



Fastbook 10

Tratamiento de Datos (Excel y SQL)

Análisis de la información
avanzada



10. Análisis de la información avanzada

Hasta ahora, hemos visto los principales elementos que debemos tener en cuenta a la hora de validar nuestros datos según su tipología, pero en la mayoría de las ocasiones estas validaciones, a pesar de ser necesarias, no son suficientes.

Como hemos visto, los datos no son nada sin el contexto y las relaciones presentes entre ellos, es decir, si os proporciono la serie {38,38,40,29,33,37,36}, difícilmente podréis entender a lo que me estoy refiriendo; en cambio, si agrego que es la temperatura media que se ha producido en Madrid en la última semana de junio, tendremos todo el contexto de información necesaria para entender la información en su plenitud.

Del mismo modo, a la hora de realizar la validación de nuestra información, deberemos prestar atención a las relaciones que existen entre los datos para poder asegurar su validez.

Por eso, en este fastbook, veremos las diferentes fases que tendremos que seguir para asegurarnos, en primer lugar, de que nuestros datos son correctos y, en segundo lugar, de que podemos empezar a trabajar con ellos de forma segura.

Autor: Breogán Cid

Etapas de la validación de información

Validaciones estadísticas

Validaciones de integridad de la información

Conclusiones

Validaciones mediante reglas de negocio

Etapas de la validación de información

X Edix Educación

Para una correcta validación de la información presente en nuestro proyecto, seguiremos una serie de etapas para **llevar un control y un orden** que nos posibiliten realizar el proceso de una forma eficaz.

Validaciones de los procesos de ingestión de información

Todo proyecto analítico parte de la **recolección de la información de una o varias fuentes de datos** que pueden estar almacenados de distintas maneras como, por ejemplo:

- Datalakes de información.
- Bases de datos.
- Ficheros almacenados en nuestro ordenador o en SFTPS.
- Webs desde las que tengamos que descargarnos nuestra información.
- APIs de proveedores.
- Archivos propios de ciertos lenguajes o programas específicos.

Pero independientemente de nuestras fuentes de datos, para empezar a trabajar con la información tendremos que **recopilarlos y almacenarlos en un único sistema**, obteniendo nuestro catálogo de información.

Los **procesos de ingestá** pueden ser diversos y estar realizados mediante el uso de diferentes lenguajes.

En la mayoría de los casos, además, serán **realizados por otros compañeros de distinto rol en la empresa** o no tendremos ni la formación ni el tiempo necesario para realizar las validaciones necesarias que nos aseguren que todos los procesos han sido realizados de la forma correcta.

Por ello, los trataremos como **elementos de una black box**, en los que solo tendremos en cuenta la entrada (fuentes de datos originales) y su salida (nuestra información almacenada en nuestro sistema).

Por eso, la primera validación, una de las más importantes, será la que asegure que **la información presente en la base de datos se corresponde con la presente en las bases de datos originales**.

En este punto del proceso, no tendremos en cuenta si la información es correcta o no, **solo deberemos validar que es la misma que teníamos en un origen**, ya que, pasados de este punto, perderemos todo contacto con las fuentes de datos originales.

Dentro de esta etapa nos centraremos en intentar validar dos aspectos fundamentales de cada fuente de información, **el volumen y el formato**, es decir, analizaremos las tablas de manera independiente, haciendo una primera validación sin considerar el resto de las tablas obtenidas en nuestro sistema.

1

Formato

La primera validación que tendremos que realizar, y que posiblemente sea una de las más fáciles de llevar a cabo, es que la estructura de nuestras nuevas tablas (da igual que se trate de tablas de base de datos, de ficheros de texto o de estructuras dentro de nuestro lenguaje) se corresponde con lo esperado, es decir, que tenemos **las mismas columnas** y que cada una de estas tiene el tipo esperado.

Una vez que realicemos esta **primera validación**, deberemos llevar a cabo todas las validaciones que hemos visto en el fastbook anterior: aquellas que nos aseguran la calidad de cada una de las variables en función de su tipo.

Recordad que tendremos que prestar especial atención a las variables de tipo fecha y a la codificación de los campos de tipo texto, para asegurar que no presentan caracteres raros por problemas de codificación.

