UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining – 2022

Caso IDM-1

Integrantes:

Caccianini Antonela
Monzón Fernanda
Piccin Germán
Viola Jesica

Considere el siguiente caso:

Una cadena de hoteles que gestiona numerosos establecimientos en varios países registra información acerca de sus huéspedes en los diferentes hoteles. La gerencia desea implementar un **proyecto de Data Mining** a efectos de extraer el máximo provecho de esos datos. Más precisamente, la cadena desea conocer mejor a sus clientes de manera de poder informarles sobre eventos especiales, promociones especiales, etc., durante su estancia como así también después de esta.

A la llegada de un huésped, se registran sus **datos demográficos** y se le pide completar un **cuestionario** cuando se le hace entrega de una credencial que le permite ingresar a distintos lugares recreativos tales como piscinas, bares, etc. Esos **lugares** son gratuitos, pero vía la credencial es posible registrar cuando un huésped hace uso de alguno de esos servicios. Además, la credencial sirve como tarjeta de crédito para pagar ciertas bebidas como así también para pagar **productos** en las pequeñas tiendas que la cadena posee. Todas esas transacciones son registradas en una base de datos.

Por medio de su canal de TV privado, el hotel informa a sus huéspedes sobre los próximos eventos, actividades y promociones especiales, etc. Sin embargo, puesto que los huéspedes pasan la mayor parte de su tiempo afuera y no mirando TV, el hotel necesita un **sistema para enviar anuncios** altamente personalizados de manera de garantizar que sólo se envíen anuncios en los que el huésped esté posiblemente interesado. Además, el sistema de data mining también tiene que dar una **sugerencia razonable sobre qué anuncios enviar a**

un huésped particular ya durante los primeros días de su estadía, cuando el sistema tiene pocas posibilidades de aprender los hábitos de ese huésped particular. Después de la estadía, se le solicita al huésped que complete un **formulario de evaluación**. Este formulario contiene una lista de preguntas en las que se debe consignar un puntaje de 1 a 5. Para cada respuesta se pueden explicar los motivos o agregar comentarios adicionales breves.

Durante el invierno, la cadena envía publicaciones de una selección de sus hoteles a antiguos huéspedes, para obtener nuevas reservas en uno de sus hoteles. La selección para este **mailing personalizado** tiene que ser hecha de modo tal que sólo las publicaciones de los hoteles más interesantes para un huésped particular son enviadas a ese huésped. Por lo expresado, es importante considerar que en la mayoría de los casos un huésped de hotel no elige muchas veces exactamente el mismo hotel.

En conclusión, el sistema de data mining debe proporcionar información para:

1. **Decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante su estadía; un subproblema de este primer problema es decidir cuáles anuncios van a ser enviados durante los primeros días de la estadía de un huésped.**

2. **Decidir sobre la selección de hoteles para el mailing privado durante el verano.**

Recorrer los diferentes pasos del proceso KDD, descubrimiento del conocimiento a través de los datos, explicando cómo se aplican en este caso. Indicar para cada paso cuáles técnicas usaría y justificar su elección.

Al arribar un huésped a nuestro hotel entendemos que contamos con cierta información tal como: datos demográficos y los datos arrojados por una breve encuesta.
Entre los datos demográficos encontramos:

- Nombre y apellido
- D.N.I
- Fecha de nacimiento
- Edad
- Estado civil
- Género

- Mail
- País de residencia
- País de orígen
- Educación
- Empleo
- Número de habitación
- Fecha de llegada y fecha de partida
- Motivo de la visita

Deberán ser completados para cada huésped. Como ejemplo: si un grupo familiar se compone de padre, madre e hijo/a deberá completarse 3 veces.

Entre los datos consultados en la encuesta para el otorgamiento de la credencial:

- Nombre y apellido
- Fecha de nacimiento
- Número de habitación
- Fecha de llegada y fecha de partida
- ¿ Es la primera vez que nos visitas?
- ¿ Cómo nos conoció?

Luego de la estadía debe completar un formulario de evaluación:

- Número de habitación
- Fecha de llegada y fecha de partida
- ¿Es la primera vez que nos visitas?
- ¿Cómo calificas la atención en recepción? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la limpieza ? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la habitación? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la calidad de los desayunos? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la calidad del restaurante? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la calidad del bar? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Cómo calificas la calidad de la cafetería? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:

- ¿En general cómo calificas la experiencia en nuestro hotel? Siendo 1 poco satisfactoria y 5 muy satisfactoria
- Comentarios:
- ¿Recomendaría este hotel? Siendo 1 no lo recomendaría y 5 sí lo recomendaría
- Comentarios:
- ¿Cómo te enteraste de este hotel?
- Deja tu mail si te gustaría recibir novedades y descuentos (opcional)

Nuestro primer desafío es decidir cuáles anuncios de TV van a ser enviados a los huéspedes durante los primeros días de su estadía en nuestro hotel, no contando con información sobre sus preferencias y usos de las instalaciones.

Para ello seguimos los 7 pasos propuestos por KDD (knowledge discovery from data). Seguir estos pasos nos llevará a tener una idea más acertada de qué anuncio será enviado a cada habitación del hotel. Lo que se buscará es encontrar patrones en nuestras dos bases, la generada en la encuesta de datos demográficos y la generada en la encuesta previa al otorgamiento de la credencial.

1. Limpiar datos (para eliminar ruidos y datos inconsistentes)
2. Integración de datos (distintas fuentes de datos son combinadas)
3. Selección de datos (datos importantes para el análisis de la base de datos)
4. Transformación de datos (donde los datos son transformados, como sumas o cálculos matemáticos)
5. Data mining (proceso inteligente para extraer patrones)
6. Evaluación de patrones (identificar patrones representados por interesantes medidas)
7. Knowledge presentation (visualización del conocimiento obtenido)

Primer paso, ***limpieza de datos***.

Respecto a la primera problemática que se presenta, nos invita a controlar la coherencia y veracidad de los datos recogidos en las dos primeras encuestas. Ejemplo de ello, imaginemos que un huésped completa en fecha de nacimiento: 15/06/1888 y en edad 34 años. Es muy probable que se haya equivocado al escribir el año de nacimiento y ese error pueda ser subsanado para asegurarnos un correcto análisis.

En la segunda problemática, nos basaremos en los datos históricos del dataset, aplicaremos este primer paso por ejemplo, teniendo en cuenta a todas las personas que nos indiquen que su edad es mayor a 90 años y constatar cruzando datos con otras fuente si esa persona sigue viva y como consecuencia recibirá o no nuestro anuncio.

Segundo paso, ***integración de datos***.

Para el punto nro 1, en este paso identificamos con qué datos podemos vincular las dos tablas que tenemos inicialmente, a saber encuesta de datos demográficos encuesta para el otorgamiento de la credencial, ellos son número de habitación y fecha de entrada y salida. Con estás claves primarias nos aseguramos de vincular correctamente ambas tablas.

En cuanto a la problemática nro 2, y estando atentos al 2do paso del KDD, integraremos los datos actuales (producidos por los huéspedes de la temporada actual) con los históricos, lo que nos permitirá a posteriori analizar la evolución temporal de los datos solicitados, y si fuera necesario hacer una reducción o ampliación de inputs.

Tercer paso, ***selección de datos***.

Para la problemática nro 1, a priori, no dejamos fuera de análisis ningún campo.

En lo referente a la 2da problemática, seguramente encontremos datos no representativos de la realidad actual de nuestros clientes, dado que tenemos los datos históricos más los datos actuales, lo cual nos brindaría un panorama más amplio sobre la evolución de los patrones de comportamiento y consumo de nuestros huéspedes. Aquí seguramente encontremos que existen datos, que no sólo pueden ser tediosos de completar por los huéspedes, sino además innecesarios. Por otro lado, si suponemos que un cliente no visita el mismo hotel en temporadas consecutivas, podemos pensar: ¿de qué hoteles vinieron los clientes actuales al hotel bajo análisis particular?, ¿que valoraron negativamente y positivamente del hotel actual bajo análisis (sistema de puntuación de 10 estrellas)?, ¿qué grado de acierto tuvimos en nuestras predicciones previas (le gustaron las amenities, las recomendaciones, las excursiones propuestas, etc.)? Las respuestas a estas preguntas nos dan un panorama a futuro sobre qué datos incorporar o retirar de nuestros análisis, como además validar o refutar todas las hipótesis que tengamos. Como ejemplo, podemos agrupar a los puntos completados por los huéspedes del formulario de evaluación acorde al NPS, Net Promoter Score (siendo 1 detractor, 2 y 3 neutros, 4 y 5 promotores) en tres grandes grupos, el primero relacionado a su experiencia dentro de las instalaciones, sea habitación, recepción, limpieza; el segundo respecto a su experiencia en comidas y bebidas, y tercero a su experiencia en general en el hotel. De esta manera podemos asignarle un promedio a cada uno de estos grupos acorde al puntaje otorgado por el huésped en cada una de las preguntas incluídas dentro del mismo. De esta manera, podríamos prestar mayor atención a aquellos promedios cuyos puntajes sean menores a 3, o mayores a 4 como un parámetro de medición para ofrecer un hotel distinto o el mismo al anterior respectivamente.

Cuarto paso, ***transformación de datos***.

En relación a la primera problemática, agregaremos una columna para el cálculo de edad con fecha de nacimiento y validarlo con el mismo campo consignado en la encuesta. Agregaremos otra columna llamada días totales restando la fecha de partida a la fecha de llegada.

Respecto a la 2da problemática podríamos crear una nueva columna que calcule el promedio de las puntuaciones que los huéspedes asignaron a los distintos parámetros.

Quinto paso, ***data mining***,

En iteraciones sucesivas de los métodos inteligentes de data mining, logramos extraer patrones de comportamiento de los huéspedes y de sus preferencias de consumo. Estos patrones, pueden parecer obvios a priori e incluso pueden aparecer patrones nuevos o

asociaciones impensadas, pero que a través de la implementación de estos pasos, cuentan con soporte estadístico y tener prometedoras probabilidades de éxito. La calidad y la capacidad de ser implementados estos patrones encontrados, serán analizados en el siguiente paso.
En nuestra 1era problemática agruparemos a todos los huéspedes de un mismo rango etario pertenecientes a un mismo grupo familiar, esposa e hijos menores de 10 años, a los cuales les enviaremos anuncios de teatro familiar nocturno.
En la 2da problemática podemos analizar los datos históricos (y actuales basados en los consumos de la tarjeta) de nuestra base de datos y así extraer patrones frecuentes y poder descubrir asociaciones y correlaciones basadas en sus gustos e intereses.

Sexto paso, ***Evaluación de patrones.***

En la primera problemática, a medida que vayan transcurriendo los días podemos observar desviaciones las cuales nos aportarán datos para modificar nuestros patrones y predicciones en relación a los anuncios a mostrar. Podemos usar técnicas como la Discriminación de datos creando grupos y así poder tener información acerca de los gustos y preferencias de los huéspedes. Por ejemplo, podemos crear grupos discriminando por rango etario y observar si los gustos y preferencias son los mismos en cada clase o hay que modificar la discriminación. Otro ejemplo, supongamos que uno de nuestros hoteles ha creado actividades destinadas a niños de entre 8 y 11 años, sin embargo, a lo largo del proceso de evaluación de patrones, los usuarios de esas actividades terminan siendo en un 95% niños de entre 5 y 7 años, lo que nos hace dar cuenta de que estas deben cambiar las edades sugeridas, y deben ser destinadas familias con niños de este último rango etario.

En la 2da problemática con respecto al feedback recibido de nuestros huéspedes lograremos encontrar esos patrones los cuales nos serán de utilidad para poder ofrecer en base a los mismos el hotel adecuado en base a sus gustos y preferencias, teniendo en cuenta principalmente el NPS y tomarlo como base de nuestra principal estrategia para el envío de mailing. Como ejemplo, podemos identificar a través del proceso de evaluación de patrones que los hombres casados de entre 50 y 55 años valoran una buena barra de tragos nocturna, entonces orientaremos el anuncio de mailing a un hotel con estas características.

Por último**, *Knowledge presentation.***

Luego de aplicar las técnicas que adoptemos y hagamos las evaluaciones pertinentes, y de enviarle los anuncios a los huéspedes que nos sugiere nuestro algoritmo, podríamos incluir en el final del mismo si su contenido le ha sido de interés o no quiere más anuncios de ese tipo, o bien en la encuesta final de satisfacción si les ha parecido útil y pertinente los anuncios y novedades del hotel, y en el caso de que la respuesta sea afirmativa, qué beneficios obtuvo con ello o qué actividades realizó de las sugeridas por los mismos.

En nuestra 2da problemática armaremos el informe con los datos recabados lo que dará como resultado qué hotel ofrecer a cada uno de los huéspedes, habiendo utilizado las técnicas de NPS ya sea agrupándolos en clusters, grupos, categorías, y a partir de allí derivar dicha información al área correspondiente del proceso de envío de los anuncios.