



Fastbook 04

Estadística Aplicada al Marketing

Transformaciones de datos



04. Transformaciones de datos

En los dos primeros fastbook nos familiarizamos con el ecosistema de datos de los departamentos de marketing. Con el tercer fastbook conocimos las principales métricas que nos ayudarán a medir nuestro negocio. Pero, ¿los datos vienen ya listos para su explotación? La mayoría de las veces la respuesta es no.

Nuestro objetivo en este cuarto fastbook es conocer las principales técnicas de tratamiento de datos que nos ayuden a consolidar la información en una única base de datos estructurada que permita su uso y análisis. Descubriremos los principales problemas a los que nos vamos a enfrentar, conoceremos distintas formas para resolverlos según el problema y la disposición de los datos, y entenderemos el concepto de ad-stock y para qué sirve.

Autora: Patricia Martín González

Tratamiento de datos a nivel temporal

El efecto ad-stock

Resumen

Test

Tratamiento de datos a nivel temporal

X Edix Educación

Hasta ahora podría parecer que los datos no iban a darnos problemas, pero no siempre es así.

Los datos no dicen nada por sí solos si no son tratados y analizados.

De hecho, podría atreverme a decir que la recopilación y tratamiento de los datos para crear una base de datos única y consistente va a ser uno de los mayores retos a los que nos vamos a enfrentar.

Missing values

En ocasiones puede haber valores que nos faltan, es decir, valores vacíos o missing values como les solemos llamar. Por ejemplo: en la serie de histórico de impresiones de SEM, solamente hay un día con missing. Otro ejemplo: el departamento de marketing nos ha compartido los datos de Facebook, pero hay algunos períodos que no tenemos información.

Carencia de histórico

No todas las fuentes que vamos a integrar disponen del mismo histórico, bien porque no ha estado activo en ese periodo (por ejemplo, un medio publicitario o la competencia) o porque no podemos recuperar la información (por ejemplo, datos de third party con menor histórico del que necesitamos).

Granularidad temporal alta

Otra posible situación es que no dispongamos de la granularidad temporal suficiente, es decir, el nivel de agregación de la serie es superior al que necesitamos. Por ejemplo: necesitamos la serie de GRPs de televisión de forma semanal, pero solamente la tenemos mensual.

Granularidad temporal baja

El caso contrario, tenemos los datos a un nivel de agregación inferior al deseado. Por ejemplo: tenemos la serie diaria de clics de display y la necesitamos a nivel semanal.

¿Cómo nos enfrentamos a estas situaciones? ¿Cómo las resolvemos?

1

Missing values

Lo primero que debemos hacer cuando nos enfrentamos a missing values es preguntarnos a qué se deben. En múltiples ocasiones la respuesta es muy fácil: **falta de ‘actividad’ en ese periodo**. En otras, puede ser por caída en el servidor, imposibilidad de recuperar los datos, etc.

Recordemos los dos ejemplos anteriores:

- Sobre la **serie de histórico de impresiones de SEM** hay un día con missing.
- El **departamento de marketing** nos ha compartido los datos de Facebook, pero hay algunos periodos que no tenemos información.

¿Podrías detectar cuál es cuál sobre las siguientes dos categorías?

Missing value = 0



Hay periodos en que no hay actividad y la forma de almacenar la información o la forma en la que recibimos los datos es con missing values. En este caso, estos valores suelen ser sustituidos por 0. Sobre el ejemplo anterior, estaríamos en el caso de Facebook: ha habido algunos períodos temporales en los que no se ha invertido, por lo que los missing values son 0.

Missing value = falta de información



Hay situaciones que provocan missing values en la serie de datos sin posibilidad de recuperar la información. Aunque no suele ser muy frecuente, la forma de resolverlo no es trivial y dependerá del medio, la duración del periodo en blanco, la causa, etc. Sobre el ejemplo anterior, el caso de SEM sería un ejemplo de falta de información: preguntando al departamento de marketing, nos informan que hubo una caída en el servidor y no se guardaron los datos por lo que resulta imposible recuperarlos. En este caso, imputaremos el promedio de los valores más cercanos.

2

Carencia de histórico

Al integrar diversas fuentes de distinta procedencia en una misma base de datos, no siempre vamos a poder tener el mismo histórico en todas ellas. Las **causas pueden ser variadas**: no se guardaba la información, no se puede recuperar, se requiere demasiado tiempo para la extracción de la información y no disponemos de él, etc.

Cuando nos encontramos ante este problema nos enfrentamos a **tres alternativas**:



Limitar el histórico del periodo de estudio. Aunque es la solución más rápida, conviene valorar la información estamos perdiendo por incluir las variables con menor histórico, ya que en ocasiones podemos perder información muy valiosa. Antes de acortar el histórico es necesario valorar si la información del periodo que vamos a recortar aporta (o no) valor, si nuestras series se comportan igual que a lo largo del tiempo, si ha habido cambios estratégicos en la compañía y es mejor coger períodos más recientes, etc.

Imputación. Imputar dicha información con la métrica más adecuada. Este suele ser el caso cuando el histórico a imputar no es mucho, ya que no compromete la fiabilidad y calidad de la información. Las técnicas que se suelen aplicar para estos casos son: replicar el mismo periodo de los años siguientes, o un promedio de los mismos, o estimar con algún modelo sencillo (por ejemplo, regresión lineal).



Desechar las variables con menos histórico. Se desechan las variables con menos histórico ya que no aportan mucho valor (la estimación resulta muy tediosa si queremos que sea robusta).

3

Granularidad temporal de mayor a menor

Cuando estamos integrando diversas fuentes en una misma base de datos, es altamente probable que el nivel de agregación temporal de las fuentes sea diferente. Estaremos en este caso cuando los datos tienen mayor granularidad (por ejemplo, mensual) de lo que necesitamos (semanal).

No todas las fuentes se miden ni se almacenan con la misma periodicidad.

Este suele ser el caso de los datos externos, y cada vez menos frecuente de los datos de medios. Nuestro cometido es **tratar las variables para poder crear la base de datos** al nivel que deseamos, y no tener que adecuarnos al nivel de agregación de los datos, aunque en algunos casos no quede más remedio. Generalmente, podremos solucionar esta situación mediante **dos técnicas** en función del problema:

- **Interpolación**

Usaremos la interpolación cuando la variable es un indicador, por ejemplo: notoriedad de marca, precio, tasa de paro, PIB, etc. La interpolación es una técnica estadística que consiste en calcular nuevos puntos a partir de otros conocidos. Aunque existen varios métodos, nos centraremos en la **interpolación lineal** que es la más extendida.

La interpolación lineal consiste en encontrar los puntos promedios entre los puntos conocidos. Dicho de otra forma: trazar la línea recta entre los puntos conocidos y calcular los valores intermedios.

Veamos dos ejemplos:



Tenemos un punto **X=0** y un punto **Y=3**. Si necesitamos conocer el punto Z como interpolación de X e Y sería **Z=1,5**.



En caso de tener que calcular dos puntos serían: **Z₁=1**, **Z₂=2**.



Matemáticamente:

Paso 1

Calculamos la longitud del intervalo a dividir.

En nuestro ejemplo sería: $Y-X = 3-0 = 3$.

Paso 2

Calculamos la distancia entre puntos: dividimos esta longitud entre $1 + \{\text{número de puntos a calcular}\}$.

- En el ejemplo a) sería $3/(1+1)=1,5$.
- Para el ejemplo b) sería $3/(1+2)=1$.

Paso 3

Calculamos los puntos. Para ello, sumamos al punto inferior (X) los incrementos.

- En el ejemplo a) el punto Z = X + incremento = 0 + 1,5 = 1,5.
- En el ejemplo b):
 - Z1 = X + incremento = 0 + 1 = 1.
 - Z2 = Z1 + incremento = 1 + 1 = 2.

- **Repartir**

Repartiremos **cuando las variables indiquen volumen**, como por ejemplo: GRPs, clics, impresiones, inversión, etc. La repartición se puede hacer de forma uniforme o de manera más avanzada usando ponderaciones.