2

Volumen

Una vez que hemos asegurado que los datos presentes tienen el formato esperado y que son correctos según su tipología, pasaremos a comprobar si estos son todos los que tendrían que estar o si hemos perdido información relevante dentro de los procesos de ingesta.

Para ello, existen **pequeños truquillos o funciones comunes** que nos permitirán comprobarlo.
¡Vamos a verlos!

Obtener el número de registros totales

Una de las primeras validaciones que tendremos que realizar es que el número de registros totales se corresponde con los presentes en las fuentes originales, puesto que, si en este punto no coinciden, ya será un indicador suficiente de que no disponemos de toda la información necesaria, aunque seguiremos realizando el resto de las validaciones para poder indicar qué información nos falta.

Comprobación de valores aleatorios

Un error habitual es suponer que, si los números coinciden a nivel global, podemos suponer que los procesos de ingesta han sido correctos, y aunque en la mayoría de los casos es así tendremos que proseguir con nuestro análisis. La siguiente validación que debemos realizar es la comprobación de ciertos registros aleatorios. Para ello, elegiremos algún campo que nos permita representar la información de forma única (la clave primaria, si estamos trabajando con bases de datos) y validaremos que la información es exactamente la misma en las dos fuentes.

Comprobación de valores vacíos o no informados

También es necesario comprobar que la cantidad de información no informada en nuestros datos se corresponde con la que teníamos presente en nuestra base original.

Realización de agrupaciones

Este proceso será muy parecido al realizado en el primer punto, cuando estábamos obteniendo el número de registros totales, pero en esta fase realizaremos las validaciones en función de diversas agregaciones o filtrados de información. Si estamos trabajando en Excel, en alguna herramienta de visualización o de BI, podremos recurrir a la creación de tablas dinámicas para realizar estas validaciones. Pero si estamos trabajando en bases de datos, recurriremos a las sentencias de agrupamiento. No hará falta realizar todas las agregaciones posibles y, en el caso de tener muchos elementos categóricos, podemos realizar filtrados para reducir la muestra analizada.

Validación de funciones complejas

Como último elemento de validación, realizaremos un par de sentencias que, aunque no tengan sentido de negocio, pueden ayudarnos a validar los datos de una manera más transversal. La idea es aplicar fórmulas matemáticas sencillas que relacionen distintas columnas de nuestros datos para obtener un único valor que podamos comparar.

Imaginemos que disponemos de la siguiente **información en nuestros datos**:

- ID.
- Nombre.
- Edad.
- Género.
- DNI.
- Código postal.

Una posible función de validación puede ser la validación de que la $\Sigma(\text{código postal} + \text{Id} + \text{Edad})$ es el mismo en las dos fuentes. Está claro que el resultado no tiene sentido, pero nos sirve para **validar la relación de los tres campos de forma fácil** y, si el resultado coincide, podremos asegurar que los datos son idénticos a los originales.

Validaciones de integridad de la información

X Edix Educación

Una vez que hemos validado que la información es correcta a nivel de tabla de nuestro sistema de datos, empezaremos a **validar las relaciones presentes** entre las distintas tablas y fuentes de información obtenidas.

