Introducción a Data Mining

Caso 1

Roberto Buffo – Noviembre 2022

En tanto que la empresa hotelera en análisis es una multinacional con locaciones en diferentes partes del mundo, se asume que la recolección de datos en cada hotel responde a una circunstancia comercial y de promoción diseñada de antemano. Así, la "definición del problema" como paso previo a la metodología KDD se restringe justamente a la recopilación de información respecto de, incluso, variables definidas de antemano. En otras palabras, no se trata de obtener registros para construir una base de datos "cruda" a modo de *start-up* a partir de la cual seleccionar aquellos significativos y analizables, sino por el contrario para alimentar otra(s) de existencia previa. De todos modos, este estudio de caso se centra en información "fresca" a nivel de cada locación de la cual se intenta extraer conclusiones valiosas mediante técnicas específicas de *data mining* a fin de optimizar el negocio hotelero. Es decir que el paso inicial del KDD, "selección de variables", se considera asimismo a priori determinado en función de una estrategia comercial preexistente.

Al momento de la registración de los clientes, el hotel dispone de los datos demográficos de los mismos, tales como: número de pasaporte o de la identificación personal que corresponda, edad, sexo, estado civil, lugar de residencia permanente, país de procedencia. El cuestionario a modo de *likelihood survey* que se solicita llenar a cada pasajero al ingresar al establecimiento en el *check-in*, debe ser escueto, sucinto y bien encuadrado, que no lleve más de un par de minutos completar y que fuerce a quien lo realiza a elegir respuestas concretas y específicas, utilizando por ejemplo un sistema de *multiple choice* a fin de seleccionar niveles de una variable categórica, ergo *amenities* y eventos del hotel, o quizás una escala ordinal artificial de *likelihood* que permita cuantificar alguna de ellas. Una pregunta perentoria en el cuestionario, decididamente previa a la selección y/o cuantificación de *amenities* y eventos, es la referente al motivo del alojamiento, ya fuera, por ejemplo, por descanso vacacional, por trabajo, por asistencia a un simposio o conferencia, etc. Cada una de estas circunstancias condiciona a priori el número de días de la estadía y por lo tanto el tiempo del que cada persona dispone para disfrutar en mayor o menor medida de lo que ofrece el hotel.

Esta definición precisa del cuestionario no es un detalle menor: se trata de que cada cliente lo complete en su totalidad, minimizando lo más posible el faltante de datos que después, en el *data mining* propiamente dicho, entorpezca el análisis y condicione la validez de los resultados frente a un número significativo y molesto de valores NA; mejor resulta aquí "prevenir que curar". Una sólida toma de información se traduce también en una "limpieza de datos" sencilla, limitada al barrido de datos erróneos: difícilmente ocurran *outliers* per se en datos demográficos o en un cuestionario bien encuadrado, a menos que se trate, justamente, de errores eventuales de tipeo por descuido, relativamente simples de enmendar. Asimismo, la definición identitaria de las variables es de fundamental importancia en el posterior análisis de datos, es decir determinar unívocamente qué tipo de variable es cada una de ellas.

En cuanto a la "transformación de datos" del proceso KDD, una tarea imprescindible es la discretización de la variable Edad, a fin de definir "grupos" etarios: resulta incluso intuitivo el hecho de que las *amenities* ofrecidas a cada uno de ellos no serán las mismas. En caso de usar escalas

ordinales de *likelihood* en el cuestionario, se recomienda uniformar los rangos a fin de evitar un posterior escalamiento. De hecho, en el planteamiento del problema que motiva este análisis, se sugieren tales escalas en el formulario de evaluación al momento del *check-out*. Debieran por tanto usarse en ambas circunstancias: *check-in* y *check-out*, sumando en el primero asimismo la directa selección de opciones en base a preguntas tipo *multiple choice* que, como se dijo, fuerzan respuestas unívocas.

En principio, la información acerca de las *amenities* y eventos de hotel se difunden por el canal privado de TV del mismo. Que tal sistema permita o no la personalización de la información según el número de habitación y en función de quiénes la ocupan como resultante de los procesos de *data mining*, depende del refinamiento tecnológico correspondiente. En rigor de verdad, resulta más realista considerar que el hotel pueda anunciar un listado genérico de los eventos próximos y/o de las locaciones de diversión internas por el sistema cerrado de televisión y que lo haga de manera personalizada mediante WhatsApp al número de celular de cada pasajero, de acuerdo a la provisión voluntaria del mismo al momento del *check-in*. Es más, este sistema puede resultar tanto más efectivo considerando que el tiempo de permanencia en la habitación tiende a ser mínimo. Si eventualmente la personalización comunicacional vía TV es factible, se sugiere emitir la información correspondiente en horarios posteriores a la cena, antes del descanso nocturno o de la preparación para una salida nocturna, cuando sea más presumible que el pasajero pueda estar viendo el canal privado del hotel.

El planteamiento del problema también indica la provisión de una credencial personalizada de acceso a las *amenities*, incluso utilizable como tarjeta de crédito para consumos internos. Es decir que, en función del derrotero particular de cada cliente dentro del hotel y de sus correspondientes actividades de acuerdo a lo que el mismo ofrece, se recolecta, con el paso de los días de estadía, información tanto más valiosa que la obtenida en el *check-in*, ya que se trata de lo que efectivamente llevó a cabo y no de lo que potencialmente hubiese hecho: la diferencia en calidad de unos datos y otros resulta evidente en cuanto a definir los gustos del cliente y a la circunscripción de oferta para maximizar su consumo en el hotel.

En definitiva: hay un proceso de *data mining* de la pesquisa inicial y otro, de mayor refinamiento, de la información recogida a lo largo de la estadía de acuerdo al uso por parte de cada cliente de las *amenities* y los eventos ofrecidos. Es más, al momento del *check-out*, no solo cuenta el hotel con los datos que provea el cliente en el formulario de satisfacción sino también con lo que realmente hizo durante su estadía, lo cual fue por cierto registrado. De este modo se dispone de la información necesaria para llevar a cabo un tercer análisis de *data mining* que permita enviar al cliente, durante el invierno y de modo específico y personalizado, ofertas hoteleras para la subsiguiente temporada estival.

Las posibles técnicas de *data mining* a aplicar en la primera circunstancia de análisis, cuando se requiere información que permita iniciar la oferta de *amenities* y eventos al cliente inmediatamente a posteriori de la formalización del *check-in*, incluye:

a) Construcción de tablas de contingencia simples, a modo de análisis exploratorio, donde se relacione una variable demográfica con la elección de determinadas *amenities* y/o eventos, a fin de identificar la existencia o no de asociaciones estadísticamente significativas entre variables categóricas mediante el uso de frecuencias (recordar al respecto la discretización de

la variable Edad). Tal relación implica el uso de la prueba de Chi-cuadrado. Estas tablas permiten comenzar a construir "perfiles" o "arquetipos" de clientes de acuerdo a las correspondientes selecciones, al menos de manera tentativa. Si el número de tablas resulta alto en cuanto a la combinación del número de *amenities* o eventos con las variables demográficas, una posible elección "cruda" inicial de cuáles elegir puede basarse ya sea en la cuantificación de *likelihood* del cuestionario de ingreso o en el costo fijo de su funcionamiento y/o montaje. De todos modos, las tablas que sí tienen significancia en este primer análisis son aquéllas que relacionen amenities y eventos con la variable categórica "razón de la estadía", sin duda de mayor importancia relativa que cualquier dato demográfico en particular.

b) Estudios de regresión logística, donde la elección o no de una determinada *amenity* o evento específico, como variable binaria dicotómica dependiente (sí o no), sea función de las variables categóricas demográficas, a modo de variables independientes. Esta técnica permite afianzar la definición de perfiles de clientes. De manera orientativa, se puede aplicar el análisis de correspondencias múltiples, recordando que todas las variables demográficas son categóricas, a fin de definir la asociación de las variables independientes que permitan dirigir la formulación de los modelos logísticos.
c) Construcción de árboles de clasificación a partir de las variables demográficas y la razón de la estadía, para predecir el etiquetado de elección o no de determinadas *amenities* y/o eventos. Esta técnica de hecho resulta más simple y directa que un análisis de regresión logística, pero su eficacia dependerá de la cantidad de datos recolectados, en cuanto a dividirlos en grupos de entrenamiento y testeo, para que finalmente los árboles sirvan de instrumentos de predicción en base a los datos recogidos al momento del *check-in*.

Estas técnicas, usadas individual o colectivamente, deben llevar a la elección de una media docena de amenities y/o eventos que puedan ser inmediatamente ofrecidos por los canales de información mencionados.

En una segunda instancia de *data mining*, se utilizan, como se enunció previamente en este reporte, los datos recolectados de acuerdo al efectivo uso de *amenities* y eventos por parte de los clientes a lo largo de su estadía. Así, la oferta personalizada se torna tanto más refinada y específica. Se mencionan en este punto:

a) "Mejoramiento" de la regresión logística, ya que ahora se trata de alimentar los modelos con datos referidos a lo que efectivamente cada cliente consumió y/o en lo que participó y no del uso potencial derivado del cuestionario de entrada. Se trata asimismo de la validación o no de la información inicial.
b) "Refinamiento" de los árboles de clasificación, en base a similar criterio.
c) Definición de asociaciones de *amenities* y eventos, es decir, la construcción de series de objetos que aparecen juntos en una base de datos a fin de generar reglas de dependencia que permitan predecir la ocurrencia secuencial de ítems en la oferta del hotel. Esta información es muy importante, ya que define el ordenamiento temporal de los avisos y promociones enviados a cada cliente mediante TV o WhatsApp.

Para el tercer análisis de *data mining*, ya finalizada la temporada estival, la empresa cuenta con una base de datos "completa", con toda la información recogida en los meses de mayor actividad,

incluyendo el cuestionario de *likelihood* del *check-out*, registros a partir de los cuales se procede a generar ofertas destinada a antiguos clientes respecto de la temporada de verano subsiguiente. El procedimiento incluye:

a) Clustering de *amenities* y eventos en función de las frecuencias de conteo de utilización y del ordenamiento de satisfacción del cuestionario de salida.
b) Selección de los ítems más significativos de cada grupo (es decir, más cercanos al correspondiente centroide) a fin de correr tanto modelos de regresión logística en función de las variables demográficas y de la razón de la estadía como árboles de clasificación. Dichos modelos, entrenados y testeados, permiten generar en conjunto "perfiles arquetípicos" de clientes a los cuales corresponderán grupos de *amenities* y eventos específicos.
c) Identificación de *amenities* y eventos semejantes en agrupamientos similares corridos para otros hoteles de la cadena.
d) Pareo de los nombres de estos hoteles con cada perfil o arquetipo de cliente asociado a los eventos correspondientes.
e) Envío a los clientes de los *dossiers* de los hoteles con *amenities* y eventos de su preferencia, por cierto cuidando de cambiar la locación y/o de repetir el hotel ya visitado, ya que raramente un cliente volverá a exactamente al mismo.