ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Data Understanding & Preparation

Minería de Datos y Modelización Predictiva

Master Big Data y Data Science. Aplicaciones al Comercio, Empresa y Finanzas

Universidad Complutense de Madrid

Curso 2022-2023

Las fases de "Data understanding & preparation" de la metodología CRISP-DM comprenden lo que se conoce como *Depuración de datos*.

Depuración de datos

La depuración de datos es el proceso mediante el cuál se **adecúa** el conjunto de datos para la fase de modelización posterior. Consiste en "limpiar" el conjunto de datos original de manera que no haya observaciones **incoherentes**, las variables se ajusten a las **especificaciones** del modelo, etc.

Fases de la depuración de datos

- Comprobación de la correcta tipología y rol de las variables.
- Análisis exploratorio de los datos y corrección de errores detectados.
- Búsqueda de datos atípicos.
- Tratamiento de datos faltantes.
- Transformación de variables.

El primer paso que se ha de llevar a cabo es la **exploración** de los datos, lo que nos indicará si hay algún problema con los mismos, para lo que es imprescindible poseer cierto **conocimiento** sobre el significado de los datos. Lo realizaremos a través de gráficos y tablas.

Aspectos a evaluar

- Comprobación de la correcta tipología de las variables. Si las variables cualitativas están representadas con carácteres numéricos, el software suele considerarlas como numéricas. Además, es recomendable tratar como variables cualitativas aquellas variables numéricas con menos de 10 valores diferentes.
- Análisis del número de datos faltantes (missings).
- Comprobación de las categorías de las variables cualitativas.
- Comprobación de los límites de las variables cuantitativas.

Aspectos a evaluar

- Análisis de la coexistencia de datos faltantes en varias variables.
- Comprobación de la representatividad de las categorías minoritarias de las variables cualitativas.
- Detección de la correcta introducción de datos faltantes codificados (“-1”, “9999”).
- Análisis de la simetría de las variables y de la posible existencia de datos atípicos.

Corrección de los errores detectados

- Variables cualitativas mal consideradas como cuantitativas.
- Datos faltantes codificados (“-1”, “9999”) no declarados.
- Valores de variables cuantitativas fuera de rango.

Corrección de los errores detectados

- Categorías de variables cualitativas erróneas.
- Unión de categorías de variables cualitativas con poca representatividad.

Búsqueda de relaciones entre variables

- Relación **continua-continua** –> Scatterplot. Coeficiente de correlación
- Relación **continua-categórica** –> Boxplot/Histograma de la continua en los grupos formados por la categórica. Diferencia de medias/ANOVA/Kruskal-Wallis
- Relación **categórica-categórica** –> Mosaico/balloonPlot. Tabla de contingencia/Chi cuadrado/V de cramer.

Un dato atípico es una observación que es **numéricamente distante** del resto de los datos.

El problema de los datos atípicos es que tienen una **gran influencia** en los resultados, si los modelos no son **robustos**.

Algunos métodos de detección de outliers son:

- **Desviación típica**: Se considerán atípicos aquellos datos que disten más de un número k (habitualmente entre 3 y 6) de desviaciones típicas de la media. Este método sólo es válido si las distribuciones son aproximadamente **simétricas** (coeficiente de asimetría entre -1 y 1).
- **MAD**: Se considerán atípicos aquellos datos que disten más de un número k (la literatura recomienda) de MADs de la mediana. *Median Absolute Deviation* (MAD) es la mediana de las distancias absolutas a la mediana. Este método es más adecuado para distribuciones **asimétricas**, pero no es válido cuando la mediana es igual a 0.
- **Rango intercuartílico**: Se consideran atípicas aquellas observaciones que se alejen más de 1.5 o 3 veces el rango intercuartílico del primer y el tercer **cuartil**. Este método está asociado a los gráficos de cajas.

Es importante destacar que, por definición, el número de datos atípicos ha de ser pequeño pues, de lo contrario, no se podrán considerar atípicos.

Dado que ninguno de los métodos de detección de atípicos es *perfecto*, lo recomendable es **utilizar varios** de ellos y sólo considerar como atípicas aquellas observaciones que sean consideradas como tal por **más de un método**.

- **Evaluar incidencia**: conocer la **proporción de datos atípicos** encontrados en cada variable.
- **Acciones**:
  - **Eliminación**: Eliminar del conjunto de datos toda observación con al menos un outlier en al menos una variable. Potencialmente puede perderse una cantidad significativa de información válida.
  - **Conversión a valor perdido**: Es habitual convertir los outliers encontrados en valores perdidos para gestionarlos posteriormente mediante algún método de imputación. Dependiendo del tipo de imputación escogida, podría cambiar significativamente la distribución de las variables.
  - **Winsorize**: Reemplazar los valores extremos por valores correspondientes a algún percentil de la distribución (95%). Se podría producir carga excesiva en un mismo valor de las colas de la distribución.

La presencia de datos faltantes en los conjuntos de datos ha de ser **analizada** pues da lugar a una reducción del número de **observaciones válidas** para los procedimientos estadísticos.

Adicionalmente, si la presencia de missings no es **aleatoria,** esta puede venir acompañada de **sesgo** en las respuestas y, por tanto, en los modelos (por ejemplo, encuestados que se nieguen a responder preguntas delicadas debido a su respuesta).

Los paquetes estadísticos más frecuentes **ignoran** la información procedente de observaciones con algún dato faltante (case deletion), por lo que la información que contienen **no se tiene en cuenta** en el modelo.

No obstante, algunas técnicas estadísticas (como los árboles o las técnicas basadas en árboles) modelizan los missings como **una categoría más**, incluyendo así sus posibles efectos predictivos y **reduciendo** considerablemente el sesgo.

Por tanto, habrá que **analizar detalladamente** los datos faltantes y aplicar una o más de las estrategias siguientes:

- **Eliminación**: tanto de variables como de observaciones.
- **Recategorización**: de los valores missing como una categoría válida. Para variables contínuas, esto implica una discretización de la misma.
- **Imputación**: es el proceso que consiste en sustituir los missing por valores válidos.

ELIMINACIÓN

La eliminación de variables y/o observaciones sólo debe llevarse a cabo cuando la proporción de missings sea muy **elevada** (superior al 50%) y, por tanto, la **pérdida de información** no lo sea tanto.

Por ese motivo, primero se lleva a cabo es un análisis de la proporción de missings:

- Por variables: Evaluar la proporción de datos atípicos por columnas, o incidencia por variable. ¿Cuáles son las variables que mayor carga de missing presentan?
- Por observación: En ocasiones existen **observaciones** con un gran número de missings (por ejemplo, por pereza del encuestado) que aportan **poca información** y que, por tanto, pueden ser eliminados.
  Para saber el número de **missings por observación**, debemos crear una variable que los cuente para, a continuación, utilizar dicha información para eliminar las observaciones “conflictivas”.

Un tentativo nuevo conjunto de datos libre de observaciones y variables con elevada incidencia de missings se podría obtener mediante filtrado por filas y columnas.

RECATEGORIZACIÓN

Cuando la presencia de missing sea **importante**, puede ser interesante definir una **nueva categoría** "Missing" de la variable (sobretodo para variables categóricas). Para las variables de intervalo, habría que **discretizarlas** antes y crear una categoría "Missing" para la nueva variable. Hay que tener cuidado con esta alternativa pues no siempre es buena idea categorizar variables numéricas.

Una **alternativa** a caballo entre la recategorización y la imputación consiste en **imputar** los valores faltantes y **crear una variable** nueva que cuente el número de variables imputadas para cada observación (alternativamente, se puede crear una por variable imputada). De esta forma, se puede trabajar con **todas** las observaciones (pues ya son "válidas") y se mantiene la información asociada a la **falta de respuesta** (es una opción especialmente interesante para variables de intervalo con una proporción de missings elevada). La variable "prop_missings" ya creada puede servir para esta función.

IMPUTACIÓN

La imputación consiste en sustituir los datos faltantes por **valores válidos**.

Existen varios tipos de imputación:

- Imputación simple:
  a) Se sustituyen los missings por algún **estadístico de localización**, como la media o la mediana para variables de intervalo, o la moda para variables categóricas. La principal desventaja es la carga excesiva de la categoría modal (caso nominal) o la subestimación de la verdadera varianza de la distribución (caso continuo)
  b) Se imputan los datos faltantes **aleatoriamente** teniendo en cuenta la **distribución** de la variable. La principal desventaja reside en la supeditación a la aleatoriedad que podría jugar malas pasadas en alguna realización (recordemos que la imputación aleatoria, de manera natural cambia en cada ejecución si no fijamos semilla de inicialización)
- Imputación por modelos: se sustituyen los missings por una predicción basada en **otras variables** del conjunto de datos. Destacan la imputación por modelos de Regresión, Cadenas de Markov, Random Forest, Knn, etc. Como principal desventaja destacamos el sobreajuste a los datos de training que puede tener lugar (los modelos potentes pueden "aprender demasiado" de los patrones del training).

En ocasiones es necesario realizar alguna **transformación** en las variables para que el modelo de **predicción** funcione mejor o se pueda plasmar la verdadera **relación** con la variable objetivo.

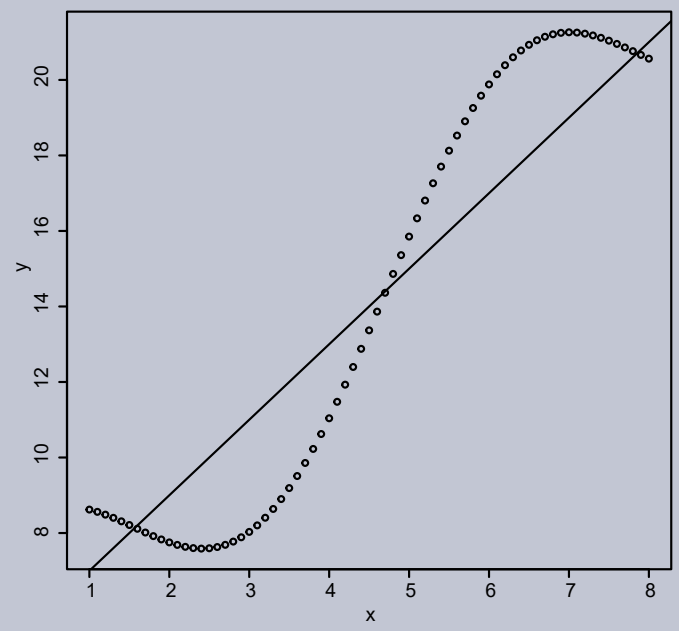
DISCRETIZACIÓN DE VARIABLES CONTINUAS

La **discretización** (en inglés, binning) permite descubrir **relaciones complejas** (no lineales) entre las variables de entrada y la variable objetivo.

Consiste en obtener una variable categórica a partir de la **división** de una variable continua.

Se persigue obtener tramos lo más **heterogéneos** posible con respecto a la variable objetivo. De esa forma, pertenecer a uno u otro grupo ofrecerá información acerca de la variable a predecir.

Un método que suele dar buen resultado es la tramificación por **Árboles de decisión**.



MEJORAR LA RELACIÓN CON LA VARIABLE OBJETIVO

En ocasiones es preferible *linealizar* la relación entre la variable objetivo y las input **cuantitativas** para que el modelo aproveche mejor el poder predictivo de éstas. Para ello, podemos transformar dichas variables eligiendo entre varias transformaciones aquella que maximice dicha relación, ya sea en términos del coeficiente de correlación (si la variable objetivo es cuantitativa) o de la V de Cramer (si la variable objetivo es cualitativa).

Las transformaciones típicas para conseguir linealidad frente a la variable respuesta (caso de la regresión lineal) o frente al logit del modelo (caso de regresión logística) son:
$X, log(X), e^X, X^2, \sqrt{X}, X^4, \sqrt[4]{X}$.

Una vez obtenidas las transformaciones, se puede elegir entre mantener las variables originales o rechazarlas.

En ocasiones, frente a variables asimétricas que puedan mejorar la linealidad frente a la respuesta mediante transformación de este tipo, es conveniente retrasar la evaluación de los outliers para hacerlo sobre la variable ya transformada.

A parte de los gráficos que ya se han generado, una buena forma de evaluar el poder predictivo de las variables input es a través del estadístico "V de Cramer". Este estadístico se calcula a partir del estadístico χ^2 y tiene la ventaja de que es capaz de capturar las relaciones no lineales existentes entre las variables.

ESTADÍSTICO χ^2

El estadístico χ^2 es un método muy común para detectar **relaciones** entre dos variables, siendo especialmente útil para relaciones **no lineales**.

Los datos de entrada requeridos para calcular este estadístico consisten en una **tabla de contingencia**, por lo que las variables de intervalo han de ser **discretizadas** previamente.

Sean C_i y D_j las clases en las que están divididas las variables de interes, entonces:

	C_1	C_2	$\dots$	C_l	Total
D_1	n_{11}	n_{12}	$\dots$	n_{1l}	$n_{1.}$
D_2	n_{21}	n_{22}	$\dots$	n_{2l}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
D_k	n_{k1}	n_{k2}	$\dots$	n_{kl}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.l}$	n

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

ESTADÍSTICO χ^2

El estadístico χ^2 toma valor 0 cuando las variables son **independientes** (no hay relación entre ellas). Matemáticamente esto equivale a decir que las frecuencias relativas de los cruces de categorías **sólo dependen** de las frecuencias relativas de las categorías:

$$f_{ij} = f_{i\cdot} f_{\cdot j}, \text{ donde } f_{ij} = \frac{n_{ij}}{n} \quad f_{i\cdot} = \frac{n_{i\cdot}}{n} \quad f_{\cdot j} = \frac{n_{\cdot j}}{n}$$

Por lo tanto, las variables que estén fuertemente relacionadas tendrán un estadístico χ^2 **alto**. Sin embargo, este estadístico no está limitado por lo que resulta complicado determinar cuando está tomando un valor significativamente alto.

ESTADÍSTICO V DE CRAMER

El estadístico V de Cramer está basado en el **estadístico** χ^2 pero tiene la ventaja de que su valor está **acotado** entre 0 y 1. Viene dado por:

$$V = \sqrt{\frac{\chi^2}{n \times \min(l-1, k-1)}}$$

El estadístico V de Cramer hace comparable la cuantificación de relaciones en tablas de contingencia de distinto tamaño.

Crear una variable aleatoria permite tener una referencia de la utilidad de las variables.