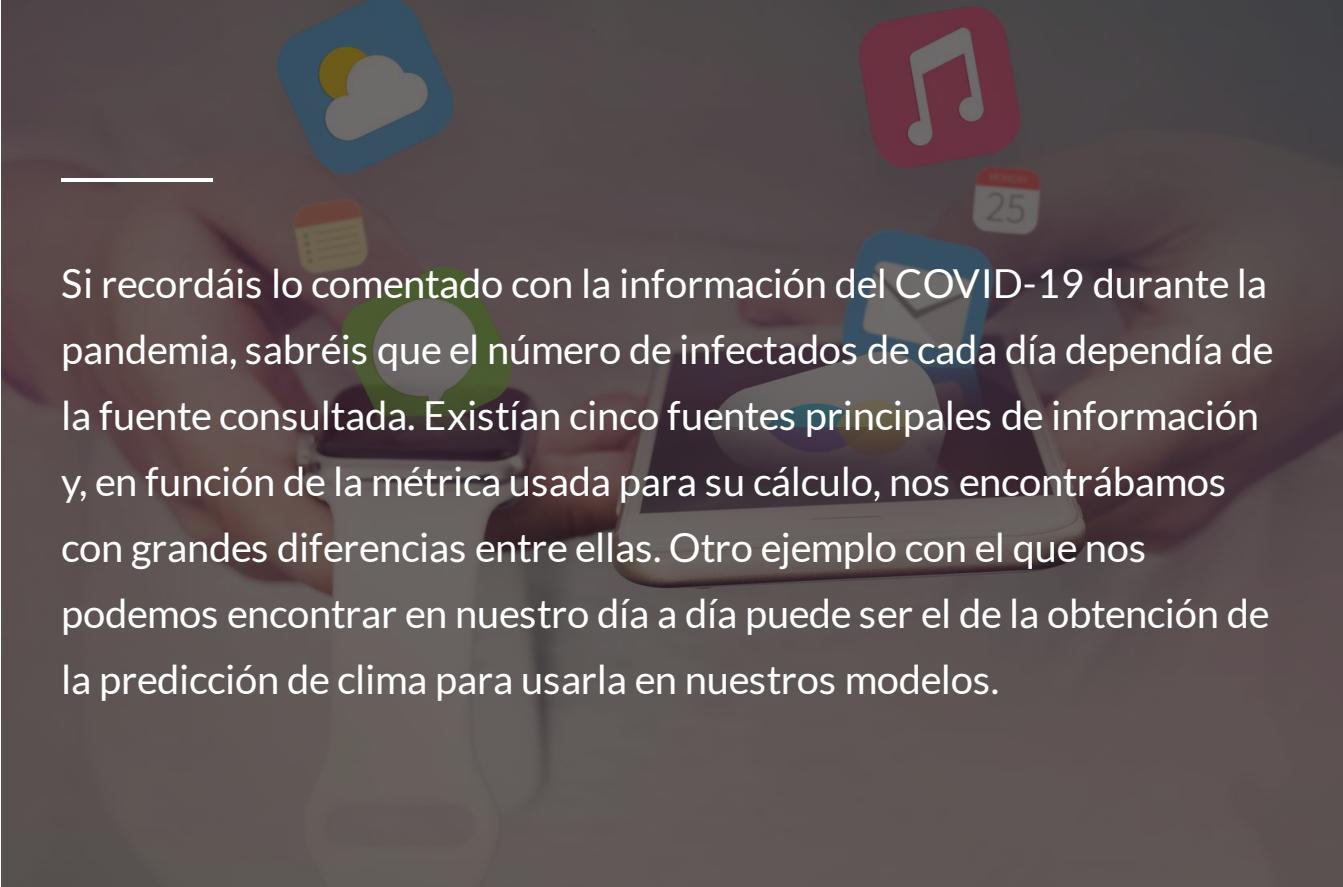
Si estamos trabajando con bases de datos, este paso será más ágil, ya que muchas de **estas validaciones serán automáticas porque nuestros SGBD las harán por nosotros**. Si este es nuestro caso, lo primero que tendremos que comprobar es que todas nuestras tablas se hayan creado con las distintas restricciones necesarias, tal como vimos en los fastbooks referentes a bases de datos.

Cuando estamos trabajando con diferentes fuentes de información, es habitual que cada una de las entidades de origen trabajen con sus propios diccionarios o nomenclatura.

Por ejemplo, si trabajamos con los **datos obtenidos del INE y del EUROSTAT** y los obtenemos a nivel de provincia, deberemos asegurarnos de que usan la misma nomenclatura y, de no ser así, tendremos que crear nuestro propio diccionario de traducciones que nos permita relacionar las dos fuentes.

Este paso es **uno de los más complicados**, ya que dependerá siempre del tipo de información con la que estemos trabajando. Y muchas veces nos encontraremos con casos que no podamos resolver de forma fácil, en los que tengamos que recurrir a métodos más complejos para unir nuestras fuentes de información, por ejemplo, imaginemos que tenemos información socioeconómica obtenida a nivel de sección censal y otra obtenida a nivel provincia, al tratarse de distintos elementos de repartición del territorio español no podremos enlazarlas de forma directa.

Además, cuando estamos **trabajando con información de distintas fuentes**, una situación muy habitual que nos podemos encontrar es que, al obtener la misma información de cada una de las fuentes, el resultado no sea exactamente el mismo.



