iab spain

LIBRO BLANCO

Inteligencia Artificial en Adtech y Martech

OCTUBRE
2022

Índice_

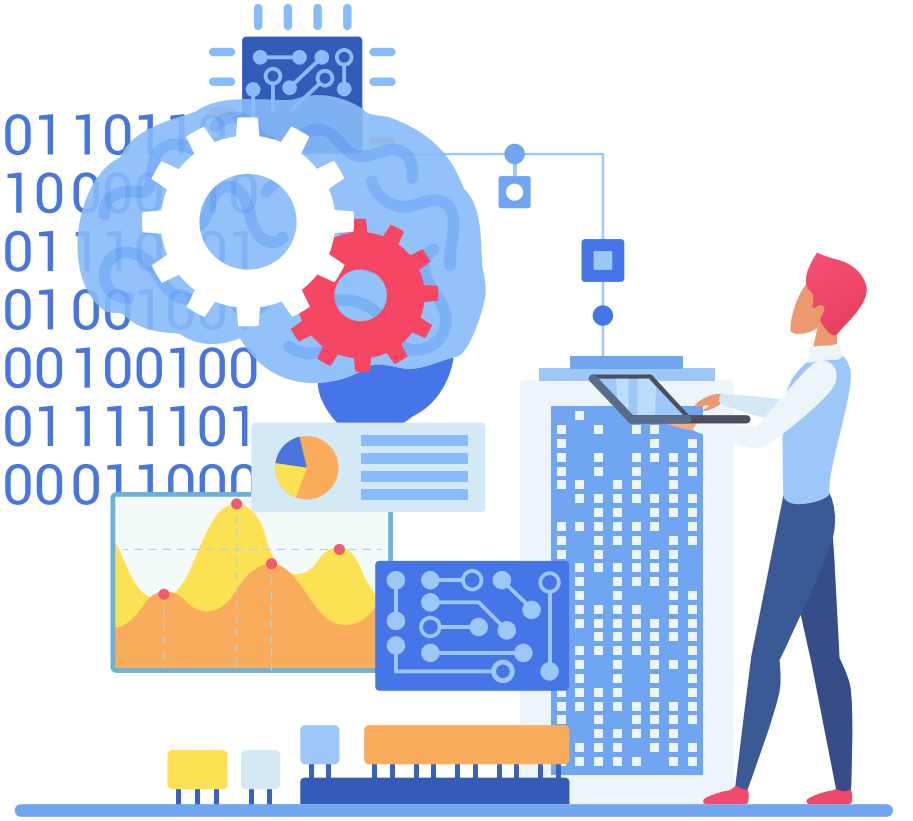
Introducción_

El objetivo de este libro blanco es sintetizar los grandes conceptos de un mundo tan extenso como la inteligencia artificial, describiendo los principales tipos de algoritmos y especialmente los de machine learning.

Y todo ello, desde la perspectiva del marketing digital, por lo que se incluyen una serie de ejemplos y casos de uso muy comunes en la industria.

Se ha pretendido buscar un lenguaje sencillo, en donde a pesar de la necesidad de usar los términos técnicos, se explique los conceptos detrás de estas metodologías cuyo uso no para de crecer gracias a su aportación de valor en:

- Mayor rapidez en los procesos de dato
- Automatización de decisiones
- Optimización de resultados
- Personalización de las audiencias
- Mayor eficiencia en los kpis de negocio
- etc



1 Tipos de algoritmos de Inteligencia Artificial (IA)_

Tipos de algoritmos de Inteligencia Artificial (IA)_

Antes de hablar de Inteligencia Artificial es necesario conocer que un algoritmo es una secuencia lógica de instrucciones que describen paso a paso la forma de resolver un problema. El objetivo de un algoritmo es definir los pasos necesarios para aprender de los datos y resolver un problema de forma autónoma.

Dentro de la inteligencia artificial se incluyen las siguientes ramas de algoritmos en función de las diferentes metodologías matemáticas aplicadas y su uso principal:

1.1- Machine learning (ML)

Hace referencia a los sistemas que aprenden automáticamente y que no se derivan de la aplicación de reglas de negocio basadas en conocimiento experto, sino que un algoritmo identifica en cada caso el patrón más acertado y este evoluciona en el tiempo aprendiendo de los datos. Dentro de las técnicas de ML, a su vez, existen 4 tipos de metodologías:

- **Modelos supervisados:** Se entrena al sistema proporcionándole cierta cantidad de datos clasificados, una vez que el sistema tiene suficientes eventos o muestra de la que aprender, es capaz de clasificar la información en base a los patrones que ha aprendido. Por ejemplo, la información acerca de los conversores de un evento (quienes sí compran, o quienes sí rellenan un formulario o…), permite saber la probabilidad de que otro individuo vaya a convertir.

  Ej de aplicación: identificar aquellos usuarios que tienen una probabilidad muy alta de realizar una compra en base a su patrón de navegación, perfil sociodemográfico, etc.

Este aprendizaje se basa en las relaciones existentes entre unas variables de entrada o explicativas y unas variables de salida o "a explicar". Dicho de otro modo, lo que haremos será enseñar a nuestros modelos qué resultado queremos obtener para una serie de valores.

Veamos un ejemplo. Somos un banco que dispone de una gran BBDD de clientes y quiere crear un modelo que le ayude a predecir qué clientes se darán de baja en el futuro.

Lo que hacen los algoritmos supervisados es tomar un conjunto de datos de muestra de los cuales ya sabemos si se han ido o no del banco y, una vez hayamos creado nuestro modelo, detectaremos si otro cliente será propenso o no a fugarse en el futuro.

Otro ejemplo. Si partimos de un conjunto de correos de los cuales ya sabemos si son o no spam, podremos crear un modelo para que futuros correos nuevos se envíen o no de manera automática a spam.

Tipos de algoritmos de Inteligencia Artificial (IA)_

Y un último ejemplo, ¿sabéis que, a partir del comportamiento de un conjunto de usuarios, somos capaces de detectar si un nuevo usuario me comprará o no el carrito de bebé?

Lo que hay que entender es que todos **estos algoritmos aprenden de un conjunto de datos de los que ya sabemos la respuesta** y nuestro modelo será capaz de predecir lo que ocurrirá ante nuevos datos.

Vamos a ver a continuación algunas de las principales **técnicas de Aprendizaje Supervisado:**

a) Regresión Lineal

La Regresión Lineal es una de las técnicas más usadas en Machine Learning. Sin duda, una de sus mayores virtudes es su simplicidad, tanto en el momento del aprendizaje como en el momento de interpretar los resultados. Se emplean cuando queremos predecir una variable continua en situaciones en las que la relación entre las variables de entrada y la variable a predecir es lineal.

Un ejemplo sencillo para entender estos modelos. Queremos predecir el Índice de Mortalidad a partir del número de cigarrillos. Para ello partimos de la siguiente tabla donde tenemos los datos de 7 ciudades:

Ciudad	Número de Cigarrillos (X)	Índice de Mortalidad (Y)
Madrid	2	0,10
Barcelona	3	1,00
Sevilla	4	1,50
Asturias	5	2,50
Galicia	7	5,00
Valencia	8	6,50
Granada	9	8,00

Si tomamos como eje X el número de cigarrillos y como el eje Y el Índice de Mortalidad podemos observar que los puntos siguen una tendencia lineal. Es decir, podemos encontrar una recta que se aproxime a esos puntos. La pregunta es, ¿cuál es esa recta?

El Algoritmo de Regresión Lineal encuentra la mejor recta, de tal forma que, cuando introduzcamos 6 como número de cigarrillos (valor que no teníamos inicialmente), el algoritmo nos predice que el Índice de Mortalidad es de 4. Es decir, a partir de la variable "*número de cigarrillos*" podemos predecir la variable "*Índice de Mortalidad*".

El caso anterior ha sido un ejemplo visual para comprender el concepto de Regresión Lineal. Hemos empleado una variable X (*número de cigarrillos*) para predecir la variable Y (*Índice de Mortalidad*). Cuando tenemos una variable como la de este ejemplo podemos graficar dicho problema en dos ejes (en el plano) y tenemos que encontrar una recta que mejor se aproxime a dichos puntos.

Bien, pues imaginemos ahora que queremos hacer lo mismo pero esta vez no sólo tendremos en cuenta la variable "*número de cigarrillos*" sino la variable "*número de cervezas*" para predecir el "*Índice de Mortalidad*". Este caso podríamos graficarlo en tres ejes (en el espacio) y tendríamos que encontrar el plano que mejor se aproxime a dichos puntos.

Obviamente en la vida real, no sólo disponemos de 2 o 3 variables, sino de decenas, cientos y miles de variables, cosa que es imposible de graficar. Pero la idea en el plano o en el espacio es suficiente para comprender estos algoritmos.

Del mismo modo que hemos mencionado la sencillez tanto en el aprendizaje como en la interpretación y es práctico para entrenar, hay que destacar que entre los grandes problemas de estos algoritmos se encuentra la necesidad de linealidad para poder ser empleados y el tratamiento de valores atípicos.

b) Regresión Logística

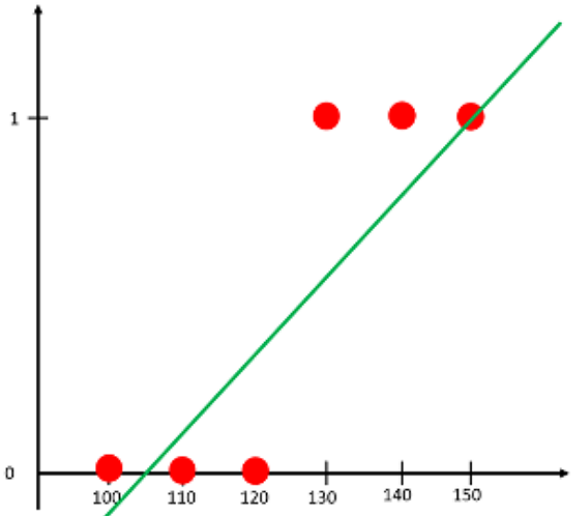
La Regresión Logística es una técnica de clasificación de Machine Learning. Los algoritmos de clasificación son aquellos que no tratan de "predecir" una variable continua, sino una variable que toma un conjunto de valores finitos, o clases de valores. Es decir, son problemas en lo que queremos predecir una variable que toma únicamente dos valores (1 o 0, True o False, luego es binario) o multiclase (tramo A o tramo B o tramo C, por ejemplo).

Supongamos que queremos predecir si una casa ha sido vendida o no en función de los metros cuadrados de la misma. Por ejemplo, imaginemos que tenemos estos datos de 6 casas.

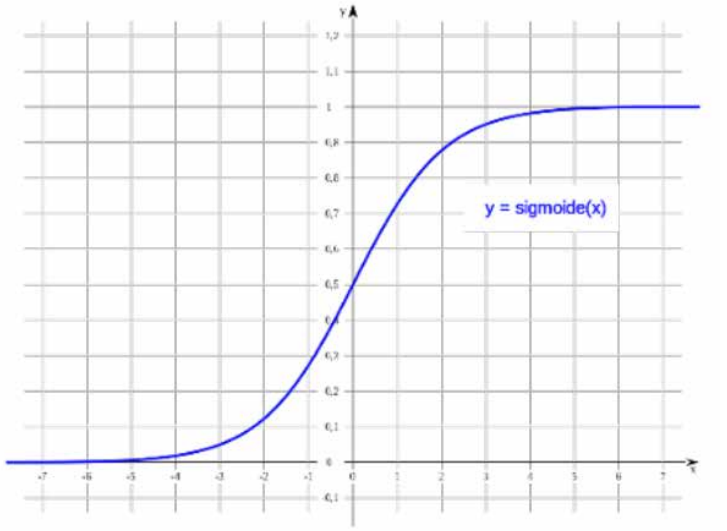
Metros cuadrados	Vendida (1=si / 0=no)
100	0
110	0
120	0
130	1
140	1
150	1

Tipos de algoritmos de Inteligencia Artificial (IA)_

Podemos observar que, para estos problemas, una Regresión Lineal no es la mejor de las opciones.



