MAESTRÍA EN CIENCIA DE DATOS

Materia: Introducción al Data Mining
Profesores: Gastón Pezzuchi
Eduardo Poggi

Trabajo Caso Número 1
Tema: Introducción al Data Mining
Alumnos:
Héctor Martinotti
Marcos Buccellato
Ángela Bastidas
Wilmer Alarcón

INTRODUCCIÓN

Data mining o *Knowledge Discovery from Data* (KDD) es el proceso de producción o "extracción" de información o conocimiento a partir de los datos. Esta actividad es vital para la toma de decisiones ya que nos encontramos en una situación de "abundancia de datos, acoplada con la necesidad de herramientas de análisis de datos que ha sido descripta como rica en datos, pero pobre en información" (Han pp. 5, 2012). El proceso de KDD puede separarse en siete etapas diferentes que se encadenan en secuencia, tiene problemáticas específicas y hacen uso de herramientas diferentes. Presentamos a continuación la aplicación de este proceso para el caso propuesto.

La situación planteada nos presenta fuentes de datos diversas y una necesidad puntual de accionar en pro de un objetivo con una serie de decisiones concretas. El objetivo del proceso KDD es convertir estas fuentes de datos heterogéneas que incluyen desde información demográfica hasta transacciones realizadas por los huéspedes de una cadena de hoteles, para poder construir un sistema de recomendaciones que le entregue a los clientes existentes ofertas de relevancia según sus necesidades, intereses y preferencias. Comencemos entonces analizando el problema siguiendo las siete etapas propuestas para un proceso de KDD.

Etapa 1: Limpieza de datos

El caso propuesto nos informa que hay un conjunto de datos diferentes de los cuales se puede hacer uso:

1. Formulario con datos demográficos del registro y un cuestionario de check-in
2. Registro de transacciones que incluye: uso de facilities del hotel, compras en locales y consumiciones.
3. Formulario de check-out con encuesta con preguntas puntuadas de 1 a 5 con un espacio para brindar de forma textual comentarios, quejas y recomendaciones.

Si bien no nos detalla el contenido de las preguntas de los formularios y tampoco se enumeran los facilities y servicios ofrecidos, tenemos suficiente información para entender las características de estos datos. En el primer caso sabemos que en el check-in probablemente nos encontremos con atributos como: ID, NOMBRE, GÉNERO,

PROVINCIA/ESTADO, PAÍS, FECHA DE NACIMIENTO, NATURALEZA DE LA VISITA, TIPO DE HABITACIÓN REQUERIDA, CANTIDAD DE INTEGRANTES, PERIODO DE ESTADÍA. Seguramente puedan consignarse más datos, pero al menos estos son necesarios para registrar a un usuario y planificar las actividades del hotel. El proceso de limpieza de estos datos está relacionado con la calidad del dispositivo de relevamiento de los mismos. No es lo mismo un formulario impreso con líneas de texto libre que un formulario Online con campos pre-categorizados. Según el caso las tareas de limpieza pueden ser varias.

Teniendo en cuenta que muchos de los campos son variables categóricas, si el proceso de carga ofrece categorías predefinidas el proceso de limpieza es simple. De lo contrario, esto implicara algunas técnicas particulares, por ejemplo, si los campos de carga fueron de texto libre, se deberá leer cada uno y comparar contra categorías existentes. Este trabajo puede ser complejo ya que seguramente pueden existir errores de escritura, lo cual puede llevar a la necesidad de recurrir a técnicas de comparación por similaridad de palabras (ej: distancia de Jaccard) para lograr identificarlas. El problema se complica más si tomamos en cuenta que aun sabiendo que todo está correctamente escrito, lo que un usuario puede cargar es altamente variable, por lo cual hay que limitar las categorías a las que son de interés. Esto se puede lograr con el uso de diccionarios que asocien grupos de palabras a categorías pre-definidas, para esto se pueden usar diccionarios propios o asistirse con técnicas de análisis de texto que permitan identificar, por ejemplo, raíces de palabras que se asocian con las categorías de interés. Dada la complejidad de este trabajo, quizás sería importante predefinir estas categorías una vez que se tenga en claro cuáles son las opciones más frecuentemente ingresadas por los usuarios y de interés para la cadena.

