotel recibe diversos tipos de perfil de cliente a lo largo del año y que cada perfil tiene unos gastos y una forma de viajar distinta, sería interesante poder identificar los diferentes perfiles que existen y qué características les definen.

En este informe se realizará un estudio de segmentación en el que se intentará agrupar los diferentes clientes que recibe un hotel en Portugal, a lo largo de los tres años que componen nuestros datos. En el informe anterior, se explicó la selección de variables que se ha realizado, teniendo en cuenta las intenciones que existían con este estudio de segmentación. Tras ese análisis exploratorio y de selección de variables, la base de datos final para este trabajo cuenta con 17 variables y se seguirá la siguiente estructura: se presentará el objetivo del trabajo y la cuestión planteada con la que se ha decidido avanzar; después se explicará el planteamiento del trabajo y directamente se comenzará a segmentar los datos; en primer lugar, calculando las distancias entre los datos para poder formar los clusters y, finalmente, analizando los segmentos resultantes.

OBJETIVO: PROBLEMA PLANTEADO

El objetivo de este estudio es conseguir segmentar o agrupar los clientes que recibe el hotel desde 2015 hasta 2017, teniendo en cuenta una serie de variables que, durante el estudio previo a este, consideramos relevantes para la realización del trabajo.

La duda que surgió al plantear el trabajo y de lo que nos dimos cuenta al empezar el estudio de las variables fue que no todas las cancelaciones afectaban al hotel de la misma forma. Se consideró que los clientes cuyos gastos en el hotel fuesen mayores, es decir, las reservas que aportasen más beneficios al hotel, serían las que causarían mayor número de pérdidas en caso de cancelación. Por ello, se consideró que una buena forma de segmentar los clientes sería en base a los gastos adicionales o los costes superiores que pudieran suponer. Si era posible identificar a los clientes que, al cancelar una reserva, supondrían más problemas para el hotel y su probabilidad de cancelación, se podrían implementar medidas para evitar pérdidas mayores.

Para ayudar a realizar una segmentación basada en esa teoría se decidió, en el informe anterior, considerar como relevantes las variables que supusieran mayores gastos para el cliente, como solicitar plazas de parking, extras en las habitaciones y comidas. Estas serían utilizadas específicamente para el estudio de agrupación y no para la parte de predicción, ya que son factores específicos que el cliente pide aparte y por los que paga un precio adicional o superior, y que para los objetivos de los siguientes trabajos son irrelevantes.

PLANTEAMIENTO DEL TRABAJO: MUESTREO

Inicialmente, al comenzar el estudio de segmentación de nuestra base de datos, la intención era utilizar la base de datos completa, pues sería la forma de obtener los datos más realistas posibles. Como la mayoría de las variables son categóricas, si se utilizara la distancia euclídea o de Manhattan (que solo pueden hacerse con variables numéricas), se perdería mucha información relevante para el estudio. Por ello, se decidió que la mejor opción para estudiar la distancia entre los datos sería aplicando las distancias de Gower.

Sin embargo, esta función es muy compleja y, al lanzar directamente la base de datos completa para calcular las distancias, se generó un error indicando que el ordenador no tenía capacidad suficiente para hacer el estudio. Cuando vimos que se tendría que reducir la base de datos para poder llevar a cabo el ejercicio, comenzamos a seleccionar cuál sería la mejor muestra que se podría escoger. Tras informarnos y la consideración varias opciones, se decidió que la mejor opción era hacer un muestreo estratificado (dividir la población total en subgrupos) y coger como muestra el año 2016 solo, ya que tanto 2015 como 2017 estaban compuestos por datos que cubrían solo medio año (de julio a diciembre en el caso de 2015 y de enero a agosto en el caso de 2017) y no podrían informar sobre todas las temporadas. Con los datos ya seleccionados, se puso en marcha el método de Gower para buscar las distancias entre las observaciones.

DISTANCIAS DE GOWER

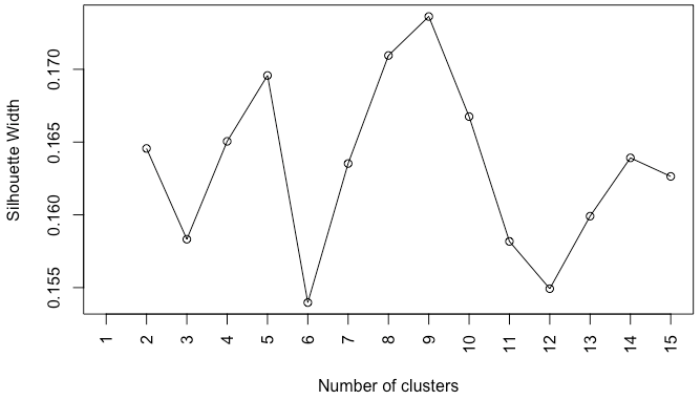
Las distancias de Gower son muy útiles porque permiten comparar variables mixtas, es decir, numéricas con categóricas. "Este método consiste en calcular todas las desemejanzas por pares (distancias) entre las observaciones en el conjunto de datos. Las variables originales pueden ser de tipos mixtos. […] La "distancia de Gower" se elige mediante "gower" métrico o automáticamente si algunas columnas de x no son numéricas. También conocido como el coeficiente de Gower (1971), expresado como una diferencia, esto implica que se aplicará una estandarización particular a cada variable, y la "distancia" entre dos unidades es la suma de todas las distancias específicas de la variable." (Montoya, Ortiz, Escobar, Galindo y Ochoa, 2018)

Con la muestra reducida, se consiguió que este proceso funcionase y fue posible crear una matriz con las distancias entre las variables. Esto es relevante, no solo para poder hacer una segmentación apropiada de los datos, sino para, previo a ello, elegir el número óptimo de clusters que se debían crear para que los datos se agrupasen de la manera más delimitada posible. Para

la mayor optimización posible, se llegó a la conclusión de que la mejor forma de dividir los datos era con el mayor número de clusters posible, así que se ejecutó el programa para averiguar cuál era el número óptimo de clusters entre 2 y 15. El resultado fue el que se ve en la figura 1.

Para analizar cuál es el número óptimo de clusters, se debía identificar cuáles eran los puntos más altos de la gráfica. En la figura 1, se pueden ver varios puntos que muestran altos niveles de optimización, pero está claro que el más alto se encuentra en el valor 9. Por tanto, se decidió que ese era el número de clusters que debíamos buscar para segmentar de la mejor manera posible la base de datos, que se haría a continuación.

Figura 1: Número óptimo de clusters

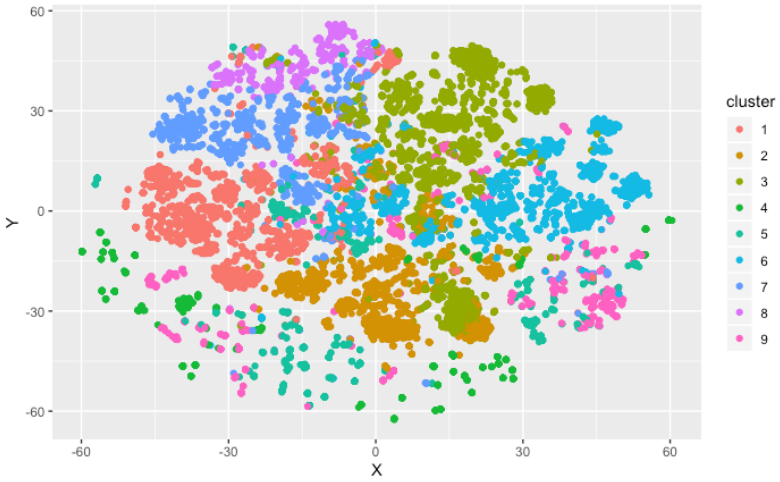


CLUSTERS

El primer paso a continuación es, directamente, dividir la muestra de los datos en 9 grupos distintos. En primera lugar, se llevó a cabo esta división mediante la función de RStudio 'pam', cuyas siglas significan *Partitioning Around Medoids* y que se refiere a la agrupación por el método K-*medoids*, una versión parecida, aunque es más robusta, al K-*means*. Este método se considera más robusto que el estudio por K-*means* ya que minimiza el número de diferencias generales emparejadas en lugar de las distancias euclídeas. Es una técnica que necesita conocer el número de clusters a crear a priori (que es nuestro caso), y es capaz de clasificar variables mixtas.

Con PAM, obtenemos un resumen de los datos estadísticos principales que componen cada una de las variables en los 9 clusters. Sin embargo, la información está explicada con demasiados datos para poder identificar las características de todos los grupos una a una y muy poco visual. Por ello, se decidió analizar los grupos de forma distinta. Con los datos obtenidos del *PAM* y las distancias de *Gower* (que se obtuvieron con la función *Daisy*), le aplicamos un gráfico de tipo t-sne para que nos muestre los resultados de los grupos, por colores, según el grupo al que pertenezcan, y lo distinguidos que están unos de otros. Los resultados que obtenemos son los siguientes, que podemos observar en la figura 2:

Figura 2: Clusters



Del gráfico, se pueden sacar una serie de conclusiones:

- Existen tres clusters (el verde, 4; el turquesa, 5; y el rosa, 9) que no están muy bien explicados, demasiado esparcidos y se pisan con otros clusters del gráfico; por tanto, como conclusión, decidimos que no se deben tener en cuenta esas agrupaciones a la hora de tomar las decisiones finales, o deben ser consideradas como menos relevantes.
- El resto de clusters están bastante bien diferenciados ya que, a pesar de algún caso en ciertos grupos, la gran mayoría están agrupados con el resto de los casos de su grupo (color), y es fácil identificar donde acaban unos y empiezan otros.
- Aunque en general parezca que no están muy bien explicados los grupos, pues se mezclan entre ellos en algunos casos y el centro en general es un conjunto o mezcla de casi todos los grupos, la segmentación es posible.

Nos habría gustado poder haber graficado los grupos con la función eclust, que marca de forma más clara los grupos, pero se descubrió que esa función usa distancias euclídeas y variables numéricas, con lo cual se perdería demasiada información necesaria para la clasificación de los grupos.

SEGMENTACIÓN FINAL

Para finalizar con los clusters, se decidió crear una tabla que incluyese las características básicas de cada grupo generado con el estudio, teniendo en cuenta la moda para todos las variables categóricas y la media y moda para las numéricas, dependiendo de cada caso.

A continuación, en la figura 3, vemos el resumen de las características principales de los grupos.

Figura 3: Tabla de características de grupo.

Cluster	ADR	Noches	Personas	Anticipacion	Meal	EstadoReserv	Country	MarketSegment	DepositType	CustomerType	MotivoViaje	Repetir	Findes
1	89.38	1	2	38.51	BB	Check-Out	PRT	Online TA	No Deposit	Transient	Agent	NoRepiten	SoloEntreSemana
2	139.90	7	2	102.73	BB	Canceled	PRT	Online TA	No Deposit	Transient	Agent	NoRepiten	Ambos
3	95.72	7	2	92.94	BB	Check-Out	PRT	Online TA	No Deposit	Transient	Agent	NoRepiten	Ambos
4	74.16	3	2	173.38	HB	Canceled	PRT	Groups	Non Refund	Transient	Agent	NoRepiten	Ambos
5	69.88	2	2	102.67	BB	Check-Out	PRT	Groups	No Deposit	Transient-Party	Agent	NoRepiten	SoloEntreSemana
6	71.33	7	2	119.99	BB	Check-Out	GBR	Offline TA/TO	No Deposit	Transient	Agent	NoRepiten	Ambos
7	80.54	1	2	13.29	BB	Check-Out	PRT	Direct	No Deposit	Transient	Independent	NoRepiten	SoloEntreSemana
8	45.52	1	1	7.55	BB	Check-Out	PRT	Corporate	No Deposit	Transient	Company	RepitenNoCancelan	SoloEntreSemana
9	69.04	4	2	214.32	BB	Check-Out	GBR	Groups	No Deposit	Transient-Party	Agent	NoRepiten	Ambos

En primer lugar, cabe mencionar que en esta tabla no están todas las variables que se han utilizado para hacer el estudio de segmentación. Esto se debe a que, hay una serie de variables que, al tener en cuenta sus características, se concluyó que no aportaban ninguna información sobre la posible segmentación de las reservas. Estas son *IsRepeatedGuest* (si el cliente ha repetido, que en todos los casos el resultado era 0, ya que se había escogido la moda para hacer esta característica) y *RequiredCarParkingSpaces* (si el cliente solicitó plaza para aparcar o no, y en todos los casos la respuesta era negativa, es decir, que ningún cliente se definía por eso).

Una vez aclarado esto, se comienzan a analizar los diferentes grupos. En base a las conclusiones que se sacaron del gráfico de clusters, se debe tener en cuenta que cualquier conclusión sobre los grupos 4, 5 y 9 no deben ser definitivas en el estudio, pues son los grupos que peor explicados por estas características están. Lo primero que se desea analizar es la cuestión planteada al principio del estudio, es decir, si hay alguna segmentación específica de los grupos que supongan más beneficios para el hotel, y si tienen más probabilidades de cancelar una reserva. Para comprobar los gastos de los clientes, nos fijamos primero en el ADR, que representa su gasto medio por reserva.

Al ver en el ADR, es evidente que los tres grupos que más gastan y que, por tanto, más beneficio le suponen al hotel, son el 1, el 2 y el 3. Las características por mencionar de estos son que, como era de esperar, pasan muchas noches en el hotel y tienen el valor máximo de la media de cantidad de personas en hacer la reserva de todos los grupos. De los tres grupos, el primero tiene una moda de 1 noche por reserva y es el que menos antelación tiene para pedir la reserva de media, mientras que el grupo 2 y 3 tienen una media de alrededor de 100 días para hacer la reserva. En el caso de los tres grupos, todos son principalmente portugueses, han hecho la reserva a través de una agencia de viajes, con estancia de tipo temporal, no repiten experiencia y se quedan todos entre semana y, en el caso del grupo 2 y 3, también el fin de semana.

De aquí se puede deducir que el grupo 1 está probablemente formado por gente que viaja por trabajo y con poca antelación, que se quedan entre semana y no repiten experiencia. El grupo 2 y 3, en cambio, son probablemente grupos que se van de vacaciones, ya que las reservas son de muy larga duración y se hacen con más antelación. Una cosa que no hemos mencionado aun y de gran relevancia es que, el grupo con los costes más altos (el 2) tiene más probabilidades de cancelar que los demás. Esto supone dificultades para el hotel, y era el aspecto principal en el que nos queríamos enfocar en nuestro estudio.

Para clasificar el resto de los grupos en base al problema planteado, se decidió mirar directamente si hay algún cluster más que tenga probabilidades de cancelar una reserva o no. Se puede ver que la única otra probabilidad de cancelación está en el cluster 4, que hemos identificado como grupo poco fiable, así que tampoco deberíamos centrarnos mucho en ello. Sí que es verdad que es el único grupo que pide dos comidas en el hotel, lo cual supondría más pérdidas de normal que otros, pero en general no es un grupo que aporte muchos más beneficios que otros al hotel; está en la media en cuanto al ADR.

Para finalizar esta parte de segmentación hablaremos del penúltimo grupo, el 8, que es el grupo con el valor mínimo de costes. El grupo 8 es el grupo que tiene el menor ADR de todos, teniendo la menor moda de noches y gente en total también. Son personas que hacen la reserva a través de una empresa y los únicos que repiten y no cancelan la reserva nunca; podemos deducir que estos también son viajes de trabajo, pero distintos a los anteriores. En este caso, la anticipación es muy poca pero los precios muy bajos, lo cual indica que sea la empresa la que organice y pague el viaje, buscando economizar al máximo este tipo de actividades. Además, el hecho de que repitan y nunca cancelen puede ser indicador de que las empresas hayan establecido algún tipo de lealtad con el hotel, de forma que el hotel asegura tener esas habitaciones siempre llenas y las empresas pagan menos por ellas. Estos grupos, en general, suponen poco riesgo de cancelación para el hotel, lo cual también es bueno saber.

El resto de clusters no muestran ningún campo o característica específica de cada uno, que indique que puedan suponer más beneficios o costes para el hotel en caso de cancelación y se encuentran todos más o menos en los rangos de media generales.

CONCLUSIÓN:

La cuestión o problema planteados al principio del estudio era si es posible identificar a los clientes que traen más rentabilidad al hotel y cuál es su porcentaje de cancelación, para poder tratar de evitar pérdidas mayores.

Una vez recopilada toda la información y habiendo segmentado los datos hasta crear 9 grupos distintos, vemos que los tres grupos que se han identificado como más beneficiarios son el 1, 2 y 3, siendo el 2 el que más rentabilidad aporta y el que, desgraciadamente, mayor probabilidad de cancelación tiene.

Una vez identificado al grupo 2 y las características que lo definen, otros hoteles podrían hacer este estudio a sus clientes para identificar este perfil e intentar crear algún tipo de prevención o alternativa para no sufrir tantas pérdidas en caso de que una de estas reservas se cancele.

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/.

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.

Montoya, A., Ortiz, J., Escobar, P., Galindo, J. Ochoa, D. (2018). *Segmentación de una Base de Datos (clusters)*. Publicado por: RPubs.com. URL: http://rpubs.com/andrestoya/391125

E.M. Mirkes (2011). *K-means and K-medoids applet*. University of Leicester. URL: http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

Autor anónimo. *INTRODUCCIÓN AL ANÁLISIS CLUSTER*. Publicado por: CEACES. URL: https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm

Shadan, M. (2017). *Mean, Median, Mode - Measures of Central Tendency*. Publicado por: Rpubs.com. URL: https://rpubs.com/mohammadshadan/meanmedianmode

Ramirez-Alan, O. (2016) *Muestreo Aleatorio Simple (MAS)*. Publicado por: RPubs.com. URL: https://rpubs.com/osoramirez/159490

Cuvero, Y. (2017) *Taller de Muestreo*. Publicado por: RPubs.com. URL: https://rpubs.com/yandiraccv/Muestreo_Clase1

Gil Bellosta, C. (2018). *KAMILA: CLÚSTERING CON VARIABLES CATEGÓRICAS*. Publicado por: Datanalytics.com. URL: https://www.datanalytics.com/2018/07/20/kamila-clustering-con-variables-categoricas/

Andrienko, S. (2019). *Project2*. Publicado por: RPubs.com. URL: https://rpubs.com/sofiandra/477664