Si recordáis lo comentado con la información del COVID-19 durante la pandemia, sabréis que el número de infectados de cada día dependía de la fuente consultada. Existían cinco fuentes principales de información y, en función de la métrica usada para su cálculo, nos encontrábamos con grandes diferencias entre ellas. Otro ejemplo con el que nos podemos encontrar en nuestro día a día puede ser el de la obtención de la predicción de clima para usarla en nuestros modelos.

En función de la fuente consultada, esta predicción puede variar significativamente, y esto **no quiere decir que la información sea incorrecta**, ya que hasta que no pasa el tiempo y podemos comprobar cuál es la que más acierta todas serían correctas.

En este punto siempre me gusta citar a uno de mis profesores de la universidad:

“Si disponemos de una única fuente de información, podremos tener dudas de su validez; cuando trabajemos con varias fuentes que nos informen de la misma medida, nunca estaremos seguros.”

- Avani Sadana

Lo más importante que debemos sacar en claro es que, si disponemos de la misma medida obtenida de distintas fuentes de información, lo ideal sería **validar que siempre coincidan**; pero este hecho muchas veces no se produce, por lo que tendremos que tomar una de ellas como principal y usar solo esa en el proyecto.

Por último, y antes de pasar a la siguiente etapa del proceso, nos queda empezar a comprobar la falta de datos esperados. Es una de las validaciones más complejas y que suele ser de las más olvidadas.

Imaginemos que tenemos una serie temporal que nos indica las ventas de un conjunto de productos en los distintos establecimientos de un supermercado para cada uno de los días de estos últimos 5 años.



¿Qué validaciones creéis que deberíamos realizar para asegurarnos de que tenemos la información completa?

A continuación, te muestro algunas de las **preguntas** que nos tendríamos que hacer:

- ¿Tenemos información para cada uno de los 5 años de histórico?
- ¿Para cuántos días distintos tengo información en cada año?
- ¿Para cuántos productos distintos tengo información en cada uno de los años?
- ¿Para cuántas tiendas distintas tengo información en cada año?
- ¿Existen tiendas que no vendan algún producto?
- ¿Cuál es la fecha mínima y máxima de venta de cada producto en cada tienda?
- ¿Existen días para los que no tenga información de algún producto?
- ¿Existen días para los que no tenga información de alguna tienda?
- ¿Existen días para los que no tenga información de algún producto en alguna tienda?

Alguna de estas preguntas, pisan fuertemente con las que realizaremos en el siguiente **paso de las validaciones** (validación con reglas de negocio), pero no está mal detectar esos periodos o conjuntos de los que no disponemos información para poder tenerlos presente.

Validaciones mediante reglas de negocio

X Edix Educación

Este paso es uno de los más complicados, ya que para realizarlo correctamente necesitaremos un conocimiento específico sobre el trabajo de nuestros clientes, que muy difícilmente será comparable al que tiene nuestro cliente. Por eso, este tipo de validaciones suele realizarse de la mano de una persona más orientada al negocio, que puede ser proporcionada tanto **por el equipo del cliente como por nuestra empresa**. A medida que vayáis ganando experiencia en proyectos del mismo sector, estas validaciones se os irán haciendo más fáciles y rápidas, pero, a pesar de ello, siempre os será de mucha ayuda el asesoramiento del personal experto.

En estas validaciones trataremos de verificar que la información obtenida en nuestras variables se corresponde con los valores esperados según **nuestra experiencia y raciocinio**. Comparto un ejemplo para poder entenderlo con facilidad: supongamos que en nuestros datos tenemos la serie de temperatura media por día en España de los últimos años. Si hemos vivido en el país podremos identificar valores erróneos a simple vista, ya que podremos indicar que 60°C es claramente un valor que no se ha producido jamás. Pero muchas veces esta validación no es tan directa.

En uno de los últimos proyectos en los que he estado involucrado, que se realizó para uno de los mayores bancos mundiales según su capitalización bursátil, me encontré en la **tesitura de tener que validar la información de algunos KPIs complejos**, propios del cliente. A pesar de tener experiencia en otros proyectos en este sector y de haber realizado otros trabajos para el mismo cliente, me resultó muy complicado realizar esta validación, y tuve que reunirme varias veces con distintos responsables de diferentes departamentos para estar seguro de que los datos eran los correctos.

