**--Data Preprocessing--**

El Capítulo 1 nos presentó a la minería de datos y al proceso estándar entre industrias para la minería de datos (CRISP-DM) proceso estándar para el desarrollo del modelo de minería de datos. En la fase 1 del proceso de minería de datos, comprensión empresarial o comprensión de investigación, las empresas y los investigadores primero enuncian los objetivos del proyecto, luego traducen estos objetivos en la formulación de una definición del problema de minería de datos y finalmente preparan una estrategia preliminar para lograr estos objetivos.

Aquí en este capítulo, examinamos las siguientes dos fases del proceso estándar CRISP-DM, comprensión de los datos y preparación de estos. Mostraremos cómo evaluar la calidad de los datos, limpiar los datos sin procesar, tratar los datos faltantes y realizar transformaciones en ciertas variables. Todo el Capítulo 3 está dedicado a este aspecto muy importante de la fase de comprensión de datos. El corazón de cualquier proyecto de minería de datos es la fase de modelado, que comenzamos a examinar en el Capítulo 7.

**2.1 Why do We Need to Preprocess the Data?**

Gran parte de los datos sin procesar contenidos en las bases de datos no están procesados, son incompletos y son ruidosos. Por ejemplo, las bases de datos pueden contener

* campos que son obsoletos o redundantes;
* valores faltantes;
* valores atípicos;
* datos en una forma no adecuada para los modelos de minería de datos;
* valores no consistentes con la política o el sentido común.

Para ser útiles para fines de minería de datos, las bases de datos deben someterse a un procesamiento previo, en forma de limpieza y transformación de datos. La minería de datos a menudo trata datos que no se han examinado durante años, por lo que gran parte de los datos contienen valores de campo que han caducado, ya no son relevantes o simplemente faltan. El objetivo principal es minimizar la basura que entra y sale (GIGO), para minimizar la basura que ingresa a nuestro modelo, para que podamos minimizar la cantidad de basura que nuestros modelos dan.

Dependiendo del conjunto de datos, el preprocesamiento de datos solo puede representar del 10 al 60% de todo el tiempo y el esfuerzo de todo el proceso de minería de datos. En este capítulo, examinaremos varias formas de preprocesar los datos para su posterior análisis posterior.

**2.2 Data Cleaning**

Para ilustrar la necesidad de limpiar los datos, echemos un vistazo a algunos de los tipos de errores que podrían colarse incluso en un pequeño conjunto de datos, como el de la Tabla 2.1.

Analicemos, atributo por atributo, algunos de los problemas que se han introducido en el conjunto de datos de la Tabla 2.1. La variable de ID de cliente parece estar bien. ¿Qué hay de zip?

Supongamos que esperamos que todos los clientes de la base de datos tengan el código postal estadounidense de cinco números habitual. Ahora, el cliente 1002 tiene este código postal extraño (para los ojos estadounidenses) de J2S7K7. Si no fuéramos cuidadosos, podríamos sentir la tentación de clasificar este valor inusual como un error, y descartarlo, hasta que dejemos de pensar que no todos los países usan el mismo formato de código postal. En realidad, este es el código postal (conocido como código postal en Canadá) de St. Hyancinthe, Quebec, Canadá, y probablemente representa datos reales de un cliente real. Lo que evidentemente ha ocurrido es que un cliente franco-canadiense realizó una compra y colocó el código postal de su casa en el campo requerido. En la era del libre comercio, debemos estar preparados para esperar valores inusuales en campos como los códigos postales que varían de un país a otro.

¿Qué pasa con el código postal para el cliente 1004? No conocemos ningún país que tenga códigos postales de cuatro dígitos, como el 6269 indicado aquí, por lo que esto debe ser un error, ¿verdad? Probablemente no. Los códigos postales para los estados de Nueva Inglaterra comienzan con el número 0. A menos que el campo del código postal se defina como carácter (texto) y no numérico, el software probablemente cortará el cero inicial, que aparentemente es lo que sucedió aquí. El código postal puede ser 06269, que se refiere a Storrs, Connecticut, sede de la Universidad de Connecticut.

El siguiente campo, género, contiene un valor faltante para el cliente 1003. Más adelante en este capítulo detallaremos los métodos para tratar los valores faltantes.

El campo de ingresos tiene tres valores potencialmente anómalos. Primero, se muestra que el cliente 1003 tiene un ingreso de $ 10,000,000 por año. Si bien es totalmente posible, especialmente cuando se considera el código postal del cliente (90210, Beverly Hills), este valor de ingreso es, sin embargo, un valor atípico, un valor de datos extremo. Ciertas técnicas de modelado estadístico y de minería de datos no funcionan sin problemas en presencia de valores atípicos; por lo tanto, examinaremos los métodos de manejo de valores atípicos más adelante en este capítulo.

La pobreza es una cosa, pero es raro encontrar un ingreso negativo, como lo ha hecho nuestro pobre cliente 1002. A diferencia del ingreso del cliente 1003, el ingreso reportado del cliente 1002 de - $ 40,000 está más allá de los límites de campo para el ingreso, y por lo tanto debe ser un error. No está claro cómo se introdujo este error, y quizás la explicación más probable sea que el signo negativo es un error de entrada de datos extraviados. Sin embargo, no podemos estar seguros y, por lo tanto, debemos abordar este valor con cautela e intentar comunicarnos con el administrador de la base de datos más familiarizado con el historial de la base de datos.

Entonces, ¿qué tiene de malo el ingreso del cliente 1005 de $ 99,999? Quizás nada; De hecho, puede ser válido. Pero, si todos los otros ingresos se redondean a los $ 5000 más cercanos, ¿por qué la precisión con los ingresos del cliente 1005? A menudo, en las bases de datos heredadas, ciertos valores específicos están destinados a ser códigos para entradas anómalas, como los valores faltantes. Quizás 99.999 fue codificado en una base de datos antigua para significar que falta. Nuevamente, no podemos estar seguros, y nuevamente deberíamos consultar al administrador de la base de datos.

Finalmente, ¿tenemos claro en qué unidad de medida se mide la variable de ingreso? Las bases de datos a menudo se fusionan, a veces sin molestarse en verificar si tales fusiones son completamente apropiadas para todos los campos. Por ejemplo, es muy posible que el cliente 1002, con el código postal canadiense, tenga un ingreso medido en dólares canadienses, no en dólares estadounidenses.

El campo de edad tiene un par de problemas. Aunque todos los demás clientes tienen valores numéricos para la edad, la "edad" de C del cliente 1001 probablemente refleja una categorización anterior de la edad de este hombre en un contenedor etiquetado como C. El software de minería de datos definitivamente no permitirá este valor categórico en un campo numérico, y tendremos que resolver este problema de alguna manera. ¿Qué tal la edad del cliente 1004 de 0? Quizás, hay un hombre recién nacido que vive en Storrs, Connecticut, que ha realizado una transacción de $ 1000. Lo más probable es que la edad de esta persona probablemente no se encuentre, y se codificó como 0 para indicar esta u otra condición anómala (por ejemplo, se negó a proporcionar la información sobre la edad).

Por supuesto, mantener un campo de edad en una base de datos es un campo minado en sí mismo, ya que el paso del tiempo rápidamente hará que los valores del campo sean obsoletos y engañosos. Es mejor mantener campos de tipo de fecha (como fecha de nacimiento) en una base de datos, ya que son constantes y pueden transformarse en edades cuando sea necesario.

El campo del estado civil parece estar bien, ¿verdad? Tal vez no. El problema radica en el significado detrás de estos símbolos. Todos pensamos que sabemos lo que significan estos símbolos, pero a veces nos sorprende. Por ejemplo, si está buscando agua fría en un baño en Montreal y abre la llave marcada con C, es posible que se sorprenda, ya que la C significa chaude, que en francés significa caliente. También está el problema de la ambigüedad. En la Tabla 2.1, por ejemplo, ¿la S para los clientes 1003 y 1004 significa individual o separada?

El campo del monto de la transacción parece satisfactorio, siempre y cuando estemos seguros de que sabemos qué unidad de medida se está utilizando y que todos los registros se tramitan en esta unidad.