Veamos dos ejemplos:

Ejemplo 1 —

Si sabemos que tenemos 35.000 impresiones/semana, pero lo necesitamos de forma diaria, dividiremos los clics entre el número de días, de tal forma que tengamos:

$$35.000 \text{ impresiones}/7 \text{ días} = 5000 \text{ impresiones/día.}$$

Ejemplo 2 —

Tenemos 620 GRPs para el mes de diciembre 2020, pero necesitamos los GRPs por semana. Para hacerlo de forma purista, primero dividiremos entre el número de días: $620 \text{ GRPs}/31 \text{ días} = 20 \text{ GRPs/día.}$

A continuación, asignaremos los días correspondientes a cada semana, de forma que tendríamos:

Semana	Nº de días (semana)	Cálculo GRPs/día * nº días
1-6 diciembre	6	$20\text{GRPs} * 6 = 120\text{GRPs}$
7-13 diciembre	7	$20\text{GRPs} * 7 = 140\text{GRPs}$
14-20 diciembre	7	$20\text{GRPs} * 7 = 140\text{GRPs}$
21-27 diciembre	7	$20\text{GRPs} * 7 = 140\text{GRPs}$
28-31 diciembre	4	$20\text{GRPs} * 4 = 80\text{GRPs}$

Si sumamos los GRPs repartidos, obtenemos los mismos que en total: 620.

4

Granularidad temporal de menor a mayor

Como recordamos del **caso anterior**: no todas las fuentes se miden ni se almacenan con la misma periodicidad. Ahora cubriremos el caso en el que la información la tenemos más desagregada del nivel que necesitamos. Suele ser el caso de los **datos provenientes de clientes** (ecommerce, datos en tiempo real), y de **medios publicitarios** (especialmente los digitales).

Para nosotros, este es el mejor de los dos casos posibles. Gracias al avance de la tecnología, es probable que sea la situación a la que nos vamos a enfrentar con más frecuencia: la **granularidad de la información** que tenemos es mayor de la que deseamos. Conozcamos las **principales técnicas**:

Promedio o media

Usaremos el promedio cuando la variable es un indicador, por ejemplo: notoriedad de marca, precio, NPS, tasa de natalidad, etc. Es el caso opuesto a la interpolación.

Un ejemplo: supongamos que tenemos el precio de venta de la leche a nivel diario y lo necesitamos a nivel mensual. Para ello, haremos un promedio de los valores de cada mes y obtendremos el valor deseado.

Suma

Sumaremos los valores cuando estemos hablando de volumen: inversión, clics, impresiones, GRPs, etc. Este caso es el opuesto al de reparto.

Un ejemplo: supongamos que tenemos los datos semanales de presión de OOH, pero los necesitamos a nivel mensual. Para ello: dividiremos las semanas en 7 días para tener el dato a nivel diario, y sumaremos los días de cada mes.

Semana	Días de diciembre	Inversión en OOH	€/7días
30nov - 6dic	6	1.001€	143€/día
7dic - 13dic	7	560€	80€/día
14dic - 20dic	7	4.984€	712€/día
21dic - 27dic	7	40.005€	5.715€/día
28dic - 3ene2021	4	35.000€	5.000€/día

La **inversión** correspondiente al mes de **diciembre** sería:

$$143*6 \text{ (semana 30/nov)} + 80*7 \text{ (semana 7/dic)} + 712*7 \text{ (sem. 14/dic)} + 5.715*7 \text{ (sem. 21/dic)} + \\ 5.000*4 \text{ (sem. 28/dic)} = 66.407\text{€ en diciembre}$$

Lesson 2 of 4

El efecto ad-stock

X Edix Educación

¿Cuánto dura el efecto de la publicidad? ¿Solamente impacta durante la emisión del anuncio? En caso contrario, ¿cuánto dura?

Quizás sea una pregunta que no te habías planteado antes, pero que a partir de ahora te recomiendo que tengas muy presente cuando quieras medir la eficacia de los medios publicitarios.

El **efecto de la publicidad sobre las ventas** (o la variable que estemos midiendo) no tiene impacto solo inmediato, sino que genera un **efecto en la memoria de los impactados** que se mantiene en las siguientes semanas, meses o años.

¿Años te parece mucho?

El efecto de una buena presión publicitaria puede durar hasta 2 o 3 años.

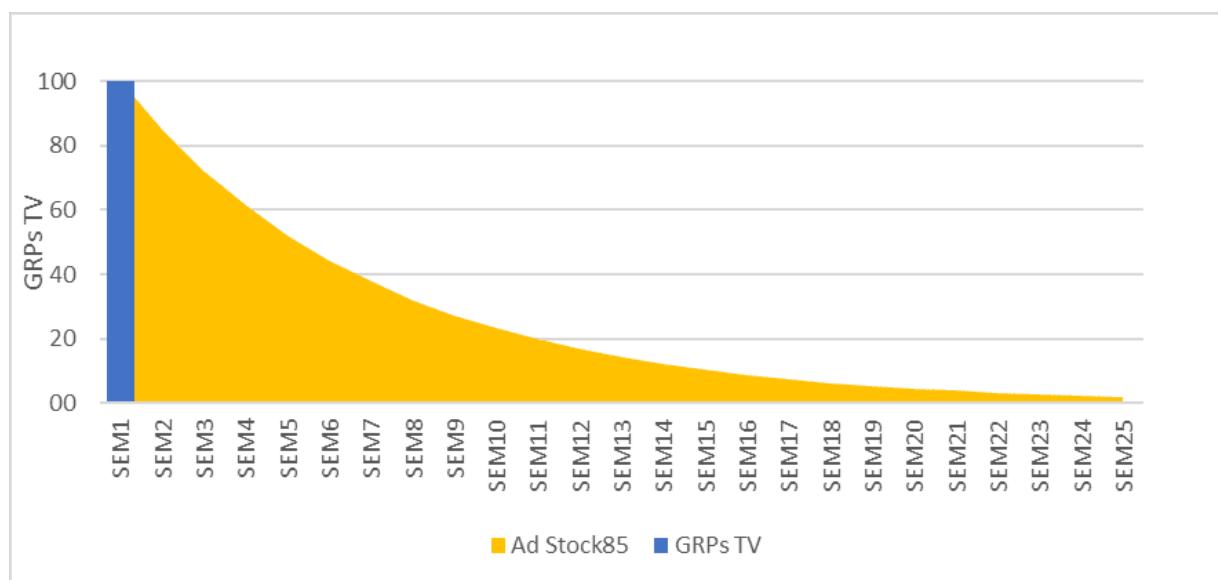
Intenta pensar en anuncios publicitarios que recuerdes, yo te comarto los que me resultan **más emblemáticos** (todos de TV) y el año en el que se hicieron: el anuncio del Atún Calvo del ‘tututum’ (2007), Coca Cola de ‘Hoy no me puedo levantar’ (2008) o, más recientes, las colonias de Invictus de Paco Rabanne (2014) o Mediamarkt con el rockero ‘Yo no soy tonto’ (2015). ¿Te sigue resultando raro que dure años?



Visualiza los anuncios: [Atún Calvo](#), [Coca Cola](#), [Invictus de Paco Rabanne](#) o [Mediamarkt](#).

El ad-stock es un efecto que permite disminuir el impacto de la publicidad de forma progresiva.

Sus valores oscilan entre 0-100 (o 0-1 si hablamos de porcentajes), donde **0 significa que cae de forma inmediata, y 100 que el anuncio sigue impactando con la misma intensidad durante los periodos siguientes a su emisión**. De forma visual sería:



La formulación matemática es muy simple. Si aplicamos un ad-stock **AD** a una variable de presión (GRPs/clics/impresiones) o inversión publicitaria **m**, la variable resultante en la semana **i** es:

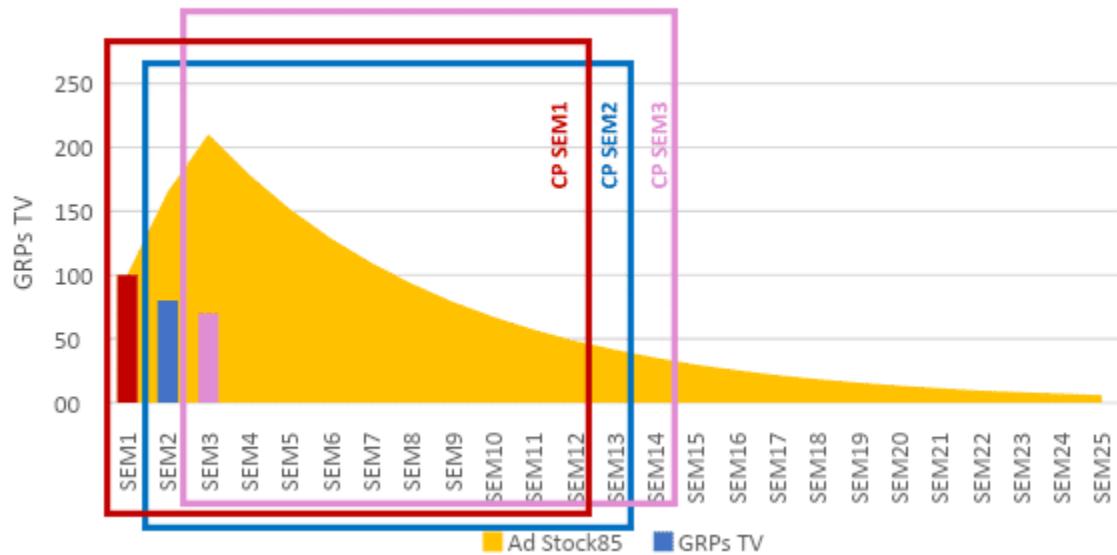
$$m_{AD}[i] = m[i] + m_{AD}[i-1] \times AD / 100$$

Veamos un ejemplo. Asumiendo un ad-stock (AD) de 50 para una serie de TV, la serie de TV con AD 50 sería:

	GRPs TV brutos 20 seg	GRPs TV con AD50
Semana 1	100	100 (no hay semana anterior)
Semana 2	80	80 (sem. 2) + 100 * 0,5 (sem. 1 con AD) = 13
Semana 3	70	70 (sem. 3) + 130 * 0,5 (sem. 2 con AD) = 135

Las variables ad-stock se suelen dividir en corto y largo plazo. El corto plazo suele contabilizarse durante las 12 semanas siguientes a la presión (3 meses si trabajamos a nivel mensual) y el largo plazo desde la semana 13 (o mes 4).

Veamos un ejemplo: si realizamos presión publicitaria en TV durante 3 semanas, y ponemos un ad-stock del 85, nuestra serie quedaría:



Como ves, hay semanas en las que conviven cortos y largos plazos. Por ejemplo, la semana 13 tendrá al mismo tiempo el largo plazo de la semana 1 y el corto plazo de las semanas 2 y 3.

El efecto del ad-stock se puede aplicar a todos los medios publicitarios, siendo los más propensos la televisión y la radio, y el menos frecuente SEM. Por ejemplo, cuando realizamos análisis de notoriedad es probable que todos los medios estén ad-stock con niveles altos.

El nivel de ad-stock de cada medio es muy variable en función de la serie a modelizar, el comportamiento del sector, la presión que se esté haciendo, etc. La televisión suele tener el AD más alto, seguido de la radio y el resto de medios digitales.

Resumen

X Edix Educación

A lo largo del fastbook hemos ido descubriendo **potenciales problemas y sus correspondientes soluciones a diversas situaciones** a las que nos podemos enfrentar en los departamentos de marketing relacionadas con el ‘formato’ de las variables.

Sabemos **medir el efecto que la publicidad provoca** en los periodos siguientes a su emisión (disminución paulatina) mediante el concepto de ad-stock. Gracias a ello, podemos **separar el efecto de la publicidad en corto y largo plazo**.

Lesson 4 of 4

Test

X Edix Educación

Te animo a que respondas el siguiente test que te ayudará a hacer un resumen más detallado del contenido.

Question

01/05

Imputación a 0:

- Missing value.

- Carencia de histórico.

Question

02/05

Interpolación:

- Transformar a menor granularidad (mensual -> semanal).
- Transformar a mayor granularidad (semanal -> mensual).

Question

03/05

Uso de promedio para problemas de granularidad:

- Para variables de volumen.
- Cuando son indicadores.

Question

04/05

Granularidad temporal de menor a mayor:

- Promedio y suma.
- Interpolación y repartir.

Question

05/05

Reparto:

- Para variables de volumen.
- Cuando son indicadores.



Respuestas: 1-A, 2-A, 3-B, 4-A, 5-A

edix

Creamos Digital Workers