Estas validaciones variarán mucho en función del tipo de datos con los que estemos, por eso, es importante entender en qué consisten los principales KPIs del cliente y los rangos de valores que pueden tener. Lo importante es empezar con los más sencillos para poder ir utilizando esta **información en las métricas más complejas**.

Supongamos que estamos trabajando para una empresa multinacional, como puede ser Inditex, y necesitamos validar la información de la **ganancia total** producida por sus ventas en España. ¿Seríamos capaces? ¿Por dónde podríamos empezar? Si tienes clara la respuesta, es que has comprendido la lección más importante de este paso.

**No tengas miedo de preguntar al especialista de negocio, ya que
es el que más sabe del tema específico.**

Pero sin recurrir a esta fuente, ¿cómo podríamos estimar un rango válido para este KPI?

Lo importante es **entender el KPI e ir validando la información necesaria para calcularla**; podemos suponer que la ganancia total de la compañía se calculará como la suma de las ganancias de cada una de sus tiendas. Por lo que tendríamos que empezar a validar...

- El número de tiendas abiertas en España por día.
- La ganancia media de cada tienda en España por día.
- Que la suma de las ganancias de cada tienda se corresponde con la ganancia total (parece obvia, pero en las validaciones no podemos dar nada por sentado).

Al ir **reduciendo la complejidad de los KPIs**, nos vamos acercando a valores más comprensibles y cercanos a nuestro día a día, lo que nos facilitará su validación, aunque de todas formas lo recomendable será recurrir igualmente al experto para validar estas nuevas métricas.

Además, como **elemento diferenciador y que nos ensalza como buenos analistas de información**, tendremos que validar algo que a mí me gusta llamar los sucesos imposibles. Es decir, debemos analizar **la información existente buscando elementos que no tengan sentido**; de hecho, este tipo de validaciones se suele usar tanto para hallar valores erróneos como para la detección de fraudes o robos de identidad.

Imaginemos que disponemos de la información de los usuarios de una compañía de telecomunicaciones, en concreto, las llamadas que realiza cada uno y los siguientes datos al respecto:

- Identificador del usuario.
- Número de teléfono del usuario.
- Número de teléfono del destinatario.
- Fecha.
- Duración de la llamada.
- Geolocalización del usuario.
- País destino de la llamada.

¿Qué validaciones podríamos realizar con esta información?

Aquí tienes un **listado de algunas de las preguntas** que nos podríamos hacer para empezar:

- 1 ¿Existen varias llamadas que se han producido por el mismo usuario simultáneamente?
- 2 ¿La geolocalización del usuario ha cambiado drásticamente en dos llamadas muy cercanas en tiempo?
- 3 En los casos en los que el país destino sea distinto al habitual, ¿ya había realizado llamadas a ese país?

4

¿Existen muchos usuarios que llamen al mismo número de teléfono de destino?

5

¿Existen llamadas fuera de su franja de llamadas habitual?

6

¿Existen llamadas de mayor duración de lo habitual?

7

¿Existen grandes periodos sin realizar ninguna llamada?

Como podemos observar, la cantidad de posibles preguntas a resolver con los datos puede ser enorme, y no tienen por qué ser un indicativo de un error en los datos, pero sí **un buen indicativo de realizar un análisis en detalle** de esas llamadas.

Lo importante en este tipo de validación es buscar valores que se salgan de lo común o que nos llamen la atención.

