CASO 1

Introducción a Data Mining

GRUPO 5

- Cardenas, Lucas
- García, Matías
- Martignoni, Franco
- Rodriguez Saá, Milagros

1. Planteo del problema

Objetivo general de negocio

Conocer de forma detallada a los clientes que visitan la cadena hotelera, de forma tal de poder realizar publicidad personalizada durante y pos-estadía.

Objetivos específicos de negocio

1) Generar publicidad personalizada sobre la cadena (eventos, actividades, promociones, etc.) a los huéspedes actuales, por medio de la TV de la habitación. Esto implica diferenciar claramente huéspedes nuevos (sobre los que no se tiene conocimiento a priori) y huéspedes conocidos, para delimitar qué anuncio resulta apropiado en cada caso.
2) Generar publicidad personalizada sobre los hoteles disponibles a aquellos huéspedes que ya terminaron una primera estancia en un hotel de la cadena, por medio del envío de un mail. En particular, qué selección de hoteles compartir. En este sentido, un huésped normalmente no volverá a elegir el mismo hotel en el que ya se hospedó y, por otro lado, puede volver a la cadena hotelera bajo un rol distinto que el de su primera estadía (por ejemplo, primero viajó por trabajo y ahora viaja por vacaciones familiares).

De los objetivos anteriores se desprende de que existen tres tipos de huéspedes distintos:

- Huésped A: huéspedes nuevos recién llegados al hotel.
- Huésped B: huéspedes conocidos, que se hospedan actualmente en el hotel, pero llevan un tiempo en el mismo.
- Huésped C: huéspedes antiguos, que ya transitaron una primera estancia en algún hotel de la cadena.

Fuentes de datos disponibles

De la información explicita provista en el planteo inicial; se distinguen 5 fuentes de datos:

1) Datos demográficos.
2) Datos resultantes del cuestionario para la adquisición de la credencial.
3) Datos de la credencial vinculados a actividades realizadas.
4) Datos de la credencial vinculados a transacciones efectuadas.

5) Datos obtenidos en el formulario de evaluación pos-estancia. Este formulario admite comentarios adicionales a los puntajes de evaluación, de forma tal que, en caso de utilizar los datos, habría que realizar un análisis y limpieza particular.

En donde los huéspedes tipo A sólo generan las fuentes 1 y 2, los huéspedes tipo B generan las fuentes 1, 2, 3 y 4, y los huéspedes tipo C todas.

Además, en todos los casos dispondremos de los datos referentes a su estadía, datos que no surgen del cuestionario sino de la contratación del servicio: qué calidad de habitación en qué categoría de hotel, cuántos días, cuántas camas, qué método de pago utiliza, si contrató desde una plataforma, qué época del año es y cómo son los precios en esta temporada respecto del resto del año, si contrató con servicio de almuerzo, cena, o solo desayuno, qué atractivos turísticos hay en esta zona en particular, si es zona de playa, montaña o urbana, etc.

También los datos de contratación y consumo históricos propios de sus estadías anteriores en nuestros hoteles. Si el huésped ya se alojó en alguno de nuestros hoteles tendremos acceso a su historial de consumo y comportamiento.

Un dato que puede ser fundamental para integrar al análisis fuentes externas de información es si su estadía está relacionada con algún evento de relevancia turística que ocurre en la zona del hotel (por ejemplo, convenciones, un recital, una obra de teatro, etc.).

2. Marco teórico: proceso de KDD

Los pasos del proceso de KDD según Han son:

1) Limpieza de datos para eliminar posible ruido y datos inconsistentes.
2) Integración de datos, combinando las múltiples fuentes.
3) Selección de datos relevantes para análisis.
4) Transformación de datos para que queden aptos para la minería.
5) Minería de datos.
6) Evaluación de patrones en relación a las medidas de interés.
7) Presentación de los resultados.

3. Aplicación marco teórico a objetivos particulares

3.1. Limpieza de datos para eliminar posible ruido y datos inconsistentes

Datos demográficos

La mayoría de estos datos pueden ser verificados Se contrastan con la documentación presentada por los pasajeros al momento de ingresar. Se verifica que los formatos y tipos de datos sean congruentes. Por ejemplo, el id de documento, pasaporte, los formatos de fechas cargados.

Datos de la credencial por uso de actividades

La limpieza de datos en este caso se acotará a identificar errores por mal funcionamiento de equipos. Un scanner defectuoso, una credencial que no funciona, molinetes o puertas descompuestas. Todas las situaciones que no permitan una carga completa y consistente del uso de amenities.

Datos de la credencial por transacciones efectuadas

En este caso la curación de datos se limitaría a estudiar los montos de compra para identificar posibles errores en la base: 1) valores de compra excesivamente grandes para productos/servicios no costosos; 2) montos de compra negativos. Por otra parte, es importante homogeneizar la moneda a utilizar, asumiendo que la cadena puede tener hoteles en diferentes países. En caso de que las compras no se registren en moneda local y en U$D, deberían obtenerse las series de tipo de cambio desde la web.

Cuestionarios de credencial y de evaluación

Los cuestionarios respondidos por los huéspedes deben ser revisados antes de ser cargados al sistema de la cadena. Los datos tendrán errores que no pueden ser verificados, y muchos de ellos estarán incompletos. **El diseño de los cuestionarios será fundamental para minimizar la posibilidad de error a través de opciones predeterminadas y poco lugar a la interpretación.** Un dato a tener en cuenta será si efectivamente completó los cuestionarios, en qué grado lo hizo y si fuera posible cuánto tiempo le dedicó a hacerlo. Algunas inconsistencias se pueden identificar de manera sencilla (por ejemplo contesta que es soltera y que viaja con su esposo, o que su elección de alojamiento se relaciona con su ubicación cercana a la playa, cuando no es el caso) y servirán para determinar si el registro tiene algún valor o si sería conveniente ignorarlo.

