CUNEF
COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS

FINANCIAL [THINKING]

Sistema de Reservas de Hotel Informe Técnicas de Agrupación y Reducción de la Dimensión

Mayra Goicochea Neyra

CONTENIDO

ABSTRACT

El sector de alojamiento es una de las industrias más ricas en datos, dado que reciben enormes volúmenes de información, a alta velocidad, con mucha variedad, veracidad y variabilidad.
Estas propiedades hacen que el análisis de datos en la industria hotelera sea complejo. Satisfacer las expectativas de los clientes es un factor clave en la industria hotelera para captar la lealtad de los clientes. Con ese objetivo, los encargados del área de marketing y los analistas de datos buscan activamente formas para utilizar esta información de la mejor manera posible y avanzar con soluciones analíticas óptimas de datos, como identificar método de clustering para la segmentación del mercado y desarrollar un sistema de recomendación (para servicios y beneficios a sus clientes de forma personalizada).[1,2]
En el presente informe, se presenta una solución de técnicas de agrupación aplicada a la información de un hotel de Algarve (Portugal), donde se aplican técnicas de análisis factorial y árbol clasificador con la finalidad de identificar los atributos más relevantes, y clustering para la agrupación del segmento de mercado.

INTRODUCCIÓN

En los últimos años, el sector de alojamiento se ha solidificado como uno de los más rentables y competitivos debido al aumento del turismo no solo local (dentro de su país) sino al turismo global, dado las nuevas regulaciones de globalización de los gobiernos en muchos países. [4, 5]

Otro de los factores importantes para su crecimiento fue el Internet, no solo porque permite conectarse a usuarios de todo el mundo, sino que guarda grandes volúmenes de información de los usuarios. Cada vez hay más empresas, como sitios web de comercio electrónico y tiendas en línea que ofrecen recomendaciones de productos [6,7] para dirigirse a clientes potenciales. Esta tendencia de proporcionar recomendaciones, como ofertas y promociones personalizadas, a los clientes a través de diversos medios, como sitios web, redes sociales en línea, televisión y teléfonos inteligentes, aumenta día a día. Sin embargo, no es factible traducir estos sistemas de recomendación existentes a la industria hotelera debido a la gran escala de la red hotelera (es decir, clientes, proveedores y propietarios) y su estricta dependencia de las tendencias económicas mundiales. Además, la industria hotelera requiere un sistema de recomendación dinámico y automatizado que hace que muchas de las técnicas existentes que se centran en los sistemas de recomendación fuera de línea sean ineficaces. Para desarrollar una solución efectiva de recomendación del cliente para la industria hotelera, es necesario utilizar adecuadamente los enormes volúmenes de datos recopilados de los clientes.
En este informe, se muestra un prototipo basado en las técnicas de análisis factorial y clustering para segmentar los clientes, que aportara al area de marketing facilidad para aplicar estrategias adecuadas y maximizar la satisfacción del cliente (posible incremento en las ventas).

OBJETIVOS

El prototipo busca resolver la problemática del área de marketing y comercial, entregando mayor valor al negocio en la captación de nuevos clientes y campañas de fidelización:

- Detección de atributos relevantes
- Identificación de segmentos de usuarios para estrategias de marketing y venta.
- Identificación de posibles grupos claves de cancelaciones de reservas.

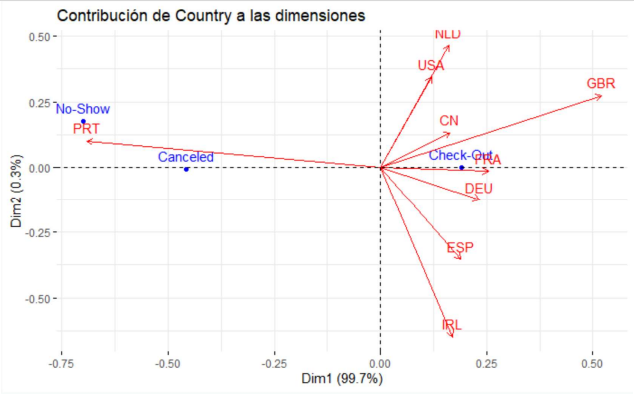
DETECCION DE ATRIBUTOS RELEVANTES

El conjunto de datos contiene 40060 observaciones de reservas, seleccionadas según su fecha de arribo, en el periodo del 1 de julio del 2015 al 31 de agosto del 2017. Cada reserva cuenta con información distribuida en 31 variables de carácter numérico y categórico, como se revisaron en el informe de operaciones

preliminares. Para el análisis de atributos relevantes, se inició con un análisis de correspondencias dado su eficiencia en la detección de similitudes entre variables categóricas. Lo usamos para revisar algunas características que se encontraron relevantes y relacionadas en el EDA. Como, por ejemplo, la variable "Country". La mayoría de reservas provienen de Portugal. Sin embargo, considero que esta información es subjetiva porque, al momento de realizar la reserva, algunos usuarios prefieren no indicar su verdadera residencia o lo ingresan por error, y en el hotel lo valida en el momento del arribo. (Ver Fig. 1 y Fig.2). El análisis de correspondencias, nos muestra que los huéspedes provenientes de Portugal son más probables de cancelar sus reservas (pero consideremos esta información como referencial dado que los casos de cancelaciones no tenemos la información validada de la procedencia)

Fig: 1.