Validaciones estadísticas

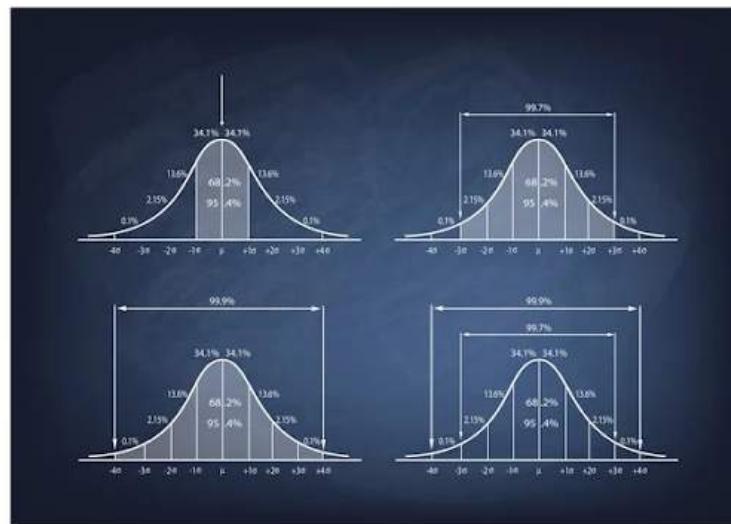
X Edix Educación

Por último, tenemos el último paso del proceso de validación, aunque habitualmente se puede incluir en los procesos de tratamiento de la información. La idea se basa en que, si no disponemos del conocimiento de negocio necesario para validar la información como correcta o incorrecta, **podemos aplicar la estadística para comparar cada uno de los datos con el resto de los del catálogo**, identificar los valores que se salen de los patrones lógicos y definir nuestros valores atípicos como los mejores candidatos a ser los datos erróneos de nuestro modelo.

En función de la necesidad, podemos ir aumentando la complejidad de nuestros análisis, pero en la mayoría de los casos con el uso de funciones de analítica básica será más que suficiente.

Podemos empezar con el análisis de la diferencia de cada medida versus su media y su mediana. Gracias a esto, obtendremos valores atípicos que pueden ser los principales candidatos a una evaluación en detalle.

Si queremos ser menos exigentes a la hora de encontrar esos posibles valores erróneos, podemos trabajar con los extremos de los valores de los percentiles, ya que muchas veces podremos asumir que, aunque **nuestros datos no se distribuyan como una normal**, los valores más factibles de ser producidos por fuentes erróneas se situarán en las colas de la distribución.



Por último, y como elementos poco usados pero que pueden resultar útiles en ciertas ocasiones, tenemos la realización de ciertos modelos de clusterización de variables o de predicción sobre alguno de los principales KPIs para obtener, de forma matemática, las distintas relaciones entre las variables, **analizando los posibles errores de predicción**. O podemos encontrar los elementos peor segmentados (según la distancia a los centroides de los clusters) para obtener los mejores candidatos a los análisis en detalle. Este método tiene como desventaja que los modelos creados no son perfectos, por lo que los resultados ocasionarán cierto ruido, pero cabe recordar que en ningún momento indicaremos que son valores erróneos, sino que son los mejores candidatos a que puedan ser erróneos, siendo siempre necesaria una validación manual de los datos seleccionados.

Conclusiones

X Edix Educación

A modo de breve resumen de lo comentado en este fastbook, me gustaría destacar **siete aprendizajes** de este tema.

- La validación es un proceso difícil, por lo tanto, es necesario seguir una serie de pasos para facilitar y agilizar el proceso.
- En nuestros proyectos, podremos trabajar con distintas fuentes de información, por lo que una de nuestras primeras tareas será **unir toda la información en un mismo sistema** para poder trabajar con ella (a ser posible en una base de datos o plataforma Big Data equivalente, en caso de necesidad).
- Es importante realizar validaciones de los procesos de ingesta para asegurar que trabajamos con toda la información disponible, pero en lugar de analizar los procesos realizados, **analizaremos los datos origen y destino** (siempre que sea posible).
- Una vez que tengamos la certeza de que los datos recopilados son los mismos que teníamos en las fuentes de origen, empezaremos a realizar las validaciones de integridad, para asegurar que las relaciones entre las distintas fuentes son las correctas.
- También tendremos que aplicar la **inteligencia de negocio** para validar el contenido de nuestra información. Como muchas veces no dispondremos del conocimiento necesario, lo recomendable es acudir a un experto del cliente para que nos dé la aprobación y nos ayude a dar el último ok a los datos.

- Como elemento adicional, podemos **aplicar la estadística básica** para suplir el conocimiento experto, analizando los datos y enfrentándolos al resto de la información disponible.
- El proceso de validación es tan complicado que difícilmente podremos asegurar nunca que los datos son correctos al 100%, ya que en la mayoría de los casos siempre encontraremos nuevas validaciones que realizar. Lo importante es **obtener la mayor certeza de validez de la información dentro del tiempo razonable**.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers