Universidad Austral - Maestría en Ciencia de Datos

Introduccion al Data Mining

Caso #1 v2

Autor: Andres Montes de Oca

Fecha: 14 de Diciembre de 2022

Caso de Estudio

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 informarlos 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 cuales 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, explicando cómo se aplican en este caso. Indicar para cada paso cuáles técnicas usaría y justificar su elección.

Objetivo

Utilizar técnicas de Data Mining en los datos que tengo sobre mis huéspedes, para entender y predecir el comportamiento de los mismos durante la estadía en el hotel. También utilizaremos estos datos para poder decidir como clasificaremos a nuestros huéspedes, teniendo en cuanta los mailings anuales con las promociones de marketing.

KDD del caso de estudio

Primero empecemos con la definición, KDD (Knowledge Discovery in Databases) es un proceso semi-automatico por el cual, mediante la aplicación de distintas técnicas y algoritmos, podemos transformar información oculta en los datos en Insights de valor para el negocio. Las etapas en dicho proceso, y como se relacionan con nuestro caso de estudio, son las siguientes:

1- **Recopilacion de Datos**: Una vez ya comprendida la pregunta de negocio que queremos responder, empzamos haciendo un relevamiento de las fuentes de datos que disponemos. Muchas veces estas no son suficientes, y tenemos que recurrir a otras fuentes, ya sean internas o externas. En nuestro caso de estudio disponemos de los siguientes datos:
    a. *Datos del Huesped:* Son los datos que ya disponemos desde que el husped realizo la reserva, ya sea de manera telefónica u online. Los mismos serian Nombre, Apellido, Pasaporte, Nacionalidad, teléfono, email.
    b. *Datos de la Estadia*: Tipo de viaje (trabajo/placer). Huespedes adicionales, tipo de habitación, fecha ingreso, fecha egreso. Estos datos también se cargan durante la reserva
    c. *Datos Formulario Egreso*: De aquí es donde voy a obtener la información la cual me va permitir clusterizar al tipo de viaje que el huésped esta haciendo, para poder ofrecerle de manera mas personalizada los servicios con mayor probabilidad de ser consumidos. Si el tipo de viaje es de placer, podemos indagar sobre cuales son las actividades preferidas del grupo familiar: Actividades diurnas como salir a correr, clases de yoga, actividades deportivas como tenis o golf, Tipo de excursiones preferidas, de agua, mas de acción, hiking, etc. También indagaremos sobre los apetitos culinarios de los huéspedes, para ofercerle un menú mas personalizado en el restaurant a la noche, que tipos de bebidas prefieren, etc.
    d. *Datos uso de Amenities*: La tarjeta otorgada al huésped dejara registro de todos los consumos que haga en el bar, restaurant, tiendas de regalos. También registra el ingreso a los distintos amenites, como la pileta, el SPA, sauna, canchas, etc.
    e. *Datos Formulario Egreso*: Mediante este formulario, intentaremos recopilar información relevante sobre la calidad de la estadía del

huésped. En la encuesta se le solicitara al huésped que califique en una escala del 1 al 5, la calidad de los servicios brindados

2- **Pre-Procesamiento de Datos**: Esta es una de las etapas mas criticas, y que mas tiempo demanda. Aquí tenemos que velar por la calidad de los datos, y no solo la cantidad. Las tareas mas comunes serian:
   a. *Manejo de Datos faltantes*: Hay que decidir que hacer con los datos faltantes. Si en una variable el ratio de datos faltantes es alto, mas del 70% por ejemplo, talvez no tenga mucho sentido incluir esa variable en el estudio. Por otro lado, si el ratio es bajo, talvez me alcance con completar los valores nulos por algún estadístico que sea representativo para esa variable (si los datos son categóricos podría usar la moda, o si son cuantitativos podría usar la media o mediana). En nuestro caso de estudio habría que prestar atención a los formularios cargados por el huésped, ya que muchas veces los mismos son ignorados y no se completan.
   b. *Manejo de Datos duplicados*: Se debe prestar especial atención a los datos duplicados, muchas veces la causa puede ser un error en el proceso de consolidación de las distintas bases de datos. Otras veces puede no ser un error en lo absoluto, y justamente representa que una misma acción sucedió igual dos veces. Por esto y por otras cosas mas, el conocimiento del dominio de negocio es fundamental. En nuestro caso de ejemplo podemos encontrar huéspedes que se alojan en mas de un hotel durante el mismo tiempo. O registros duplicados en los datos de consumo de la tarjeta de amenities, generado por un mal funcionamiento de alguno de los lectores de tarjeta.
   c. *Manejo Datos Erroneos*: Suelen ser valores muy raros, que no representan la realidad. Podría ser una persona de mas de 100 anos en el registro de ingreso, o un grupo familiar de dos dígitos muy grande. Estos errores suelen causados por un error en el ingreso de los datos, generalmente errores del tipo humano. Para reducir el ratio de los mismos, deberíamos diseñar los distintos formularios de manera correcta, el uso de campos abiertos.
   d. *Manejo de Outliers*: La diferencia entre Datos Erroenos y Outliers, es que los últimos, si bien son atípicos, si son datos reales. Muchas veces los Outliers son justamente los datos de estudio, y nunca deben ser eliminados o ignorados. Se los debe manejar con cuidado, empleando técnicas robustas para su estudio. En nuestro ejemplo, un huésped que pase mas de 90 dias por ano en alguno de nuestros hoteles, seguramente sea identificado como Outlier, pero también como un usuario VIP. Es de suma importancia para el negocio tener constantemente identificados a estos usuarios, para bajar la tasa de turover (churn)

3- **Transformación de Datos:** En esta etapa es donde se buscan características útiles para nuestro análisis, realizando distinto tipo de modificaciones y transformaciones en los datos de entrada. El output de esta etapa es un Dataset limpio, que será usado como input para la creación del modelo. Hay varios tipos de transformaciones que se podrían realizar en nuestros datos, cada una con distintos objetivos:
    a. *Normalizacion de Variables Continuas:* Para que el funcionamiento de muchos algoritmos de Machine Learning sea optimo, muchas veces es mejor escalar las variables continuas de alguna manera, como la Normalizacion Z o la normalización Min-Max. Cabe destacar que esto solo modifica la escala de los datos, es decir los valores numéricos, pero la forma de la distribución se mantiene exactamente igual.
    b. *Análisis de Correlación*: Muchas veces distintas variables de nuestro dataset hablan sobre lo mismo, pudiendo estar fuertemente relacionadas. Para confirmar esto, podemos hacer un análisis de correlación entre las variables numéricas. En nuestro caso se me ocurre
    c. *Discretizacion de Variables Continuas*: De esta manera podemos transformar una variable numérica a una variable categórica ordinal, en nuestro caso podríamos discretizar la variable edad de nuestros huespedes, para dividirlos en subgrupos mas fácilmente interpretables (Niños, Adultos, Seniors). Hay que tener en cuenta que cuando se discretiza una variable se pierde un nivel de información, ya no sabemos mas las edades exactas de los huéspedes, solo su rango etareo
    d. *Data Aggregation*: Podemos consolidar la información de cierta manera que facilite su comprensión, realizando algún tipo de resumen que facilite la interpretación de los datos. En nuestro ejemplo podemos agregar la suma total de los gastos del huésped en el restaurant,
    e. *Encoding de Variables Categoricas*: Las variables categóricas, que suelen ser de tipo cadena de caracteres, deben someterse a un proceso de conversión para poder ser representadas en el sistema numérico. En nuestro caso de estudio, podemos mencionar los distintos tipos de amenities del hotel, si el tipo de viaje es de negocios o no, el estado civil del huésped, etc.
    f. *Análisis de Componentes Principales(PCA):* Es una técnica que mediante la aplicación de combinaciones lineales en las variables cuantitativas, permite la reducción de dimensionalidad de las mismas, en set de datos donde estas estan altamente correlacionadas. Esta reducción de la dimensionalidad nos puede ayudar en la simplificación del problema, como también en la visualización del mismo, ya que al ser humano se le complica la visualización de datos de mas de 3 dimensiones. En nuestro caso de estudio podríamos utilizarlo en el análisis de los distintos tipos de gastos que hizo el huésped.

4- **Data Mining:** Esta es la etapa fundamental del proceso de KDD, es la etapa en donde se utilizan los distintos métodos inteligentes que nos permiten encontrar insights y patrones ocultos en los datos.
    a. *Clasificación:* Son parte de la familia de modelos del Machine Learning Supervisado, en los cuales nosotros sabemos con anticipación el resultado de la variable Target. Esto nos permite medir que tan buenos son los desarrollos de nuestros modelos, ya que tenemos contra que comprar. En el caso puntual de la clasificación, la variable Target es una variable categórica. En nuestro caso, un claro ejemplo donde se usaría es en la construcción del modelo de mailing de las promociones de verano, el cual nos ayudaría a identificar cuales son los hoteles que mas probabilidades tienen de ser visitados por nuestros huéspedes durante sus próximas vacaciones. Existen distintos tipos de algoritmos de clasificación, como los Arboles de Decision, la Regresión Logistica, Modelos Bayesianos, Redes Neuronales, entre otros.
    b. *Regresión Lineal:* También son parte de la familia de modelos de Machine Learning supervisado, pero a diferencia de los Modelos de Clasificacion, aquí la variable target es una variable cuantitativa. En nuestro caso, un claro ejemplo seria poder predecir los gastos que diariamente van a tener nuestros huéspedes en el restaurant. Otros casos donde estos modelos también pueden ser útiles son las Series Temporales, esto por ejemplo nos podría ayudar en predecir cuando es que un huésped va a volver a hospedarse con nosotros.
    c. *Análisis de Correspondencia:* Es otra técnica similar al Análisis de Componentes Principales, pero en lugar de hacerse sobre variables cuantitativas se hace sobre las cualitativas o categóricas. La misma nos ayuda en la comprensión de las relaciones que existen (o no), entre las distintas variables categóricas del set de datos. En nuestro caso de ejemplo, lo podríamos usar para relacionar si el tipo de viaje que hace nuestro huésped se relaciona con el valor del score de satisfacción, indicado en el formulario de egreso (bajo, medio, alto). A diferencia del PCA, este análisis es solamente de carácter exploratorio, facilitándonos la comprensión de los datos pero no nos sirve como input para otro modelo, como si lo hace el PCA.
    d. *Clustering:* Este tipo de modelos nos permite segmentar nuestros por similitud, pero sin disponer de los valores del Target, ósea que pertenece a la familia de modelos de Machine Learning No-Supervisado. En nuestro caso, se podría implementar en la segementacion inicial de clientes, cuando no tenemos mucha información sobre ellos y solo disponemos de los datos llenados en la reserva y de los datos del formulario de ingreso. Esto lo usamos luego a la hora de decidir que publicidades de TV enviamos por el canal privado a cada huésped, según a que cluster pertenece.

5- **Presentacion y Evaluación de Patrones:** En esta etapa es donde evaluaremos si los insights obtenidos durante el análisis son de utilidad para las decisiones de negocio que necesitamos tomar. Se utilizan muchas graficas y conceptos de Story-Telling, ya que los mismos ayudar en facilitar la interpretación de los mensajes que queremos transmitir. En nuestro ejemplo podriamos crear visualizaciones de distintos tipos como:
    a. *Histogramas:* Para verificar la normalidad de las distribuciones de mis datos cuantitativos, como por ejemplo los gastos en el restaurant de mis husepedes
    b. *Gráficos de Torta:* Visualizando que porcentaje del total de huéspedes entre Mujeres y Hombres
    c. *Gráficos de Barra:* Para visualizar de manera sencilla los países de origen mas frecuente entre mis huéspedes
    d. *Tablas Contingencia:* Para analizar la asociación entre el sexo de los huéspedes, y que tipo de amenities les son de interés
    e. *Matriz de Confusion:* Es una tabla de 2x2, la cual nos ayuda a visualizar la cantidad de clasificaciones correctas (TP y TN), e incorrectas (FN y FP). En nuestro caso la usaremos cuando midamos el impacto que tuvo el mailing de verano enviado a nuestros clientes, visualizando en la diagonal principal los huéspedes correctamente clasificados (los que prefije que volvían y volvieron, como los que predije que no iban a volver y no lo hicieron), y en la diagonal secundaria los clasificados erróneamente.