Fig: 2.



En el análisis exploratorio del informe 1, se observó que la mayor tasa de reservas las tiene el segmento de mercado de los OTAs (como booking, expedia, entre otros) y que tenían mayor número de cancelaciones. El análisis de correspondencias nos muestra que hay cierta relación con las cancelaciones (Ver Fig.3) según la dimensión 1 que explica el 99% de la varianza (Ver Fig.4). En cuanto al segmento directo, está muy relacionado a que la reserva se efectué y no se cancele (checkout).

Fig: 3.

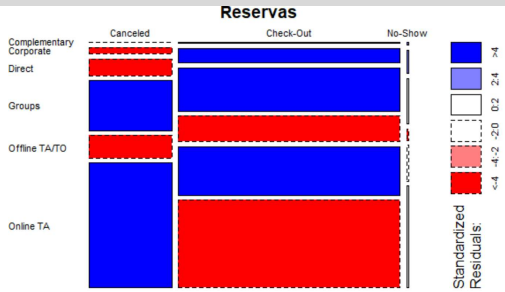
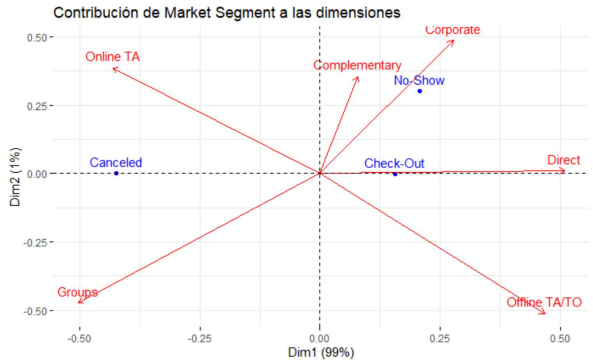


Fig: 4.



El análisis de correspondencia también nos muestra la que las reservas por OTAs son generalmente sin pago (Ver Fig. 5). Tal vez, este sea uno de los puntos relevantes que el área de marketing podrá aplicar para sus planes a futuro, de prever en este nicho de mercado más cancelaciones mediante otras estrategias.

Fig: 5.

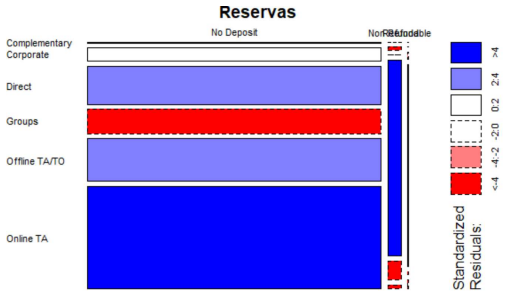
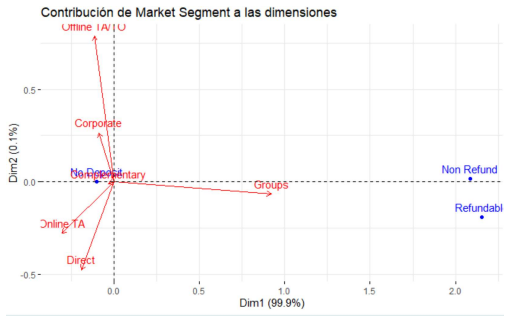
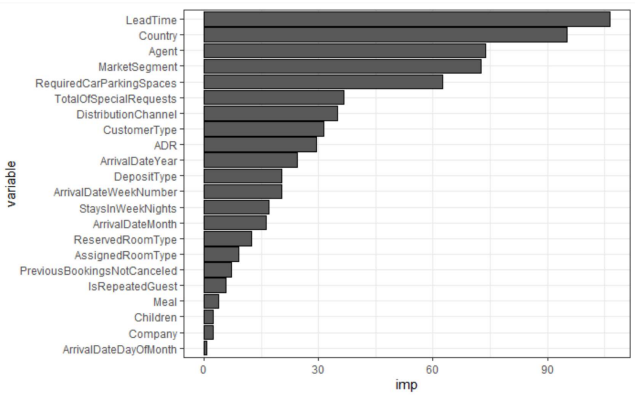


Fig: 6.



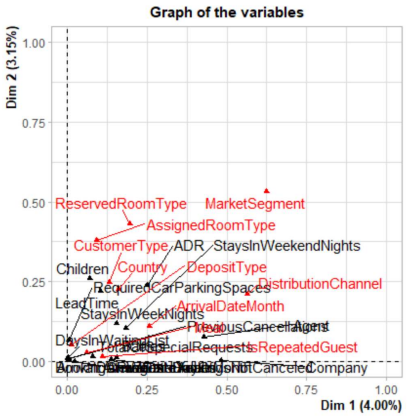
El análisis de correspondencias nos entrega gráficos y métricas basadas en el estudio de dos variables categóricas, pero el prototipo debería revisar el conjunto de las 31 variables. Es por ello que se evaluaron otras técnicas de la asignatura, resulto más adecuado un árbol clasificador. ACP (Análisis de componentes principales) es una técnica potente para reducción de dimensión, sin embargo, solo se puede aplicar en variables numéricas y el dataset cuenta con atributos relevantes de orden categórico, como lo demuestra el análisis de correspondencias. El árbol de clasificación (basada en la técnica de los árboles de decisiones) encontró que las variables más importantes son ***LeadTime***, ***Country***, ***Agent***, ***MarketSegment*** y ***RequiredCarParkingSpaces*** con índice mayor a 50% (Ver Fig. 7)

Fig: 7.



También existe un método de análisis factorial mixto. FAMD es un método de componentes principales dedicado a analizar un dataset que contiene variables cuantitativas y categóricas. Hace posible analizar la similitud entre los individuos considerando ambos tipos de datos. Es un algoritmo que combina el análisis de componentes principales (PCA) y el análisis de correspondencias múltiple (MCA) [3]. El resultado no fue eficiente, porque con dos dimensiones se puede explicar el 7.15% de la varianza, no es óptimo. (Ver Fig.8)

Fig: 8.



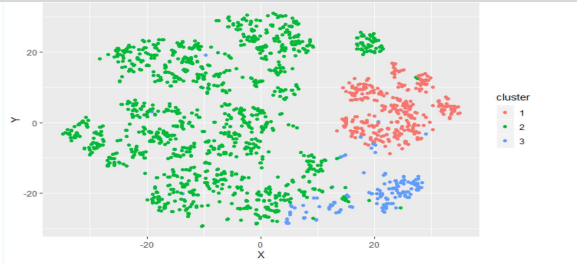
IDENTIFICACION DE GRUPOS (CLUSTERS)

Para el análisis cluster, se utilizó la distancia Gower por su robustez en data mixta (numérica y categórica). En resumen, Gower calcula las

diferencias parciales dependiendo del tipo de variable. Con Gower se evaluaron 4 métodos de clustering: Jerárquico Aglomerativo, Jerárquico Divisivo, y los no Jerárquicos PAM y Fuzzy. Se estimó que un número adecuado es entre 2 a 3 grupos según la anchura de la silueta.

Para seleccionar el método adecuado para el prototipo, se consideró que los grupos se discriminen bien y a una cantidad manejable de grupos para que el área de marketing pueda gestionarlos. El método elegido es el jerárquico aglomerativo, porque los clusters estimados discriminan bien y provee un numero óptimo de 3 grupos (Ver Fig. 9).

Fig: 9.



Los Métodos jerárquicos son los más eficientes a comparación de los no jerárquicos, porque las observaciones con más distancia Gower los clasificaron en dos grupos distintos a comparación de PAM y Fuzzy. (Ver Fig.10)

Fig: 10.

	hclust.clust <int>	IsCanceled <int>	LeadTime <int>	ArrivalDateYear <int>	ArrivalDateMonth <fctr>
1211	1	0	132	2017	May
744	3	0	0	2015	November

Finalmente, se encontró que el clúster 2 (con 31.77%), según el método jerárquico aglomerativo tiene más incidencia de cancelaciones. (Ver Fig.11)

Fig: 1.

```
"Cluster 1:  14.89 % Cancelaciones"
"Cluster 2:  31.77 % Cancelaciones"
"Cluster 3:  17.51 % Cancelaciones"
```

CONCLUSIONES

- El análisis de correspondencias es una técnica muy robusta para revisar las relaciones entre las variables categóricas. Las variables LeadTime, Country, Agent, MarketSegment y RequiredCarParkingSpaces son las más importantes de las 31 variables por que explica la varianza de la muestra eficientemente (según el árbol clasificador)

- La distancia Gower demostró que es muy eficiente para calcular la distancia de datos que están explicados en variables del tipo mixto.

REFERENCIAS BIBLIOGRAFICAS

- [1] Multivariate Analysis I, Practical Guide to Cluster Analysis in R, Alboukadel Kassambara.
- [2] Clustering, innovation an d hotel competitiveness: Evidence from the Colombia destination, Orietha Eva Rodriguez-Victoria, Francisco Puig, Migel Gonzalez-Loureio, publicado en International; Journal of Contemporary Hospitality Management (2017).
- [3] Multivariate Analysis II, Practical Guide to Principal Components Methods in R, Alboukadel Kassambara.
- [4] Tourism and sustainability: Development, globalization and new tourism in the third world, Mowforth, M., & Munt, I. (2015). Routledge
- [5] UNWTO (2010) Tourism and Poverty Alleviation, (link)
- [7] Integrating web mining and neural network for personalized e-commerce automatic service. Expert Systems with Applications, 37(4), 2898-2910. Chou, P. H., Li, P. H., Chen, K. K., & Wu, M. J. (2010).
- [6] Cluster analysis in marketing research: Review and suggestions for application. Journal of marketing research, Punj, G., & Stewart, D. W. (1983)

ANEXO

- Archivo RMarkdown en Github (MasterDS2019>Tecnicas de Agrupacion y Reduccion de la Dimension>Research>ARD-Examen20-Informe2MGN.Rmd).