Sin embargo, la Regresión Logística es la mejor opción, ya que, sin entrar en detalles matemáticos, en vez de emplear una recta para aproximar los puntos, emplea la llamada Función Sigmoide que tiene esta pinta:



Del mismo modo que con la Regresión Lineal, es un modelo sencillo de entender y explicar. También tiene otras ventajas como que es poco propenso al sobre ajuste y es rápido de entrenar. Sin embargo, como ocurría con la Regresión Lineal, es problemático ante datos no lineales y puede sufrir con valores atípicos.

c) Árboles de decisión

Los árboles de decisión son muy frecuentes ya que en contexto de negocio facilita la explicación e interpretación de los resultados del algoritmo. El propio formato en el que se nos presentan los resultados (árbol) se puede trasladar directamente a un conjunto de reglas de negocio que directamente se pueden aplicar.

Vamos a ver un ejemplo muy sencillo de lo que hacen estos algoritmos.

Supongamos que queremos ver los individuos más propensos a comprar el carrito de bebé del que hemos hablado antes.

Como cualquier algoritmo de aprendizaje supervisado partimos de un conjunto de usuarios de los que ya sabemos quiénes han comprado o no nuestro carrito. Por ejemplo, vamos a tomar a 10.000 usuarios como base para que nuestro modelo aprenda. De todo ese conjunto sabemos que el 15% ha terminado comprando el carrito. Inicialmente tenemos el siguiente nodo:

Lo que hacen los Árboles de Decisión es sumergirse y estudiar todas las variables de las que disponemos (*género, edad, salario, etc*) y encontrar aquella que mejor separa o discrimina ese 15% que teníamos identificado que sí ha comprado. Continuamos con el ejemplo para entender esto mejor. Supongamos que el modelo ha detectado que la mejor variable para separar ese porcentaje inicial es la edad, ahora tenemos dos grupos, un grupo que supera los 35 años (*Grupo 1*) y otro que está por debajo de los 35 (*Grupo 2*). Si miramos ahora el Grupo 1, vemos que el porcentaje de compra es del 30%, mientras que el Grupo 2 es ahora del 7%.

Llegados a este punto, nuestro árbol continuará dividiendo del mismo modo. Es decir, volverá a estudiar cuál es la variable que mejor separa estos grupos y se volverá a ver el porcentaje de compra en cada nuevo grupo. Imaginemos que tenemos ahora la siguiente ramificación.

Con dos ramificaciones ya tenemos una visión general de los perfiles a los tenemos que dirigirnos para que compren un carrito.

De este modo hemos encontrado 2 perfiles diferentes de personas que compran 7 veces y 4 veces (respectivamente) el carrito más que la media global de usuarios.

Entre las ventajas que tenemos con este tipo de modelos se encuentran el que soporta relaciones entre variables no lineales y, como hemos comentado, el enorme sentido de negocio que aportan.

¿Sus desventajas?, aunque no lo hemos mencionado en modelos anteriores, existe un problema general común a todos y es el sobre ajuste y el especial cuidado que hay que tener con ello. Otro defecto es que son sensibles a pequeñas variaciones en los datos, es decir, si tomamos una muestra representativa para enseñar a nuestro modelo obtenemos un árbol y si tomamos otra muestra que también es

representativa puede que obtengamos un árbol totalmente distinto al anterior.

- **Modelos no supervisados:** Estos sistemas tienen como finalidad la comprensión y abstracción de patrones en la información que se analiza, un caso común es la técnica de clustering. El sistema identifica los atributos que hacen similares a un mismo conjunto y lo diferencian de otro conjunto.

Ej de aplicación: Las variables que caracterizan al grupo de clientes que sí causan baja de una compañía y que son capaces de discriminar o diferenciar respecto a lo que no se dan de baja.

A diferencia de las técnicas de modelos Supervisados, estos aprenden únicamente a partir del conjunto de datos de entrada sin saber el resultado qué queremos obtener y son los propios algoritmos los que nos darán información. Con estas técnicas se busca resolver dos problemáticas:

1. **Agrupación**
2. **Reducción de la dimensionalidad**

Supongamos que jamás en nuestra vida hemos visto ni oído hablar de las películas de superheroes. Un amigo nuestro nos muestra la siguiente imagen:

No tenemos ni la más remota idea de quiénes son esos personajes. Sin embargo, nos damos cuenta de que puede que existan dos grupos de personas en base a los patrones de vestimenta y comportamiento.

Tipos de algoritmos de Inteligencia Artificial (IA)_

Es muy común nombrar a este tipo de aprendizajes ***"Clustering"***, y es uno de los problemas más importante dentro del campo del Aprendizaje No Supervisado. Hemos sido capaces de hacer dos grupos (1.agrupación) sin la necesidad de tener ningún conocimiento sobre ellos (2.reducir la dimensionalidad del problema).

Aquí es donde encontramos uno de los principales inconveniente de estos algoritmos y es que no tenemos ningún ejemplo de respuestas con el que podamos comprobar si lo que hemos hecho es correcto o no. Incluso existe otro gran problema y es que no sabemos el número exacto de grupos que existen, pero con criterio de negocio se define como de granular se desea discriminar a X grupos.

Sin embargo, la principal ventaja que tienen estos modelos de aprendizaje frente a los Supervisados es que los datos son menos costosos de conseguir. Es lógico, ya que, si queremos predecir que, si un usuario va a comprar el carrito de bebé o no mediante un Algoritmo de Aprendizaje Supervisado, es necesario que previamente tengamos una gran cantidad de usuarios de los que sepamos si han comprado o no. Estos algoritmos son empleados para encontrar patrones de comportamientos entre los datos que introducimos.

- **Reinforcement learning:** el algoritmo aprende por medio de prueba y error hasta alcanzar la mejor manera de completar una tarea dada.

  Ej de aplicación: Los sistemas de recomendación de contenido (películas, series, etc) aprenden a adaptar su respuesta en base a "recompensas" (cada vez que alguien ve un determinado contenido) para que resuelva y proponga la mejor alternativa sin programarlo específicamente para que lo realice de una forma determinada.

- **Deep learning (DL):** las redes neuronales son algoritmos que se desarrollan a través de niveles jerárquicos. En el nivel inicial de la jerarquía la red aprende algo simple y luego envía esta información al siguiente nivel. El siguiente nivel toma esta información sencilla, la combina, compone una información algo un poco más compleja, y se lo pasa al tercer nivel, y así sucesivamente.

Tipos de algoritmos de Inteligencia Artificial (IA)_

Ej de aplicación: El reconocimiento de imágenes donde es posible identificar objetos a gracias a que iterativamente cada nivel jerárquico determina formas, colores, etc. para finalmente consolidarlo todo y concluir si la imagen contiene un animal, una ciudad, etc

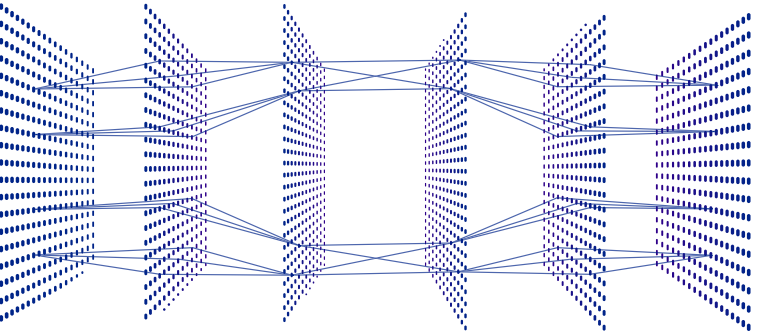
Las Redes Neuronales constituyen posiblemente la familia de modelos más famosos en los últimos años. Tienen por objetivo construir un modelo capaz de reproducir el método de aprendizaje del cerebro humano

Los modelos de aprendizaje que hemos comentado hasta ahora funcionan de forma más sencilla: metemos un conjunto de datos y nuestro modelo aprende de ellos. Sin embargo, una red neuronal es bastante complicada.

Las redes se estructuran en capas, debiendo existir al menos una capa de entrada (asociada a las variables explicativas) y una capa de salida (asociada a la variable target). Una red neuronal puede poseer de forma opcional una o varias capas ocultas que ofrecen a la red gran flexibilidad en los tipos de relaciones input – output que puede manejar, permitiendo reflejar relaciones más complejas (no lineales) entre las variables. El campo del **DEEP LEARNING** no es más que algoritmos basados en cientos y miles de Redes Neuronales conectadas entre sí.

En cuanto a las ventajas que tiene ese modelo está la gran flexibilidad en los datos de entrada. Sin embargo, a pesar de ser a nivel matemático y conceptual el algoritmo más potente en el campo del Machine Learning, tiene sus desventajas: la gran complejidad de aprendizaje cuando más cosas se necesiten aprender y, lo más importante para este negocio, no permite interpretar los resultados. Es decir, tomando como ejemplo el Árbol de Decisión el cual recordemos que nos permite "andar por las ramas" para encontrar e interpretar los resultados, lo que ocurre dentro de las capas ocultas de una Red Neuronal no podemos verlo, o mejor dicho, es demasiado complejo para observarlo simplemente y reducir fácilmente a unos pocos parámetros su explicación.

En pocas palabras, una Red Neuronal es muy posible que resuelva mejor el ejemplo de los usuarios que comprar el carrito, es decir, acertará con mayor probabilidad que un arbol. Sin embargo, como no podemos saber los motivos por los que un usuario es más propenso a comprar, no podremos extraer lógicas de negocio y, en este caso, la Red Neuronal no sería el mejor algoritmo para este problema.

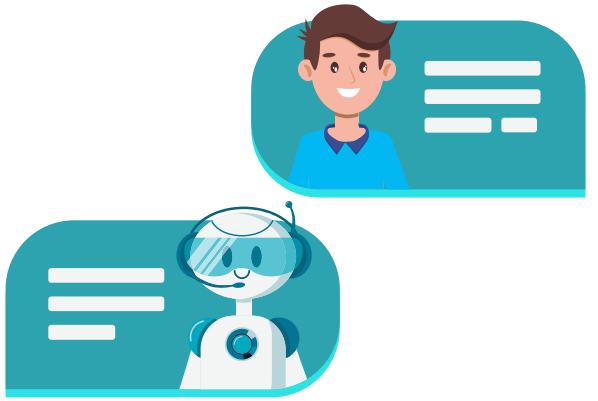


1.2.- Procesamiento de lenguaje natural (NLP)

Este tipo de algoritmos se usan fundamentalmente para la:

- **Extracción de información:** identificación y conversión del texto en datos estructurados.
- **Clasificación:** categorización de la información aplicando análisis morfológico y sintáctico.
- **Traducción:** incorpora un análisis semántico y pragmático para incorporar la correcta interpretación previa a la traducción del idioma.

Ej de aplicación: los chatbots que atienden las peticiones de los usuarios en apps, webs, etc



1.3.- Robotics

Este tipo de algoritmos se usan fundamentalmente para:

- **Implementación**, a través de mecanismos de software, de procesos que automaticen tareas repetitivas para ahorrar tiempo y minimizar errores.

Ej de aplicación: extracción de datos manualmente de los ad server con carácter periódico para copiar, pegar y agregar a una hoja de cálculo en la que se realicen cálculos para realizar el seguimiento de una campaña. Se pueden sustituir por procesos de lectura de la fuente a través de APIs que actualicen todos los días el informe a una hora determinada.

1.4.- Speech

Este tipo de algoritmos se usan fundamentalmente para:

- **Speech to text:** transformar voz en texto

Ej de aplicación: clasificar el texto obtenido del archivo de sonido para clasificar tópicos que representativos del contenido y poder clasificar a todos los usuarios que lo escuchan como interesados en estos tópicos, para así generar audiencias de interés. Así un Podcast de noticias que es muy genérico puede clasificarse en tópicos en función de los temas que se traten en cada capítulo (internacional, local, economía, ..) o temas más concretos por las noticias y etiquetar a los usuarios que consumen cada capítulo.

1.5.- Text to speech: Transformar texto en voz

Ej de aplicación: Capturar un audio de un usuario a través del micrófono para convertirlo en texto y realizar una búsqueda en internet (Google, Bing,...).

1.6.- Optimización y planificación

Este tipo de algoritmos se usan fundamentalmente para optimizar el uso de un recurso que es finito, y donde un mejor uso puede generar ahorros de costes o beneficios:

- Análisis para realizar la programación de medios más eficiente posible adaptada a la oferta y demanda real con los recursos disponibles.

Ej de aplicación: Se dispone de un presupuesto anual para comunicación con el que se busca maximizar el retorno de la inversión en el kpi de negocio. La optimización proporciona cuánto se ha de invertir en cada medio y en qué momento del año conforme a ciertas restricciones marcadas.

Ej de aplicación: A partir de una parrilla de TV donde el espacio publicitario abierto en segundos tiene un máximo permitido por la legislación vigente. La venta de publicidad, generalmente se hace en base a GRPs (audiencias), con lo que una colocación óptima de los pases publicitarios, atendiendo al objetivo de buscar audiencia permite minimizar el número de pases publicitarios y así colocar más campañas de publicidad en un mismo espacio, rentabilizando así la parrilla de Tv. Por ejemplo, dos pases de 1 GRP ocupan 40 segundos (20s+20s) y un pase de 2 GRPs 20s, llegando a la misma audiencia usando la mitad de segundos.

1.7.- Visión

Este tipo de algoritmos se usan fundamentalmente para:

- Reconocimiento de imágenes/vídeo:

Ej. De aplicación: Reconocimiento de imágenes para evaluar contenido de vídeo y taguearlo con atributos que permitan salvaguardar los criterios de brand suitability y brand safety de los anunciantes. Por ejemplo, una película de catástrofes puede ser tagueada como apropiada para anunciantes de seguros, pero no para anunciantes de viajes.

- Tratamiento de vídeos

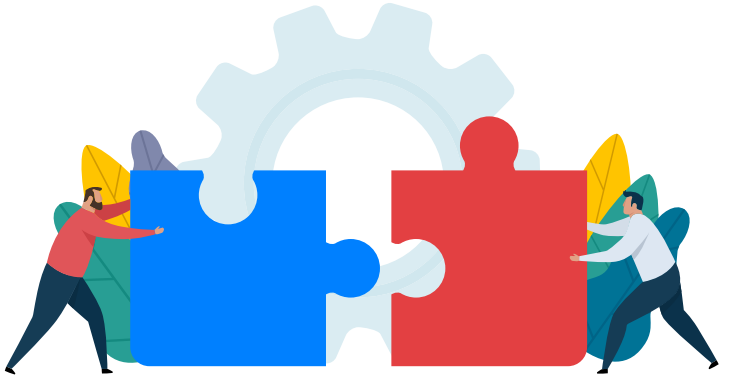
Ej. De aplicación: Identificación de la exposición de la marca en piezas de vídeo o imágenes, a través del reconocimiento del logotipo cuando no se trata de anuncios publicitarios propiamente.

2 Puesta en marcha de los algoritmos_

Puesta en marcha de los algoritmos_

Antes de empezar cualquier proyecto de ML (Machine Learning), hay que tener bien definidos y claros una serie de puntos, de manera que podamos garantizar una buena definición de proyecto ML, estos puntos son:

1. Entender y definir el problema a resolver
2. Analizar y preparar el dato
3. Aplicación de algoritmos
4. Reducción de errores
5. Predicción de resultados



2.1.- Entender y definir el problema a resolver

Lógicamente, todo empieza con un problema, sin problemas no habría necesidad de solucionar ninguna situación.

El problema suele plantearlo negocio, y no porque quiere plantear problemas porque sí, sino que es algo que es de vital importancia para la organización y que se supone puede dar nuevos ingresos. Este punto es muy importante, hay que tener una aproximación positiva al problema porque resolverlo de manera correcta puede comportar nuevos ingresos para la compañía.

¿Y desde negocio cómo podemos hacer para entendernos mejor con nuestros compañeros de data?

Para ello es muy importante que determinemos el problema a resolver en todos los parámetros que podamos definirlo. Para ello hay una manera de tender ese puente y hacer entender muy bien cuál es nuestro resultado esperado y es mediante la elaboración del caso de uso. En un caso de uso no pueden faltar la siguiente información:

1. Definición del problema
2. Definición de las fuentes de datos
3. Definición de los resultados esperados
4. Definición de deadlines
5. Definición del formato del entregable

Ahora que sabes todo esto, intenta pensar en un caso de uso para este ejemplo: queremos ser capaces de diferenciar las variedades de una misma flor en función de sus características.

- **Definición del problema:** diferenciar flores
- **Definición de las fuentes de datos:** nuestro banco de datos de nuestras flores y si necesitamos hay que evaluar la compra de datos fuera
- **Definición de los resultados esperados:** clusters de tipo de flores
- **Definición de deadlines:** mínimo 4 semanas, 1 para hacer, 3 para entrenar
- **Definición del formato del entregable:** Etiqueta para cada variedad

Dicho esto, por tanto, nuestra meta será crear un sistema basado en ML capaz de clasificar las distintas variedades de una flor en función de sus características y esta es la conclusión a la que podemos llegar porque la definición del caso de uso ha sido completa.

Si nuestro problema a resolver es demasiado complejo, siempre podemos descomponer este en otros problemas más sencillos que nos permitan resolverlo con mayor facilidad.

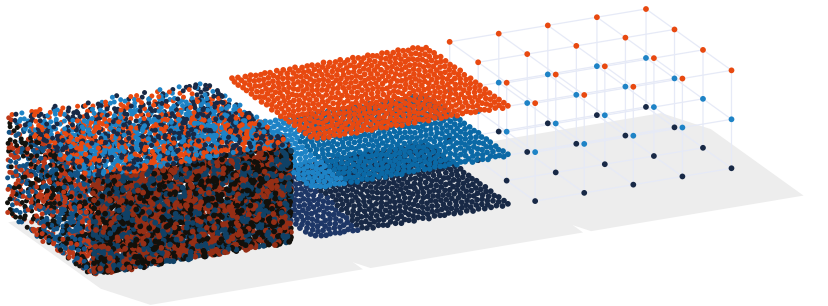
Como dijo el matemático George Pólya: "Si no puedes resolver un problema, entonces hay otro problema más sencillo que podrás resolver, encuéntralo".

2.2.- Analizar y preparar el dato

Una vez tenemos nuestro problema bien definido, quizá una de las preguntas iniciales que podemos hacernos sea ¿Cuántos datos necesito para hacer un proyecto de ML?

Esta pregunta no tiene fácil respuesta, pues depende del caso de estudio, de manera general podemos decir que, a más datos, mejor, y en todo caso manejar como referencia los siguientes mínimos:

- Como punto inicial, deberíamos tener mínimo cerca de 1.000 ejemplos de media.
- Para la mayoría de los casos, deberían existir entre 10.000 y 100.000 ejemplos.
- Para los casos más complejos (por ej, traducciones, grandes dimensiones de datos, proyectos de deep learning), necesitaremos mínimo entre 100.000 y 1.000.000 de ejemplos. Como norma básica general podríamos decir que por cada dimensión de datos que tengamos, necesitaremos cerca de 10 veces más de datos.
- Cuanto más complejo sea el problema, más datos serán necesarios.



Puesta en marcha de los algoritmos_

Una vez sabemos que tenemos la cantidad de datos necesaria para iniciar nuestro proyecto, quedaría conocer los pasos o procesos.

1. **Recogida de datos:** deberemos recoger información relevante para el modelo que queremos desarrollar.

2. **Limpiar los datos:** es muy común que los datos a tratar no estén todos rellenos o con valores homogéneos en cada columna, especialmente si estos datos se han recogido de encuestas respondidas directamente por personas o se han agrupado datos de sistemas distintos. Así pues, suele ser bastante normal que haya que realizar una labor de limpieza de datos para dejarlos preparados de cara a una primera evaluación:

   - Decidir qué hacemos con valores que vienen a nulo o sin valor. Esto puede ir desde eliminar filas, imputar un estadístico (media, moda) o dejarlos vacíos si el algoritmo a usar lo soporta.

   - Valores que están fuera del dominio del atributo. Ej. Para una columna que lleve el CPM encontrar valores negativos.

   - Valores que vienen incompletos o mal escritos. Para una columna con la provincia valores como madriz, Madrí, mad, etc.

   - Valores que no están normalizados; un algoritmo clasificará como valores distintos madrid, Madrid o MADRID para una provincia.

3. **Preparación de los datos:** es importante reconocer y minimizar posibles sesgos en nuestro dataset, para ello:

   - En primer lugar, habrá que ordenar de forma aleatoria los datos para evitar que el orden de los datos afecte al modelo.

   - Examinaremos los datos para detectar posibles asimetrías, que ayudarán a identificar y rectificar posibles sesgos que afecten a los resultados.

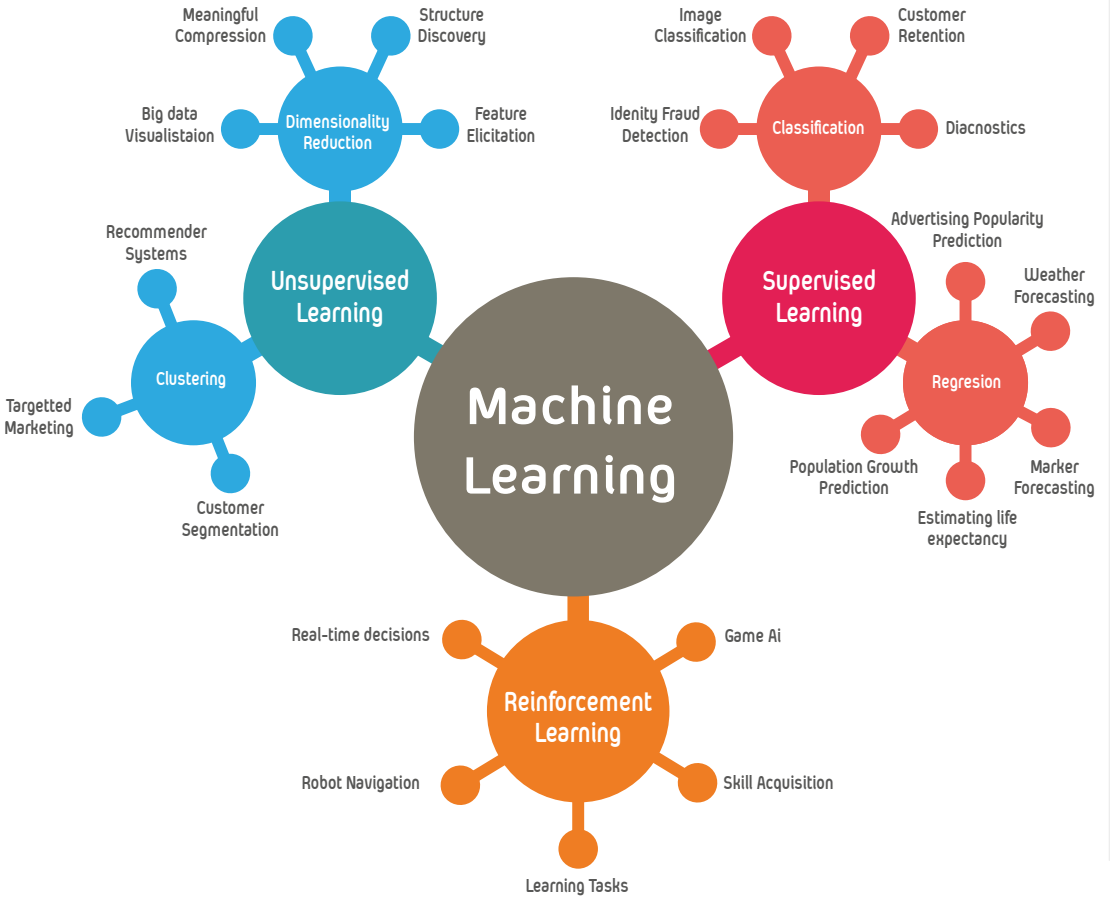
   - División de los datos en el conjunto de entrenamiento y de evaluación. Generalmente se hace con un 80%/20% o 70%/30% respectivamente. Con el primer 80%-70% de los datos se entrena el modelo y con el 20%-30% restante se evalúa si las predicciones se corresponden con los datos.

   - Otros puntos a tener en cuenta son los relativos a la calidad del dato, como son la veracidad, validez, integridad, consistencia, precisión y actualidad.

Cuanto mejor estén preparados los datos, más eficiente será el modelo.

2.3.- Aplicación de algoritmos

Dependiendo de la finalidad de nuestro proyecto, deberemos elegir un modelo u otro, por ejemplo, para proyectos relacionados con textos, funcionaran mejor unos modelos, que para otros destinados a la detección de imágenes.



Para ello siempre podemos evaluar distintos algoritmos para ver cuál es el más acertado para nuestro problema para poder decantarnos por el que lo resuelva de forma más precisa. Es importante entender que en ML no existe un modelo específico que sea capaz de responder al 100% a cada problema planteado, por ello tendremos que buscar aquel algoritmo capaz de reducir al máximo las probabilidades de sobre ajustar el modelo (overfitting), así como reducir la varianza, de forma que consigamos una mayor precisión.

Encontramos por tanto dos fases:

1. **Entrenamiento:** Es la parte esencial del proceso de Machine Learning, es donde usaremos la parte de nuestros datos destinada a enseñar al modelo para obtener el output deseado. Este proceso requiere un proceso interativo y de experimentación, por decirlo de una forma sencilla, sería comparable a aprender a montar en bicicleta.

2. **Evaluación del modelo:** Una vez entrenado el modelo es necesario ponerlo a prueba con el conjunto de datos que hemos reservado para el testeo. Esto permite detectar si los objetivos para el que se creó el modelo se consiguen o no. En caso de que no se consigan, habrá que revisar los pasos anteriores hasta identificar el motivo del fallo del modelo para rectificarlo. Si la evaluación no se realiza correctamente, el modelo podría no cumplir el objetivo para el que fue diseñado, lo que podría generar un grave problema (ej: modelo creado para evaluar la optimización e inversión en medios de una campaña publicitaria 360).

2.4.- Reducción de errores

Una vez tengamos la confirmación que la evaluación es satisfactoria, procederemos al ajuste fino de los parámetros del modelo, que consiste en optimizar los resultados de forma que sean cada vez más precisos.

2.5.- Predicción de resultados

Una vez alcanzado este punto podremos considerar que nuestro modelo está preparado para ser aplicado en casos reales. El modelo es capaz de extraer sus propias conclusiones, siendo el verdadero reto del modelo igualar el juicio humano a la hora de evaluar los distintos escenarios.

2.6.- Cumplimiento normativa datos

Desde el 25 de mayo de 2018 entra en aplicación el RGPD (Reglamento General de Protección de Datos), suponiendo el cambio más significativo en la legislación de protección de datos de la Unión Europea desde el año 1995

En este nuevo reglamento se establecen una serie de definiciones que tratan de aclarar y facilitar el cumplimiento de la norma. Entre estas definiciones quizá las más importantes son:

- **Datos personales:** toda información sobre una persona física identificada o identificable. (Entendiéndose por identificable cualquier método o variable como un identificador, un nombre, datos de localización, uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona).
- **Interesado:** persona sobre la que se estén tratando los datos personales.
- **Responsable del tratamiento:** quien determina los fines y medios del tratamiento de datos personales.
- **Encargado del tratamiento:** quien trate datos personales por cuenta del responsable del tratamiento.
- **Destinatario:** quien recibe o al que se le comuniquen datos personales.
- **Tercero:** quien trate datos personales siempre que no sea el interesado, el responsable de tratamiento, el encargado de tratamiento o las personas autorizadas para tratar los datos personales bajo la autoridad directa del responsable o del encargado del tratamiento.

- **Tratamiento:** cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales.

- **Elaboración de perfiles:** toda forma de tratamiento automatizado de datos personales con el fin de evaluar, analizar o predecir aspectos personales de una persona física.

- **Consentimiento del interesado:** toda manifestación específica, libre, informada e inequívoca por la que el interesado acepta al tratamiento de sus datos personales, mediante declaración o clara acción afirmativa.

Por otro lado, también contamos con un **Código de Buenas Prácticas en protección de datos para proyectos de Big Data** desarrollado por la AEPD (Agencia de Protección de Datos Española) en el que se recogen las principales implicaciones derivadas de los tratamientos, el origen del dato, la calidad y conservación de los mismos, la obtención del consentimiento de los interesados, la transparencia que se debe ofrecer en la información previa facilitada a los mismos, etc… Se puede consultar aquí.

2.7.- Presentación de resultados

En cuanto a la presentación de resultados, suele haber un gap entre la capa de negocio que patrocina el proyecto o iniciativa y el equipo DS que lo aborda. Esto se debe a la disparidad de perfiles y conocimientos y a que hay una tendencia natural a explicar los detalles de la resolución de cada problema, que ya de por sí suele ser complejo. Esto se complica porque desde el lado que debe escuchar no se suele disponer del tiempo adecuado para dedicar a los detalles del problema en cuestión.

Para evitar este efecto, hay algunos puntos que se pueden tener en cuenta en la presentación de resultados:

- Basarse en un documento gráfico; la presentación con elementos gráficos suele ser más sencilla de entender que el observar tablas de números. Aquí vale la pena dedicar tiempo a detalles como los gráficos a elegir, colores, formatos de números, acompañar de kpis resaltados, explicaciones o conclusiones que se auto expliquen…

- Utilizar KPIs de negocio, o del dominio del problema para explicar la solución y el resultado, dejando en segundo plano algunas métricas estadísticas que necesitan de por sí una explicación.

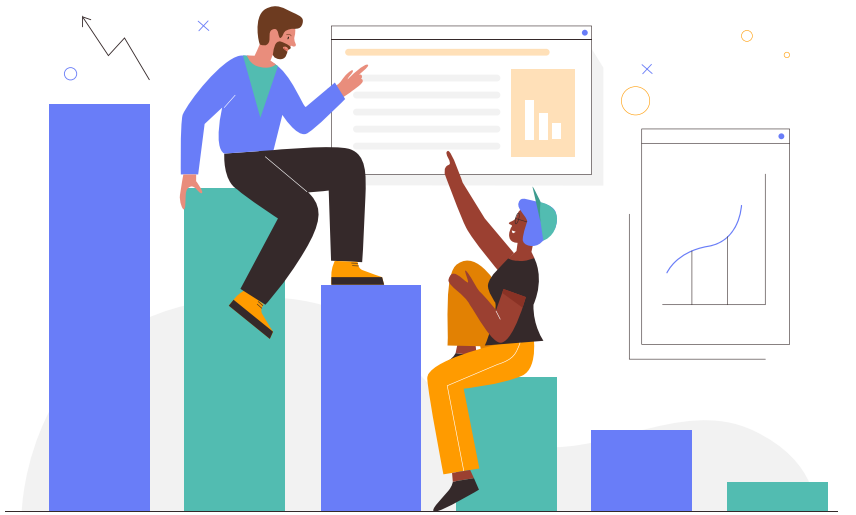
Puesta en marcha de los algoritmos_

- Concreción y brevedad; es preferible hacer una presentación breve con un buen entendimiento de la solución y que evite los pasos intermedios para llegar a ella, ofreciendo la posibilidad de profundizar al final de la presentación. Es preferible empezar con la solución y explicar después hasta llegar al detalle requerido.

- Contar los resultados; funciona mejor, explicar la presentación, de forma que se puedan dar detalles no incluidos si es necesario que directamente enviar un documento con la explicación de la solución.

- Evitar tecnicismos; hay que tratar el storytelling desde un punto de vista de negocio, no entrando en los detalles de los algoritmos utilizados si no es necesario. Al final hay que intentar dar una explicación plausible para la capa de negocio, que refuerce el resultado obtenido.

- Adaptar la presentación a la audiencia en cuestión; depende de los asistentes, debe adaptarse la presentación al grupo que lo va a recibir, ya que a veces puede ser un grupo de departamentos distintos y a los que el problema les toca de forma tangencial. Puede parecer una obviedad, pero tener parte de la audiencia que se descuelga de la presentación o hace preguntas que a otros les parecen obvias puede disminuir el efecto del resultado a presentar.



3 Aplicaciones de marketing de la IA_

LIBRO BLANCO **Inteligencia Artificial en Adtech y Martech**

Aplicaciones de marketing de la IA_

A continuación, vamos a describir usos comunes de la inteligencia artificial (IA) en las necesidades del marketing, sin profundizar ahora en la técnica concreta empleada:

3.1.- Prospects

- **Identificación y construcción de audiencias:** determinar los factores clave que describen y diferencian a un segmento o target para poder diseñar la comunicación más adecuada.

Ej: Cómo son los potenciales compradores de coches eléctricos de gama alta y cuáles son los medios de comunicación más afines para impactar con la campaña.

- **Retargeting:** en el ecosistema digital es la identificación de un individuo en diferentes plataformas o sites, y en base al interés que ha mostrado por comportamientos pasados, impactarle con publicidad relacionada.

Ej: Un usuario visita un website de viajes interesándose por París como destino sin contratar ningún viaje. Ese mismo usuario, posteriormente y en otro site, se encuentra con un anuncio en el que se le ofrece un viaje a París.

- **Look alike:** búsqueda de personas con perfiles similares a los de una audiencia de interés para cierta comunicación.

Ej: Una audiencia, que es el target más afín para comprar moda online, es identificada por ciertos comportamientos e intereses y se busca dichas características en otras cookies o personas para comunicar los productos de moda a los que sean más afines o similares al segmento de referencia.

3.2.- Clientes

En el ámbito de marketing o CRM, se lleva trabajando desde hace décadas en el desarrollo de modelos de propensión a realizar una determinada acción (compra, baja, etc) para identificar a través de modelos matemáticos aquellos clientes con propensiones más altas y así priorizar sus acciones de comunicación.

- **Cross-selling o venta cruzada:** son las acciones dirigidas a conseguir que un cliente compre o contrate algún producto o servicio adicional al que ya tiene con la marca.

Ej: Un cliente que tenga un seguro de auto con la compañía y presenta una propensión alta a contratar una póliza de otro producto, por ejemplo hogar, es priorizado en las acciones de comunicación del producto de hogar.

- **Up-selling:** son las acciones dirigidas a conseguir que un cliente contrate un nivel de servicio o producto superior al que ya tiene con la marca.

Ej: Una contratación de una tarifa de telefonía en la que el cliente sube su cuota mensual basándose en su mayor propensión a incrementar su consumo para priorizar dichas acciones.

- **Churn & Retention (fuga y retención):** es la identificación de clientes en riesgo alto de cancelar el producto o contrato con la marca y las acciones priorizadas con esos clientes para retenerles.

Ej: Un cliente tiene sus ahorros y nómina vinculada a un banco y está valorando cambiarla a otro banco. El banco actual identifica esta situación a través de modelos de propensión a la fuga y prioriza las acciones para retenerle.

3.3.- Prospects y clientes

Existen aplicaciones desde el punto de vista de data que son similares en cuanto a enfoque y metodología tanto para prospects como para clientes:

- **Insights & profiling:** Descubrimiento de los atributos más relevantes en los comportamientos, intereses, características, etc de las audiencias o targets.

Ej: Aplicación de técnicas de clustering para identificar que distingue: 1) a los clientes con mayor valor potencial de la compañía para realizar acciones de fidelización 2) a los prospectos o potenciales nuevos clientes más afines al producto de la marca y por tanto los prioritarios para dirigir la comunicación.

- **Personalización:** se refiere a la adaptación del mensaje y oferta en función de cada target, de forma que en lugar de existir una única comunicación existen tantas como perfiles se hayan identificado.

Ej: Un usuario de una plataforma de comida a domicilio que es un habitual consumidor de menú familiar los domingos por la noche, recibe un correo con una oferta de menú adaptado a estas características en lugar de la oferta estándar para audiencias que no han podido ser perfiladas.

- **A/B testing:** Un caso particular dentro de la personalización es la realización de pruebas con dos versiones diferentes de contenido o landing page considerando el resto de los elementos y condiciones de contorno iguales.

Ej. Comparar resultados de varios formularios para valorar cuál de ellos convierte mejor.

- **Creatividades dinámicas:** Un caso particular dentro de la personalización es la adaptación de la pieza de comunicación para optimizar los resultados.

  Ej. Comparativa de la tasa de clic o conversión de varias piezas creativas diferentes del mismo producto en el mismo site y similar número de impresiones y exposición.

3.4.- Medición del impacto de la comunicación

Los enfoques más comunes a la hora de medir el impacto de la comunicación basados en algoritmos son los siguientes:

- **Atribución digital (MTA o multi touch attribution):** partiendo de los impactos digitales que recibe una cookie o persona, se evalúa cuál es la importancia de cada uno de los canales digitales que han asistido para que finalmente se produzca una conversión o no. Esto se realiza en base a modelo matemáticos (cadenas de Markov, Shapley value, etc) que analizan el path de las cookies, y no bajo criterios de negocio como por ejemplo la medición de last clic que ignora la participación de otros canales offline y digitales en partes previas del funnel. A día de hoy, debido a las restricciones legales para trazar toda la actividad de las cookies y a la gran penetración de los walled garden, no se dispone de los paths completos que permitan llevar a cabo modelos robustos.

- **Atribución multimedia (MMA):** desde una perspectiva multimedia y no basada en cookies, se desarrollan modelos matemáticos que calculan la atribución de las interacciones llegando a asignar estadísticamente el efecto de los soportes , de las creatividades, la relevancia del momento de la comunicación, etc y así poder optimizar los planes tácticos de comunicación.

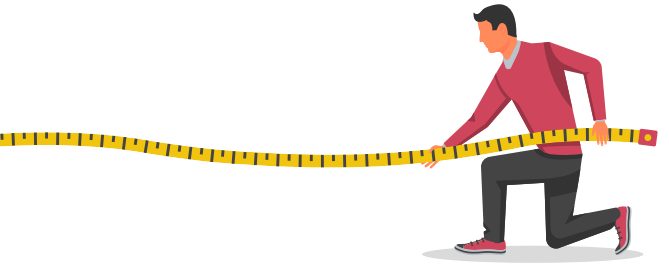
- **Marketing mix modeling (MMM):** se trata de modelos econométricos basados típicamente en regresiones lineales múltiples que permite estimar para periodos de varios años el ROI de los medios y su contribución al kpi analizado así como entender el efecto de tendencias de mercado, competidores, etc

- **Agent based models (ABM):** estos modelos hacen una simulación de un mercado virtual para inferir los factores que determinan una categoría (fabricantes de coches, refrescos, etc) y poder cuantificar los factores que explican el volumen de una determinada marca. A diferencia de las 2 alternativas previas, son capaces de considerar variables a nivel de target/segmentos y con información experta no necesariamente registros reales.

En cualquiera de estos enfoques es importante considerar todos los tipos de medios para conseguir una medición correcta y completa de la contribución de todos los tipos de interacciones sobre las personas:

- **Pagados:** televisión, paid search, programática, exterior, radio, etc
- **Ganados:** vídeo online orgánico, social orgánico, etc
- **Propios:** emails, posicionamiento orgánico, etc

Partiendo de cualquiera de los enfoques previos se pueden realizar algoritmos matemáticos que permitan realizar simulaciones y/o predicciones de cuál será el kpi en los próximos meses, etc dados ciertos escenarios de inversión en comunicación. Y también es posible resolver modelos de optimización, por ejemplo, dado un objetivo de negocio y unas restricciones ¿cuál es el mejor mix de inversión para maximizar dicho kpi objetivo?



3.5.- Visualización

En los últimos años ha aumentado el interés por todo lo relativo a la visualización de datos y a la extracción de insights, especialmente debido al gran volumen de información que se genera en el mundo digital (no sólo las webs, sino toda la conectividad y registro adicional de información en todas las transacciones de la vida diaria).

Por ello, algunas utilidades básicas del tratamiento y análisis de los datos son:

- Los dashboards. Entornos de reporting que ayudan a entender lo que ha ocurrido y tomar decisiones de negocio
- Entornos de consultas ad-hoc. Se han extendido las herramientas que a través de interfaces de usuario, evitando la necesidad de conocer código o ser capaz de programar, permiten realizar análisis de datos al aplicar ciertos filtros
- Funcionalidades de data Discovery para realizar exploración de datos también a través de interfaces de usuario

4 Casos de uso_

4.1.- Miembros del hogar y match con categorías IAB

Objetivo, definición y etapas

Objetivo: Establecimiento de una metodología para la identificación y tratamiento de conjuntos de dispositivos en lo que llamaremos **households ID (HHID)**.

Definición: Household, de forma muy breve, como aquella agregación de dispositivos cuyos usuarios tienen una relación de convivencia y, por lo tanto, representan intereses de consumo comunes.

Es evidente que tendremos muchos modelos distintos de household, y por ello, la metodología propuesta considera esta diversidad en prácticamente todas sus etapas.

Definición de estapas en el proceso:

1. Identificación de los hhid
2. Gestión de los hhid
3. Enriquecimiento de la BBDD de hhid
4. Explotación de los hhid

1.- Identificación de los HHID

Aproximación observacional para la identificación de los hogares. A partir de los datos de consumo y navegación diarios se definen una serie de reglas que permiten ver relaciones entre dispositivos.

Dichas relaciones permiten establecer lazos de convivencia entre los usuarios de dichos dispositivos, ademas estas reglas también están vinculadas a diferentes definiciones de hogar. Esta aproximación permite la inclusión de nuevos modelos de hogar capturando, de esta manera, las complejidades de los modelos de vida de nuestra sociedad.

Utilizamos las siguientes variables:

- Datos de dispositivo (user_agent): para determinar el tipo de dispositivo que está visualizando.
- Datos de red (ip): para identificar dispositivos que comparten la misma conexión a Internet.
- Datos de conexión asociados a la dirección IP.
- Datos de geolocalización de alta precisión provenientes de dispositivos móviles.
- Usuario, en el caso que existan datos de registro de usuario para el acceso a algunos contenidos.

A partir de la aplicación de las distintas reglas se genera un modelo probabilístico sobre el cual se identifican los hogares mediante el establecimiento de un umbral por encima del cual se consideran los dispositivos parte de ese hogar.

Se genera un dataset con los resultados de la aplicación de las reglas con las relaciones entre dispositivos de modo que al final pueden "sumarse" y generar de esta manera un dataset global sobre la que se establece el umbral.

Por otro lado, el modelo se complementa con la aplicación de técnicas no supervisadas para hallar grupos de dispositivos que quedan fuera de las reglas y definiciones propuestas.

Legalmente, es necesario que todos los dispositivos que se incluyan en este proceso tengan concedidos los permisos de análisis y vinculación de dispositivos.

2.- Gestión de los hhid

Se genera una base de datos en la que se almacenan los hhid juntamente con los identificadores de los dispositivos que formarán parte de cada hogar, así como la probabilidad de que ese dispositivo forme parte de ese hogar.

Adicionalmente, también se almacenan etiquetas descriptivas del hogar en cuestión. Dichas etiquetas pueden provenir de la identificación de los hhid, por ejemplo, a partir de distintos algoritmos para su identificación, o de etapas posteriores de enriquecimiento. Por último, también contamos con campos descriptores del HHID.

La base de datos de hhid cumple con las siguientes condiciones:

- Cookieless: las variables a tratar tienen son de origen 1st party data.
- Cumplimiento de privacidad TCF: el sistema considera la aceptación del usuario de la finalidad especial de "vinculación de dispositivos" tal como está definido en la especificación TCF
- Multidevice (móvil, Tablet, PC, Smart TV): esta gestión es compatible con todos los tipos de dispositivos.
- BBDD unificada de hogares con dispositivos asociados: ofrece al publisher la BBDD unificada de todos sus usuarios agregada por hogares.

3.- Enriquecimiento de la BBDD de los hhid

Una vez identificados los hhid, se pueden enriquecer los datos disponibles mediante distintas fuentes.