Eventos de interés turístico en la zona

Los datos de estos eventos pueden tener diferentes fuentes y su carga al sistema de la cadena debe considerarse según cada caso, fuentes especializadas o de noticias locales accesibles desde la web. De estas fuentes de información se debe limpiar los datos que no correspondan a eventos de interés turístico, verificar que el etiquetado sea correcto con el tipo de evento.

3.2. Integración de datos, combinando las múltiples fuentes

Aquí se creará un datawarehouse que tendrá los datos obtenidos de las diferentes fuentes para tenerlos disponibles en una versión unificada. Se relacionan los datos a través de sus claves primarias para identificarlos con los huéspedes de manera unívoca. Y se deja disponible para luego en los pasos de selección y transformación necesarios para la minería.

3.3. Selección de datos relevantes para análisis

Para definir la personalización de la publicidad en el hotel trabajaremos principalmente con los datos demográficos, los datos de uso de su credencial.

Para definir la estrategia de mails usaremos los datos demográficos, los datos de los cuestionarios previo y posterior y los datos de eventos de interés turístico.

3.4. Transformación de datos para que queden aptos para la minería

Se verifica que los datos tengan su formato correspondiente.

Se generan nuevas variables resumiendo datos relacionados al huésped: por ejemplo, se calcula la duración de la estadía, el uso tiempo de cada amenity, el consumo de bebidas resumido en precio, tipo y cantidad, cantidad de menores si viajó con sus hijos.

Se limpian los inputs de comentarios de huéspedes, por ejemplo, se lleva todo a minúsculas, se eliminan los espacios dobles, etc.

En caso de que los datos transaccionales de la credencial no estén en la misma moneda, se realizan las paridades necesarias.

3.5. Minería de datos

El proceso lo comenzaríamos por un análisis exploratorio de los datos para encontrar tendencias, dispersiones y cantidades.

Generamos gráficos para su mejor comprensión: tipos de huésped según el motivo de su estadía (descanso o trabajo), sexo, si viaja solo, en pareja o con familia, las edades, los días de estadía, si es día de semana o fin de semana y feriado, meses del año. Graficar amenities por su consumo relativo, las clases de bebidas más populares entre los pasajeros, etc.

Ya después se puede analizar si hay algunas relaciones respecto del comportamiento y evaluar si existen correlaciones entre variables.

En relación a los ítems particulares que menciona Han en el proceso de minería de datos:

- Caracterización y discriminación. Va en línea con la delimitación del tipo de huésped previamente descripta.
- Extracción de patrones frecuentes. Además de encontrar tendencias, pueden establecerse reglas de asociación; como por ejemplo qué amenities se consumen juntas.
- Clasificación y regresión. Como se requiere de datos etiquetados, podría aplicarse más fácilmente a los clientes ya conocidos (huéspedes B y C de la sección 1). Ejemplos de casos de clasificación: predecir qué grupo de actividades preferidas tienen los huéspedes (relajación, turismo aventura u otras), cuál es su nivel de consumo durante las estadías a partir de los datos transaccionales (consumo medio, alto, bajo). Y así enviar publicidades más a tono con sus intereses o posibilidades.
- Análisis de conglomerados: como se realiza a partir de datos no etiquetados, este enfoque serviría más para los clientes nuevos (huéspedes A de la sección 1). Con los datos existentes para los mismos, pueden establecerse clústeres con clientes que compartan ciertas características, e idear una publicidad de TV interesante para ellos.
- Análisis de valores atípicos. Delimitar huéspedes que tienen un perfil extraño, con outliers en sus variables, para estudiarlos en detalle.

3.6. Evaluación de patrones en relación a las medidas de interés

En este paso lo importante es analizar los insights relevantes para nuestras necesidades de publicidad personalizada.

Por ejemplo descubrimos que existe una relación entre la época del año y el tipo de huésped, que en navidad la proporción de ocupación por familia es muy superior al resto del año. En la temporada de eventos corporativos los huéspedes usan más el spa.

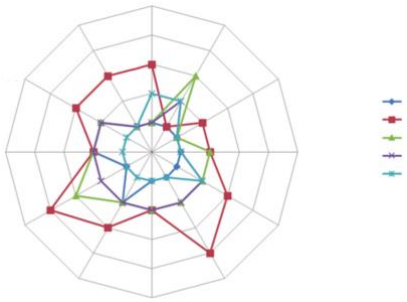
3.7. Presentación de los resultados

Se arman tablas con las medidas de estadística descriptiva, detallando por variable.

Se presentan los agrupamientos, se les da un nombre descriptivo y se plantea su publicidad óptima de acuerdo a los destinatarios tipificados de cada anuncio.

Presentaríamos distintos gráficos generales para visualizar la distribución de los huéspedes según sus consumos.

Para la presentación de datos y comparación de los patrones de huéspedes creemos que sería útil el armado de gráficos de estrella por cluster.



Se arman tablas de frecuencia de envío de publicidad detallando qué mensajes se enviaron a qué cantidad de huéspedes, qué anuncios de tv se mostraron y los resúmenes estadísticos de ambas decisiones.