Para el caso de los países y las ciudades, las opciones son más sencillas ya que probablemente se utilicen listas predefinidas, a diferencia de las categorías, no hay un criterio de definición de estas que responda a motivos particulares, el universo está definido con claridad. Para el caso de las franjas etarias, probablemente, si se parte de la fecha de nacimiento, se pueden construir rangos de franjas etarias para convertir esta variable en una de tipo intervalo o tipo numérico. Lo mismo puede hacerse para el caso de las variables de tiempo de estadía de ser necesario, hay que considerar que esta acción no es reversible, es decir, que hay pérdida de datos al realizarla, por lo cual, la misma puede ser diferida a la etapa 4 de transformación de datos. También es cierto que es redundante tener la fecha de nacimiento y la edad, por lo tanto, esto puede dejarse para etapas posteriores.

El dato de la cantidad de integrantes del grupo que se registra es más complicado pero rico en información. En primer lugar, nos da la pauta de que tipo de acompañantes tienen los huéspedes, no es lo mismo una pareja que un grupo familiar o personas solas, esto puede caracterizar el tipo de visitantes que frecuentan el hotel. Por otro lado, cada registro de personas en un mismo grupo (probablemente identificado por la transacción comercial asociada y/o el número de habitación), permite establecer relaciones entre huéspedes que nos pueden no solo permitir el análisis de los datos de forma individual, sino que también nos permita hacer el análisis de datos en forma grupal. La limpieza de estos datos implicara poder establecer correctamente estas relaciones determinando que campos son los que definen la relación.

En este proceso es también donde se deben salvar todos los errores e inconsistencias producto de la carga del formulario. De nuevo, la cantidad de estos errores y problemas asociados dependerá de la calidad del instrumento de relevamiento; Supongamos que el formulario de entrada daba un campo para la provincia/estado y otro para el país, pero el formulario no presentaba esto de forma condicional (provincias dependiendo de la elección de país), entonces podemos encontrarnos con inconsistencias en los datos. En este caso probablemente debamos usar la nacionalidad como clave y perdamos el dato de provincia (al fin y al cabo, tendremos la información del pasaporte que nos da la nacionalidad como dato objetivo).

Si bien algunas de estas tareas podrían asociarse a las etapas de integración, el objetivo de esta etapa es lograr un set de datos depurado con variables definidas, la etapa de limpieza es relevante.

Para el caso de las transacciones, el proceso debería ser más sencillo. El caso propuesto nos informa que hay un sistema automático que registra las transacciones de cada cliente de forma automática por lo cual los errores y las inconsistencias deberían ser minimizadas; Claramente errores de lectura, cancelaciones en el uso deberían ser analizadas para poder obtener realmente el set de los datos de lo realmente consumido, hay que entender también que esta información lleva asociados datos que hay que asegurase que estén relacionados, por ejemplo los productos y servicios, que seguramente estén tipificados por categorías y tengan datos asociados. Lo interesante de estos datos es que no sólo nos dicen “que” consumió sino “cuando” lo que nos permite pensar en un orden temporal.

Por último están los datos del formulario de check-out. Afortunadamente el caso nos indica que este formulario tiene una parte de preguntas con una escala de puntos asociadas. Lo cual nos permite entender que cada pregunta responde a variables

ordinales. En cuanto a los campos para agregar comentarios, difícilmente sean apropiados en esta instancia hacer mucha limpieza. Quizás se podría hacer un filtrado de campos que tengan palabras sin sentido o estén vacíos y marcarlos como “NA”, para simplificar el trabajo posterior; Pueden también aparecer inconsistencias en esta etapa que incluyan preguntas contradictorias (por ejemplo preguntar si está satisfecho con el servicio en general y que la respuesta sea 5 y luego en las puntuaciones individuales de cada servicio sean 1) pero esto debería evitarse en gran medida si el cuestionario está bien redactado.

EL resultado final de este proceso deberían ser fuentes de datos limpias, consistentes y con atributos bien categorizados. De este análisis nos damos cuenta la importancia del correcto diseño de los instrumentos de relevamiento. Los formularios deben estar bien calibrados y los procesos de relevamiento transaccional deben permitir identificar casos problemáticos como cancelaciones o similares.

Etapa 2: Integración

En esta etapa es que se integran fuentes de datos diferentes y distintas bases de datos, como vimos en la etapa 1, tenemos tres fuentes diferentes de datos y adicionalmente tenemos datos de diversos hoteles con los que cuenta la cadena y en este paso se suele construir los datawarehouse. Lo primero que hay que hacer es relacionar las bases y guardar la información de forma que simplifique el análisis, para empezar, hay que determinar cómo establecer estas relaciones, sabiendo que cada cliente tiene un identificador único, este relacionamiento es sencillo de realizar. Ahora, en este caso la información va a ser integrada con fines de análisis de datos, no para operar o transaccionar, por tal motivo podemos reorganizar la información de forma diferente.

Una de las características de los datos transaccionales es que deben estar normalizados (independientemente de la forma normal) para evitar inconsistencias y redundancia de datos, pero para el registro con fines de análisis, la redundancia de datos no es un problema, por lo tanto, podemos generar datos que no sean normales (sabiendo que en la etapa 1 se atendieron las inconsistencias y problemas de cada fuente por separado). Es decir que podemos tener tablas de datos con muchos atributos para un mismo registro, esto puede no ser necesario, pero esta consolidación de datos facilita las consultas posteriores sobre todo en la etapa de selección de datos. Por tal motivo cambiar la estructura de los datos simplifica las tareas posteriores. Si el sistema fue bien diseñado y la limpieza bien realizada los problemas de consistencia de datos entre bases deberían ser mínimos. Ahora esto no es una obviedad, supongamos que el sistema de datos de check-in se hace con un sistema de un vendor particular y el registro

de transacciones se hace con otro sistema propietario dependiendo de la integración transaccional entre ellos puede haber información inconsistente entre ellos, transacciones no asociadas entre registros de diferentes bases de datos o supongamos que la encuesta de check-out se realiza por un sistema Online que no está integrado a las bases del hotel, en este caso habrá que identificar los registros de una y otra base de datos a partir de algún criterio (ID, NUMERO DE PASAPORTE, entre otros) y realizar el relacionamiento entre los registros de las diferentes bases. Las técnicas a utilizar en esta etapa va depender fuertemente del tipo de datos propios de cada sistema y de cómo haya sido diseñada la integración en el uso de los datos.

El resultado final de esta etapa son registros propios de cada huésped que integran los datos de las diferentes fuentes donde cada registro contiene toda la información pertinente, aunque pueda ser de manera redundante.

Etapa 3: Selección de datos

En esta etapa es donde se procede a la selección relevante de datos para el análisis, esta selección va a estar muy marcada por el objetivo buscado; Si bien el objetivo del análisis puede ser transversal a todas las etapas es en esta en particular donde el “para qué” tiene un fuerte impacto. A partir de esta selección es donde se podrán responder ciertas preguntas, al mismo tiempo la restricción de dimensiones irrelevantes para el análisis puede simplificar mucho las tareas posteriores, en este caso el objetivo es poder construir un sistema de recomendación de anuncios, para el mismo hay muchos datos que pueden no ser relevantes.

El primer criterio de selección seguramente deba recurrir al conocimiento experto del dominio, si bien es cierto que puede haber patrones de datos que pueden depender de variables que a priori pueden parecer irrelevantes, es posible usar un primer filtro que tenga que ver con lo que ya se sabe que no es relevante. Por ejemplo, supongamos que una de las preguntas del check-in fuera el color de las valijas del huésped o se le asignara a cada una un identificador (por ejemplo para que los asistentes las identifiquen y no las confundas), o supongamos que en las preguntas del check-out se le preguntaran cuestiones como que le pareció la marca de jabón de baño del hotel, estas preguntas podrían ser obviadas de la selección.

Pero si se tienen dudas sobre que se corre un riesgo de prejuzgar la importancia de atributos específicos en el análisis, se puede hacer un análisis exploratorio de datos

para buscar correlaciones o entender el peso de los diferentes atributos al momento de explicar las diferentes posibles correlaciones univariadas o multivariadas. Hay diferentes técnicas estadísticas que nos pueden permitir seleccionar o descartar diferentes atributos (ej. histogramas o diagramas de cajas, así como medidas descriptivas, distribuciones y frecuencias de las variables). Puede también darse el caso que haya campos que tengan problemas propios, por ejemplo, demasiados NA, o que se haya descubierto que estos datos siempre son ignorados o contestados de forma falaz por los huéspedes, es decir, que son poco confiables. Esto también puede ser un criterio de selección de datos.

El resultado de este proceso será un set restringido de datos con las variables que son más relevantes para el objetivo de análisis, descartando atributos que son irrelevantes, poco confiables o que no afectan o no están relacionados con el resto de los datos relevantes.

Etapa 4: Transformación

En esta tapa es que se procede a la transformación de los datos con el fin de poder analizarlos, como mencionamos en la etapa 1 hay tareas como la conversión de variables numéricas a intervalo a través de la discretización de datos puede realizarse en esta etapa. Este proceso implica perdida de datos, pero se dejan datos más relevantes para los procesos de análisis y toma de decisiones posteriores.

Por otro lado, en esta etapa es donde realizan agregaciones de datos. Por ejemplo, podemos generar sumas de consumos totales por rubro, cantidad de huéspedes por país, tipo de visita, tipo de grupo, entre otras; o actividades más detalladas como análisis de hospedajes de individuos particulares a lo largo de los diferentes hoteles de la cadena, lo importante en esta etapa es definir las segregaciones y las métricas necesarias para construir estos datos. Las combinaciones y agregados (con diferentes funciones como suma, promedios, máximos, mínimos, etc.) pueden ser muchas para lograr esto probablemente se deba trabajar buscando asociaciones de datos con análisis exploratorios consultas OLAP, correlaciones entre variables, componentes principales para encontrar composiciones de variables de interés explicativo, entre otras. Acá es importante resaltar una cosa, al agregar y sumarizar, por ejemplo entre tipos de cliente o tipos de transacciones realizadas, podemos utilizar los atributos clasificatorios relevados o podemos descubrir nuevos conjuntos clasificatorios producto de data mining. Por lo tanto, hay un proceso iterativo, por ejemplo, podemos descubrir un grupo de clientes que son aquellos que gastan mucho en bebida, vienen solos y en

viajes de negocios y otros que vienen en familia y gastan solo en servicios recreativos, etc, por tanto podemos querer transformar los datos acordes a esos patrones detectados en iteraciones previas, incluso podemos encontrar categorías nuevas para incorporar nuevas categorías discriminatorias a los formularios de check-in (por ejemplo darnos cuenta que hay categorías de tipo de visita nuevas). El sistema de recomendación final seguramente deba pasar por algunas iteraciones antes de llegar al producto final, el resultado de esta etapa es un set enriquecido de datos con agregaciones, transformaciones de atributos y definición de nuevas variables.

Hasta esta etapa es que se trabaja en el pre-procesamiento de los datos para poder descubrir patrones de interés en las siguientes etapas, este proceso no necesariamente es previo, sino que puede llegar a definirse a partir de iteraciones sucesivas y la experiencia en la práctica; En particular la etapa de transformación puede tener muy buena realimentación de las etapas posteriores.

Algunas variables que se podrían tener como resultado de esta etapa son las siguientes:

- Tipo de Cliente nuevo o antiguo.
- Nivel adquisitivo.
- Cantidad de visitas: cantidad total de hospedajes.
- Cantidad de hoteles visitados.
- Cantidad de ciudades visitadas.
- Promedio de días de hospedaje.
- Cantidad de visitas en pareja.
- Cantidad de visitas con familia.
- Cantidad de visitas de negocios.
- Cantidad de visitas de vacaciones.
- Gastos totales del cliente.
- Gastos en habitación del cliente.
- Gastos en snacks del cliente.
- Gastos totales del cliente en viaje de negocios.
- Gastos totales del cliente en vacaciones.
- Cantidad de campañas de email en las que se incluyó al cliente.
- Cantidad de clics únicos en email hechos por el cliente.
- Actividades que realiza el cliente dentro del hotel.

Etapa 5: Data mining

En esta etapa es donde comienza la aplicación de métodos inteligentes de análisis de datos para encontrar patrones en los mismos con la finalidad de producir información significativa para dar solución al problema de negocio, en este caso lo que queremos es producir un sistema de recomendación, entonces lo que vamos a tratar es de identificar patrones de consumo, demográficos y de satisfacción que nos permitan identificar grupos de clientes para el cual poder realizar anuncios personalizados. Lo que queremos es afinar el target de los anuncios para lograr captar la atención con contenido relevante para el cliente y evitar el superfluo que termina desviando su atención no logrando ventas efectivas de productos. Queremos lograr un clustering general de los huéspedes con patrones asociativos entre las variables diferentes.

Para esto podemos comenzar entendiendo el tipo de información que tenemos, si vamos al caso de las transacciones, nos vamos a encontrar que podemos realizar análisis de "canasta de productos" es decir que podemos realizar comparaciones en las similitudes de productos, teniendo en cuenta que la canasta de posibles productos es muy grande, seguramente nos interese entender que productos compraron en común y no tanto la "distancia" entre las diferentes canastas, para esto aplicaremos probablemente medidas de similaridad como la similaridad de coseno o la distancia de jaccard. con esto podremos comparar diferentes conjuntos de transacciones por huésped para poder encontrar patrones de compra. Siguiendo con las transacciones, no solo son datos de canasta de producto sino que son datos de secuencia, nos indica "cuando" se realiza la compra, esto puede permitirnos extender la identificación de patrones para ver cuando se consumen los diferentes productos a lo largo de un día (hay que considerar el problema de la variable cíclica del tiempo para la comparación), Podemos tratar de realizar también regresiones para encontrar patrones de gasto y consumo, entender y estimar cantidades consumidas en relación con otras variables; En el caso de las variable demográficas podemos encontrar patrones que están relacionados con los datos espaciales, por ejemplo reconocer que hay regiones enteras con patrones similares en función de diferentes variables (por ejemplo encontrar que los clientes del cono sur americano de una franja etaria particular tienen cierta conducta común) es decir que la distancia y contigüidad espacial puede ser relevante para el análisis. En lo referente a los datos de satisfacción podemos encontrar asociaciones y correlaciones entre grupos que nos muestren patrones de preferencias particulares. Si también consideramos los campos de comentarios, reclamos y sugerencias, podemos realizar un análisis de texto y detectar sentimientos positivos o negativos en general como

insumo para trabajar sobre la detección de patrones, también podemos clasificar los textos semánticamente para localizar focos problemáticos y así identificar ofertas alternativas, etc.

Lo que nos interesa en esta etapa es identificar la mayor cantidad de patrones posibles en los datos en conjunto para poder tener como insumo para la construcción del sistema de recomendación.

Etapa 6: Evaluación de patrones

En esta etapa es que se definen, de todos los patrones identificados, aquellos que son relevantes para el problema en cuestión. En este caso, que patrones me permiten identificar diferentes grupos de clientes para el targeting de anuncios. Para identificar estos patrones podemos seguir el criterio propuesto por han:

> *… un patrón es interesante si es (1) fácilmente comprensible para los humanos, (2) válido en datos nuevos o de prueba con cierto grado de certeza, (3) potencialmente útil y (4) novedoso. Un patrón también es interesante si valida una hipótesis que el usuario buscaba confirmar. Un patrón interesante representa el conocimiento. (Han págs. 21, 2012)*

De los resultados anteriores pueden haber surgido patrones complejos que son incomprensibles desde la perspectiva del dominio de negocio que pueden ser irrelevantes para el problema en cuestión, con un nivel de probabilidad bajo de que se presenten o identifiquen o que sean cosas ya sabidas (por ejemplo, que los grupos de personas que vienen con niños suelen ser grupos familiares y vienen en viaje familiar). Entonces es en este momento que deben seleccionarse aquellos patrones que cumplen con estas condiciones enunciadas para que resulten útiles para resolver el problema.

Un primer criterio para seleccionar patrones relevantes es el criterio de interés, es decir la relación entre el patrón encontrado y la cantidad de casos a los que aplica. Un patrón que aplica a una muy baja proporción de los huéspedes quizás no sea de interés. Otro criterio es el nivel de confianza de que un patrón se presente y por otro lado hay criterios propios del dominio de negocio que tienen que ver con accionabilidad, es decir un patrón que me permita tomar una acción relevante. Por ejemplo, supongamos que encontramos un patrón entre quienes consumen café y el nivel de satisfacción con el servicio de habitación, este es un patrón interesante pero no me da ninguna posibilidad de accionar en la oferta de comerciales en la TV del hotel.

En esta etapa es que se selecciona patrones que sirvan concretamente para la toma de decisiones en el problema de negocios, es importante mostrar los resultados en un formato entendible, por esta razón las técnicas de visualización son importantes para que los resultados sean útiles y fácil de interpretar.

Etapa 7: Presentación

En esta etapa es donde se presentan los resultados a los usuarios para la toma de decisión. Sin embargo, en este caso la presentación implica la ejecución de los modelos de recomendación de comerciales. Así que es en este punto donde podemos responder las preguntas propuestas por el ejercicio.

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

Para el primer problema el análisis realizado puede ofrecer una solución, los patrones detectados permiten identificar los tipos de cliente objetivo para los diferentes tipos de anuncios en función de su conducta. Así que al registrarse al hotel y comenzar a transaccionar, el sistema puede ir asociando al huésped con una de las categorías detectas e ir ofreciéndole el contenido adecuado. En cuanto al problema específico de que ofrecerle antes de que el cliente transaccione lo suficiente, es decir en los primeros días, los datos de registro y eventualmente el historial del cliente nos pueden proveer los datos para poder darle una oferta adecuada en esos primeros días. Cuando la conducta efectiva del huésped se manifieste con cierta claridad se puede afinar mejor el resultado de la oferta, sin embargo,

Si el cliente es totalmente nuevo en la cadena con los datos del check-in como tipo de viaje (negocios o vacaciones), acompañantes (viaja solo, en pareja, con familia o con amigos), días de hospedaje permitirá realizar una categorización del tipo de cliente para definir anuncios desde el primer día de alojamiento.

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

En cuanto a la selección de hoteles para enviarle al mailing, los patrones de consumo de clientes similares y el historial del mismo cliente son los que determinan que oferta es la adecuada. En particular seguramente la ubicación geográfica indicará por ejemplo cuando tiene periodos vacacionales y cuando no, la franja etaria probablemente el tono

del mensaje y el tipo de oferta, el tipo de consumo que ha hecho y el nivel de gasto indicarán el nivel adquisitivo del cliente, pero estos sencillos discriminadores probablemente se complementen con patrones más elaborados que podrán ser identificados en función de la información registrada para el huésped. En cuanto a la captación de clientes nuevos por mail, las bases de datos de prospectos podrán compararse con los patrones definidos para poder encontrar prospectos similares a cada categoría objetivo-detectada y así entregar la oferta adecuada.