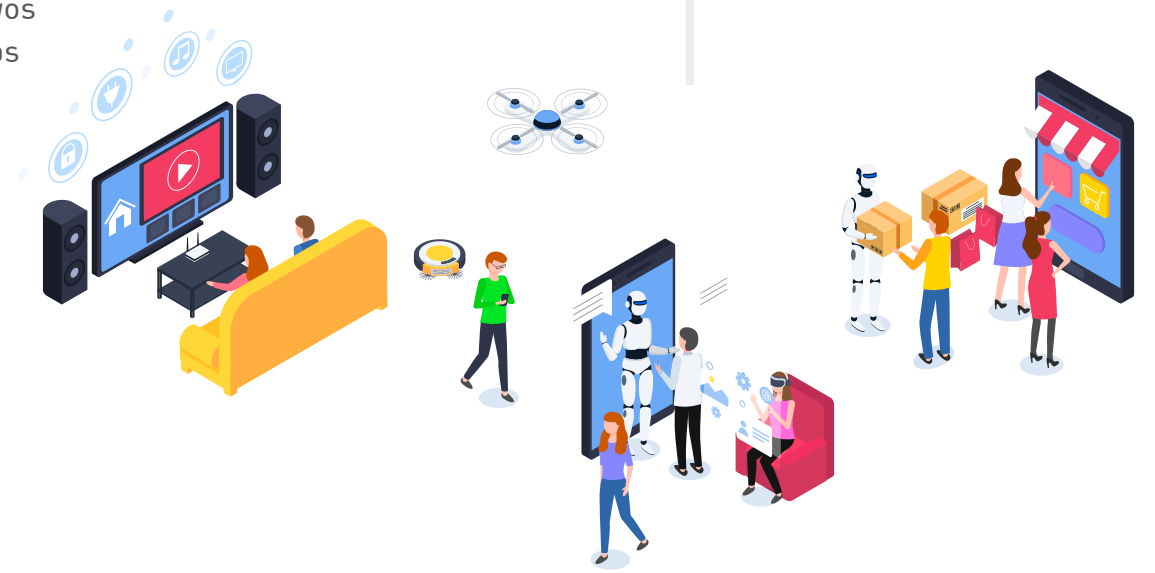
A modo prospectivo, se pueden incluir para el enriquecimiento de la BBDD de los hhid:

- Datos de comportamiento y navegación
- Datos de intereses
- Consumo de contenidos
- Perfiles de hogares
- Taxonomía IAB
- Datos declarativos
- Fuentes externas

4.- Explotación de los hhid

La base de datos de los hhid está accesible mediante un dashboard que puede crear segmentos de interés a partir de las variables y etiquetas existentes. Dichos segmentos pueden exportarse en distintos formatos, siendo posible también la integración con adservers.

- Campañas multicanal orientadas a hogares
- Categorización de los hogares para afinar targetización de las campañas
- Segmentación de dispositivos según criterios de hogares
- Acceso API a la BD de hogares para analíticas in-house



4.2.- Puja de espacios publicitarios y audiencias

La publicidad programática corresponde uno de los avances más notorios en la industria publicitaria, permitiendo automatizar los procesos de compra y venta de inventario a través de plataformas con un sistema de precio basado en puja y que se realiza impresión por impresión, todo ello en el espacio de pocos milisegundos. Es lo que se conoce como RTB.

En esta metodología de compra intervienen tres actores principales: una plataforma de compra o DSP (Demand Side Platform), una plataforma de venta ó SSP (Supply Side Platform) y un mercado donde se transacciona, el AdExchange.

Una de las mayores ventajas que aporta el RTB es que el anunciante, a través del DSP, podrá pujar y comprar impresión por impresión, posibilitando acceder al inventario en función del valor del usuario, la página web en la que se encuentre, y los parámetros KPI aplicados a la campaña, emitiendo una puja por cada request y pagando un precio 'justo' para cada uno.

Es en el DSP donde se aplican modelos de Machine Learning para establecer cuál debe ser el precio de puja adecuado dentro de los parámetros de campaña.

Dado que el modelo de negocio de los DSP es, en la mayoría de los casos, basado en un revenue share de la inversión, la propia utilización de IA en los mismos beneficia, no sólo al anunciante, sino a los ingresos del propio DSP, por lo que el desarrollo de elementos predictivos dentro del DSP ha evolucionado en mayor medida.

Los elementos principales sobre los que se ejecuta un modelo de predicción de puja en un DSP serán:

- **Presupuesto de campaña**, que será el que establecerá el framework inicial para el particionado del mismo y la distribución de este para cada estrategia. Una parte del presupuesto de campaña será igualmente reservado para capturar información previa aleatoria del stream antes de optimizar el algoritmo.

- **Consecución del KPI** de campaña dentro del presupuesto, definiendo la consecución de este (ej.: CTR, clic, o cualquier otra interacción de un usuario) y su comportamiento con relación al precio de puja. Se pueden definir diferentes modelos lineales y no lineales para la predicción del KPI, destinados más a la explotación de este con pujas elevadas por impresiones de alto valor, o bien aumentando el número de pujas con un menor coste.

  El algoritmo deberá adecuarse a elementos de campaña relacionado con el KPI como son la frecuencia y recencia del clic o la interacción del usuario.

- **Análisis del precio de mercado.** En este apartado ha habido muchos cambios, y ha sido necesaria la adaptación de los algoritmos a un sistema de pujas de First Price, en detrimento del antiguo modelo basado en Second Price, y a la información disponible sobre pujas ganadoras ya que en todos los casos el Adexchange informa al DSP ganador y el precio es conocido, pero de igual manera, si no gana la puja, no se recibe información sobre el precio ganador.

  Aún así, cuando se pierde la puja, existe la posibilidad de trabajar con los precios de puja perdedores y los win rates para alimentar con el algoritmo con más información sobre los precios de puja en el mercado.

- **Capacidad de procesamiento de QPS.** El DSP deberá ajustar su capacidad de escuchar bid requests de los adexchanges (traffic shaping) para reducir los costes a la par que se cumplen los requisitos básicos de funcionamiento: consecución de los KPI's y cumplimiento de los budgets de campaña.

Cómo funciona

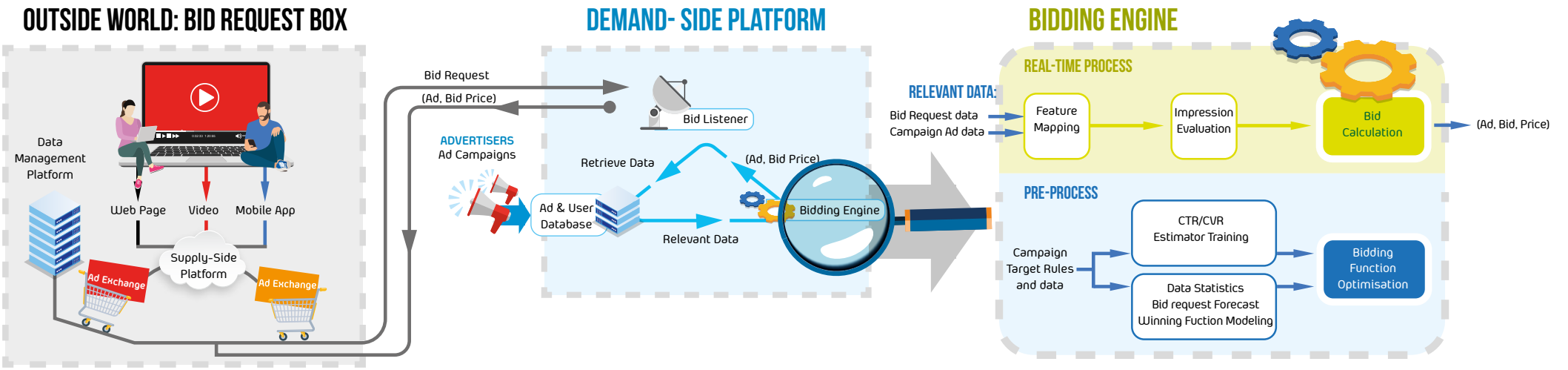
El modelo sobre el que un DSP trabaja viene determinado porque se ejecuta en stream de información por lo que requerirá analizar elementos sobre un contexto histórico de datos para diferentes rangos de temporalidad, junto con el feedback de su aplicación para la campaña activa. Es por ello por lo que es frecuente la utilización de modelos de regresión bayesianos para resolver el cálculo del precio de puja, actualizando sus parámetros con la nueva información recibida.

La data que más comúnmente se utiliza para el entrenamiento del modelo es:

- Bid requests, comprendiendo todos sus elementos (site, emplazamiento, usuario, dispositivo, geo).
- Características de las creatividades.
- Win notice y el precio ganador.
- Loss notice y la puja enviada.

De igual forma se alimenta de datos almacenados relativos a usuarios, tanto de la pertenencia a determinadas audiencias, como el valor promedio de un usuario en determinado momento del día.

La determinación del precio de puja también estará condicionada por el tipo de publicidad que pueda servirse (personalizada o no personalizada), una vez recogido el consentimiento desde el bid request y en función de la información disponible relativa al comportamiento de un usuario (por la existencia de una cookie o su ID).



Fuente: Optimal Real-Time Bidding for Display Advertising
KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining
August 2014 – Varios Autores.
Septiembre 2022 - Rediseñado por IAB Spain

Como uno de los valores de entrada del modelo de cálculo de la puja es frecuente que se incorpore la salida de la ejecución de otro modelo predictivo para la propensión de KPI (ej.: propensión al clic de un usuario).

En un entorno sin cookie de tercero y en el que no se pueda identificar al usuario de ninguna otra forma, el valor de un usuario y el precio de puja que establecerá el motor de puja será menor que cuando sea plenamente identificable, dado que existen menos elementos para el análisis de la puja y se reduce la precisión en la obtención del KPI de campaña.

4.3.- Segmentaciones de audiencias

La **"clusterización"** de audiencias puede ayudar **a extraer información y patrones estadísticos a partir de grupos de audiencia**, agrupados en función de una serie de características similares, basadas en su comportamiento y consumo.

Esta técnica de machine learning no supervisada nos permite agrupar puntos de una distribución en grupos o clústeres en función de su similitud.

La eficacia de esta técnica viene dada por la forma de los clústeres (covarianza de la distribución) o por si el resultado de la clusterización viene expresado como la pertenencia categórica a un clúster, o si lo expresamos como la posibilidad de pertenencia a cada clúster.

Para aplicar efectivamente las técnicas de clusterización, primero debemos abordar la transformación de los datos que podemos usar para formar la distribución de puntos. La información que podemos usar puede ser, por ejemplo:

- **Texto.** Existen algoritmos para tratar información escrita y transformarla en un vector que dé cuenta de las palabras que contiene y su frecuencia. También es posible mantener la similitud entre los vectores correspondientes a diferentes textos con cierto grado de similitud.

- **Información categórica.** Por ejemplo, tags, ya sea como una asignación pesada con diferentes valores, o no.

- **Asignaciones de escala.** Como por ejemplo fechas, asignando valores más pesados a fechas más recientes, o información geográfica, denotada por parejas de valores de latitud y longitud.

- **Imágenes.** Podemos usar algoritmos para vectorizar el contenido de las imágenes, como por ejemplo la información en los canales de la imagen.

Una vez obtenida la serie de vectores en las dimensiones de las transformaciones que hemos elegido, podemos aplicar alguna técnica de clusterización para generar los N clústeres de audiencias similares.

Este número N de clústeres se deduce de la información previa que tenemos sobre la distribución, o puede asignarse arbitrariamente y ver qué número de clústeres ofrece mejores resultados, existiendo algoritmos para estimarlo rápidamente.

Las técnicas pueden variar en función de la forma de los clústeres, para lo que podemos utilizar diferentes métodos. Los métodos más comunes son:

- **"K-means".** Método de clusterización basado en centroides aleatorios, que mediante iteraciones sucesivas se van recolocando, y reasignando los puntos cercanos hasta minimizar las distancias de cada punto al centroide del clúster.
  Este método es escalable para grandes fuentes de datos y garantiza la convergencia de los centroides, aunque solo es eficaz con estructuras lineales y es sensible a la aparición de puntos externos a la distribución (ruido). La forma de los clústeres resultantes es aproximadamente redonda, y puede no ser conveniente según la covarianza de la distribución.

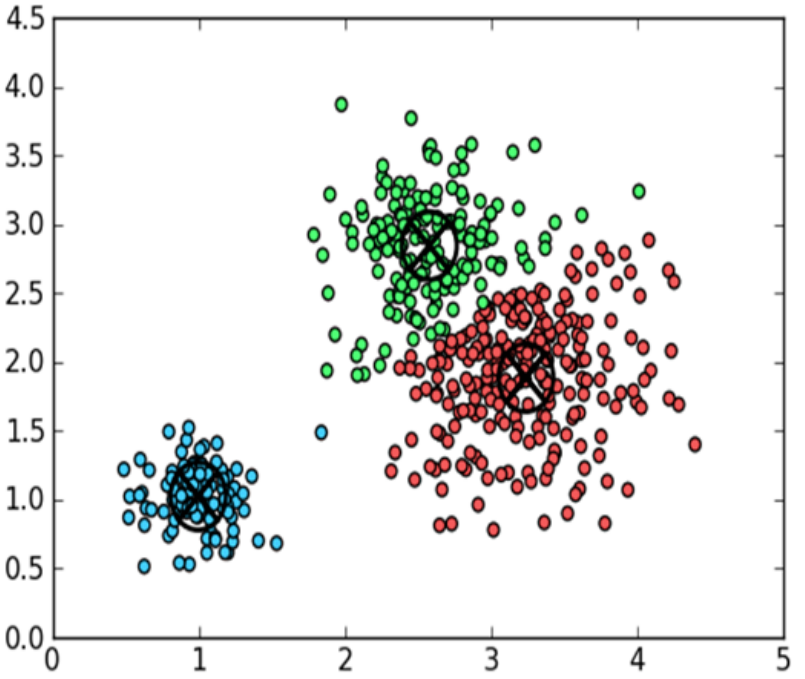


Ilustración 1. Método "K-means" con centroides visibles.

-**"Gaussian mixture model"** (GMM). Esta aproximación a la clusterización de los datos asume que la distribución es una combinación de un número finito de distribuciones gaussianas. Este modelo puede verse como una generalización del método "K-means" para incorporar la información sobre la varianza en la estructura de datos.

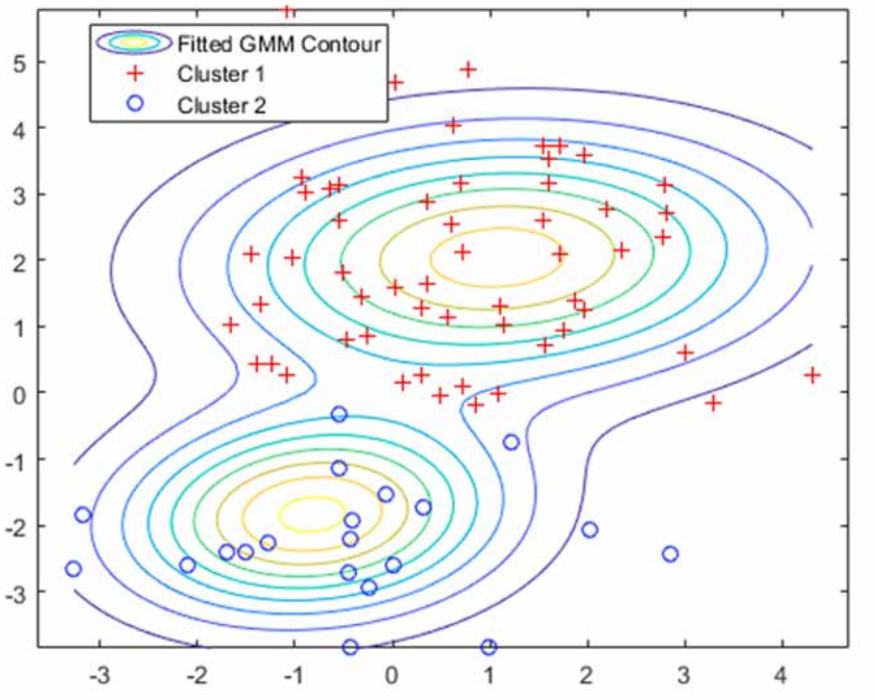


Ilustración 2. Método de mezclas Gaussianas.

Aunque este método nos va a dar cuenta de las posibles formas diferentes de la distribución, es conceptualmente más complejo.

Comparativamente ambos métodos obtienen resultados que dependen de la forma de la distribución y la conveniencia, como vemos en los ejemplos a continuación:

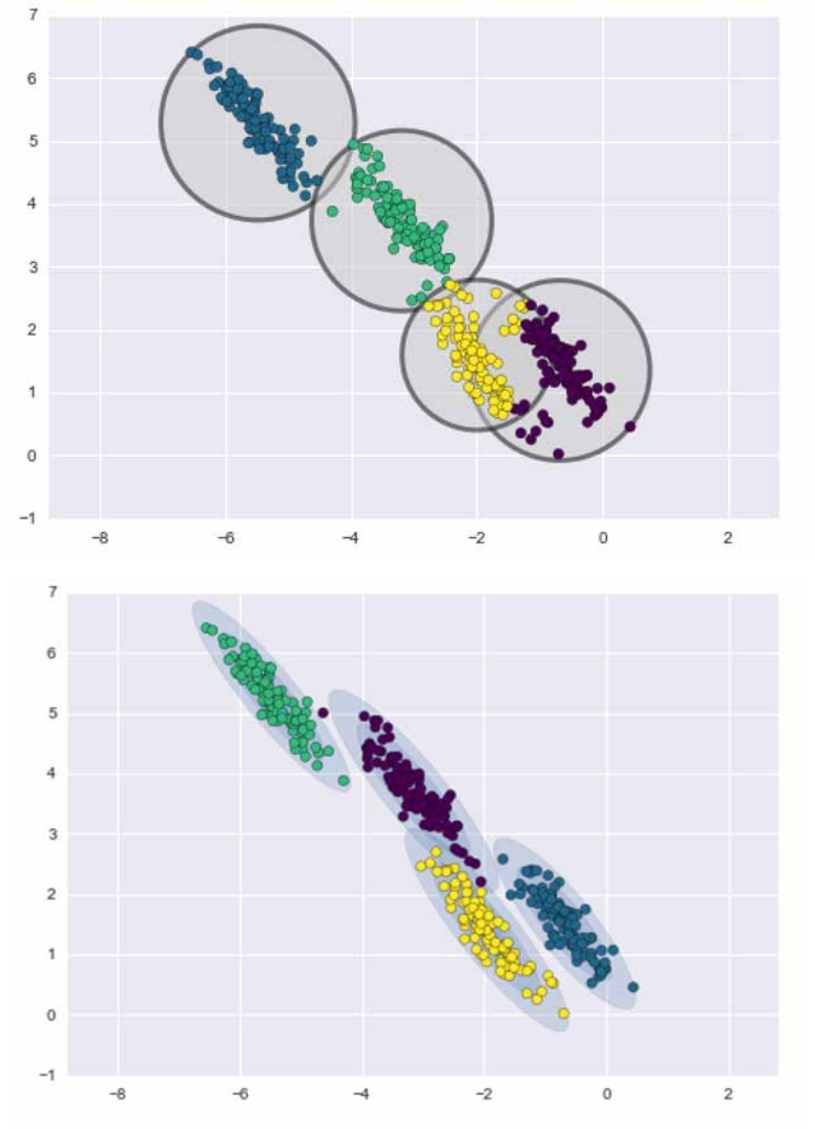


Ilustración 3. Clusterización de la misma distribución. Método "K-means" (izquierda) y método GMM (derecha).

En las imágenes se muestran dos distribuciones idénticas, y tras aplicar ambos métodos se observa que hay una serie de puntos entre los clústeres cuya pertenencia a su clúster cambia según qué método se use.

Además de los posibles métodos que se pueden utilizar para clusterizar datos, existen algoritmos cuyo propósito es acelerar la convergencia de la posición de los centroides, o para asignar el número de clústeres que más conviene al estudio de los grupos de audiencia.

Por último, una vez asignados los clústeres queda analizar la información que contiene cada uno. Esto podemos hacerlo aplicando las transformaciones inversas a las primeras que hicimos, pero esta vez sobre datos promedio de cada clúster.

A nivel de aplicación de este método a un modelo de negocio, habría que garantizar la anonimidad de la audiencia. Esto podemos conseguirlo estudiando la distribución y su dimensionalidad, de forma que podamos elegir un número N de clústeres adecuado, que garantice un mínimo de integrantes en cada clúster, de manera que en vez de que podamos extraer patrones estadísticos del grupo, no de los individuos.

Por ejemplo, podríamos obtener una serie de tags pesados, correspondientes a los que más aparecen en cada clúster y que denoten los intereses generales del clúster en su totalidad, de modo que para los próximos usuarios que, por sus características en las dimensiones

de la distribución, sean asignados a ese clúster, tengamos una serie de características para definirlos según su grupo.

En resumen, las técnicas de clusterización nos van a permitir agrupar la audiencia en grupos cuyos intereses promedio podemos extraer sin conocer a priori información precisa sobre los individuos que componen la audiencia, ni los grupos que forman en la distribución según las dimensiones que hemos elegido.

4.4.- Activación de audiencias basada en la geolocalización

La geolocalización es clave para entender los hábitos, intereses y necesidades de las personas.

Para el sector retail el potencial de la geolocalización es enorme, ya que abre muchas puertas de personalización, segmentación de audiencias y activación que les permite optimizar y eficientar sus campañas.

Haciendo uso de modelos de Machine Learning, en un cliente de retail, hemos desarrollado un modelo que nos permite optimizar la planificación de campañas, en base a visitas a tienda (online y física).

En base a los productos, los resultados históricos obtenidos, la estacionalidad y otros factores contextuales, la plataforma bajo la que se implementan estos modelos de Machine Learning, recomienda a cada director de tienda cuáles son las mejores zonas y presupuestos óptimos para lanzar la campaña en base a todos los datos históricos analizados.

La implementación de estos algoritmos, sobre su plataforma de activación, les ha llevado a obtener las siguientes mejoras en el último año:

- +67% incremento tráfico a punto de venta
- -37% CPC en Paid Social
- X3 CTR en prospecting

4.5.- Atribución cookieless

Hace no tanto, 3-4 años, muchas marcas tenían entre sus objetivos principales la puesta en marcha de una solución avanzada de atribución digital cross plataforma basada en modelos matemáticos para medir la efectividad de las campañas, y así poder mejorar el proceso de optimización de su actividad digital.

Sin embargo, desde entonces la situación ha cambiado significativamente y pocas marcas barajan hoy en día implementar un sistema de atribución digital que realmente cubra toda su actividad digital.

Pero ¿por qué este cambio? ¿por qué ya no se contempla la atribución digital basada en modelos matemáticos?, básicamente debido a tres factores:

- El primero fue el RGPD, en aplicación desde mayo de 2018, que ha traído mayor rigor en la gestión del consentimiento explícito de las cookies para poder registrar su actividad.

- El segundo es el avance continuo de los walled gardens. Fundamentalmente Google, Facebook y Amazon que aglutinan cada vez más inversión digital y restringen cualquier posibilidad de acceso a la visión de la actividad de una cookie dentro de sus plataformas.

- El tercero es el crecimiento de las políticas de privacidad para facilitar el bloqueo de la huella digital que los usuarios dejamos a través de los navegadores. Firefox y Safari (Apple) fueron los primeros en promoverlo, y Google finalmente se ha visto en la tesitura de seguir está tendencia del cookieless world.

Estos factores han hecho que disponer de información completa o consistente a nivel de cookie no sea posible porque:

- Las cookies entre walled gardens y fuera no se pueden consolidar.
- No siempre es posible identificar a un mismo usuario cuando cambia de dispositivo / navegador.
- El uso de la navegación en modo incógnito.
- La evolución de los propios navegadores que están limitando los pixeles, inhabilitando los que son de terceros y limitando las cookies a aquellas de 1st party aún más para los próximos meses.
- No siempre se dispone de los clics orgánicos atribuidos salvo a través de las herramientas de analítica.
- etc.

En este contexto, y dado que la atribución digital a través de modelos matemáticos se basa en las cookies, la realidad es que no se dispone de datos adecuados para poder realizar este tipo de análisis.

Pero entonces ¿qué ocurre con la medición y atribución digital basada en reglas de negocio como last clic, etc?. Actualmente, el estándar entre las marcas es utilizar los módulos de atribución del adserver (Google Campaign Manager, Sizmek, Adform, etc) así como sus herramientas de web analytics (Google Analytics con su módulo de Attribution y Adobe, fundamentalmente) para asignar la atribución a determinados canales. Sin embargo, las conversiones medidas a través del ad-server están también impactadas por el hecho de no poderse medir el post-impression en todos los casos por lo que las herramientas de web analytics están ganando peso para considerar y analizar todas las conversiones. Es decir, de alguna forma se ha vuelto a la casilla de salida, se puede medir la evolución del kpi de conversión a través de web analytics, pero el disponer de la traza o el customer journey de la cookie previa a la conversión se ha limitado notablemente frente a la

situación de hace unos pocos años. En paralelo, en los últimos meses Google ha confirmado lo que la industria ya sabía, que el last clic no es suficiente como medida de atribución y está impulsando su solución algorítmica de Data Driven dentro de su plataforma o walled garden. Pero dichos aprendizajes y algoritmos por plataforma están sesgados porque no consideran todos los impactos publicitarios, tanto digitales como no, que recibe una persona. Es decir, ignoran el efecto de la comunicación recibida fuera de su plataforma y toda la conversión se atribuye a las activaciones en su plataforma.

Además para resolver las limitaciones de algoritmos de atribución basados en información parcial de cookies, aparecen algoritmos capaces de medir el impacto de la comunicación no basados en cookies que además permiten considerar todos los impactos multimedia y no solo los digitales. Este tipo de metodologías de Machine Learning para realizar la atribución multimedia o crossmedia son fundamentales ya que NO:

- Dependen de cookies o de datos basados en IDs.
- Analizan sólo desde la perspectiva de 1 plataforma, ya que la suma de las atribuciones por plataforma NUNCA genera el total de conversiones de una compañía.
- Ignoran que para las marcas con cierta trayectoria, no todo depende del efecto de la comunicación en el corto plazo, si no que la construcción de marca posibilita tener conversiones no dependientes de los medios, lo que se conoce como línea base.
- Ignoran el impacto de los canales propios y ganados, centrándose solo en los canales pagados.

En definitiva, este tipo de algoritmos permite dar respuestas a diferentes cuestiones:

- ¿Qué parte del KPI no depende de los medios (línea base o baseline)?
- ¿Cuánto KPI está aportando cada medio o disciplina, tanto offline como online?
- ¿Cómo de eficiente es cada medio?
- ¿Cuánto KPI está aportando cada soporte de cada medio?
- ¿Cómo de eficiente es cada soporte?
- ¿Qué elementos funcionan mejor en la planificación : día, creatividad, etc?

Y de esta forma ajustar el plan de activación de la comunicación para optimizar los resultados.

Entendimiento Negocio/Tecnología
Objetivo: es poner en evidencia las posibles fricciones que hay en llevar proyectos de tecnología y más en detalles de machine learning y cómo resolverla.

¿Data es tecnología o marketing?
Must: definir la paciencia para negocio versus inteligencia artificial. Como llevar estos insight a negocio

5 Glosario de terminos_

Glosario de términos_

Adexchange:
Se podría definir como un marketplace digital de anunciantes y soportes, donde se compran y venden impresiones publicitarias digitales mediante pujas en tiempo real. Normalmente, se suele utilizar para vender impresiones de publicidad en display, vídeo y mobile.

Adserver:
Un adserver es básicamente un servidor de anuncios. Cuando hablamos de adserver nos referimos a un "Third Party adserver". Por definición los adservers son tecnologías que se contratan a terceros para gestionar la publicidad digital. Se trata de un tipo de software especializado, usado tanto por agencias como por soportes, al que se accede vía browser, y que se encarga de servir los anuncios en los diferentes espacios publicitarios disponibles en los medios online.

Análisis morfológico:
Básicamente consiste en determinar qué clase de palabra o categoría gramatical forma cada palabra en una frase. Es importante no confundirlo con el análisis sintáctico, donde lo que se analiza es la función que cumple una palabra en una oración.

APIs:
API significa "interfaz de programación de aplicaciones". En el contexto de las API, la palabra aplicación se refiere a cualquier software con una función distinta. La interfaz puede considerarse como un contrato de servicio entre dos aplicaciones.

Arquitectura Data Mesh:
Una arquitectura data mesh busca optimizar el valor de la data de tipo analítico atacando varios aspectos críticos: el cambio constante de los datos disponibles, la proliferación de fuentes y consumidores de datos, la diversidad en los tipos de transformaciones y procesamientos para cada caso de uso y la agilidad necesaria.

Para ello se focaliza en la descentralización y la distribución de responsabilidad en los perfiles más cercanos a los datos para soportar escalabilidad y cambio continuo. Y esto se realiza trasladando los diferentes dominios empresariales y de negocio a la gestión de los datos. Migramos la arquitectura tradicional a una arquitectura de datos centrada completamente en el ownership de los diferentes dominios.

Bid request:
Una solicitud enviada desde una plataforma de suministro (SSP) o intercambio de anuncios a un licitador (parte de una DSP). La solicitud contiene diversos datos sobre el contexto del anuncio (contenido de la página, URL, etc.) y el usuario (por ejemplo, datos de cookies). Sobre la base de los datos de la solicitud de oferta, el ofertante puede cotejarla con los criterios de la campaña del anunciante y decidir si ofertará o no por la impresión.

Brand suitability:
Su objetivo no es solamente evitar que se asocie a las marcas con contenidos perjudiciales en entornos inadecuados, sino que la publicidad llegue a los lugares con mayor probabilidad de captar la atención del público objetivo.

Centroide:
En geometría, el centroide o baricentro de un objeto X (una forma geométrica continua), es

la intersección de todos los planos (en las dimensiones que correspondan) que dividen X en dos partes de igual tamaño (distancia, área, volumen, etc). En el caso de una distribución discreta es el punto que minimiza la distancia al centroide de todos los puntos de la distribución.

Chatbots:
Un chatbot es un programa informático que simula y procesa conversaciones humanas (ya sea escritas o habladas), permitiendo a los humanos interactuar con dispositivos digitales como si se estuvieran comunicando con una persona real.

Chief Data Officer:
Responsable de asegurar la calidad de dato y el buen uso del dato en la organización. Lidera las iniciativas de gestión de datos entendiendo el valor del dato y su aporte de negocio.

Clusterización:
Es una técnica utilizada en minería de datos (dentro del área de la Inteligencia Artificial) para identificar de forma automática agrupaciones (clústeres) de elementos de acuerdo con una medida de similitud entre ellos. Esta técnica también se conoce como segmentación.

Convergencia:
Define el proceso tras el cual, al iterar un proceso matemático, partiendo de variables aleatorias, terminamos llegando siempre al mismo valor. En ese caso el algoritmo converge a ese valor.

Covarianza:
En estadística la covarianza es una medida de la variabilidad entre dos variables. El signo de la covarianza muestra la tendencia en la relación lineal de las variables, aunque su magnitud es difícil de interpretar ya que no está normalizada.

Data Analyst:
Se encarga de procesar en último término la información proporcionada por los científicos de datos y obtener conclusiones que ayuden a mejorar resultados por lo que tienen un conocimiento muy cercano al negocio o al caso de uso que se está trabajando. Es capaz de traducir las salidas de estos modelos/ algoritmos en conocimiento y valor para el negocio facilitando la toma decisiones. Más allá de una formación concreta, se trata de la habilidad para interpretar los datos y darle una perspectiva y aplicación de negocio y proporciona vías de visualizar los datos de forma amigable y relevante. Algunas de las herramientas que manejan son Excel, PowerBi, Tableau, Data Studio, etc

Data Arquitect:
Diseña y elabora la arquitectura necesaria para almacenar la información incluyendo la estructura de los programas con los que se recogen, procesan y almacenan la información. Por tanto, diseña las bases de datos y también las gestiona, dimensionándolas y configurándolas de manera que las queries sobre las mismas son ágiles. Algunas de las herramientas que manejan son MapReduce, Hive, Pig, Spark, Flink, SQL (MySQL, PostgreSQL) y NoSQL (Hive, Redis, ElasticSearch, etc), Java, Scala o Python.

Glosario de términos_

Data Engineer:

Se encarga de proporcionar los datos de una manera accesible y apropiada a los usuarios y Data scientists. Es un perfil especializado en en la ingesta de datos desde las fuentes primarias a bases de datos. Desarrolla y explota técnicas, procesos, herramientas y métodos que deben servir para el desarrollo de aplicaciones Big Data. Tiene un gran conocimiento en gestión de bases de datos, arquitecturas de clusters, lenguajes de programación y sistemas de procesamiento de datos. Maneja con soltura entornos clásicos de bases de datos relacionales (MySQL o PostgreSQL). Algunas de las herramientas que manejan son Google Cloud Platform (GCP), Amazon web services (AWS), Microsoft Azure, etc .

Data Orquestration:

Puede ser confundido como un rol pero en realidad es un proceso y / o disciplina que ayuda las organizaciones a administrar datos de una manera que reúna los datos correctos para el propósito correcto ha sido un tema de administración de sistemas durante décadas, aunque de maneras que no son tan efectivas como se pensó inicialmente. Esto puede hacer que en las organizaciones, según necesidades, alguien del antiguo departamento de sistema se pase a ocupar de Orquestration haciendo nacer un rol que podríamos denominar como data orquestrator.

Data Product Owner:

En el core de este cambio de paradigma está pasar a entender la data como producto. Seguimos contando con equipos de data engineers transversales, responsables de la infraestructura a nivel global, pero cada dominio pasa a tener un responsable, el Domain Data Product Owner. Este perfil es responsable, dentro de su dominio, de:

Calidad y riqueza de los datos

- Disponibilidad de la data para su consumo
- Reducción de tiempos de peticiones de datos responsabilidad de su dominio
- Satisfacción global de sus clientes internos
- Ciclo de vida de los datasets en su dominio
- Definición de los KPIs de los data products

Para cumplir con su misión este perfil necesita contar con un conocimiento profundo del dominio del que es responsable, de sus clientes y de cómo consumen éstos los datos.

Data Scientist:

Analiza, trata y modeliza los datos para construir los algoritmos que responden a la necesidad de negocio. Este perfil requiere combinar conocimientos matemáticos/ estadísticos y conocimientos informáticos para ser capaz de extraer conocimiento e información valiosa de los datos, ya que deben tener una visión general del proceso de extremo a extremo y realizar la construcción de modelos analíticos y algoritmos acorde a la necesidad de negocio. Algunos de los lenguajes de programación más comunes que manejan son R o Python, entre otros.

Data Steward:

El rol de un administrador de datos tiene la tarea específica de mantener el control de datos en las iniciativas de gestión de datos maestros y gobierno de datos en el día a día. La

Glosario de términos_

administración de datos es necesaria para que la implementación y la gestión de datos tengan éxito. Un ejemplo de lo que pueden hacer para lograr esto es redactar las reglas de calidad de datos con las que se miden sus datos.

Datos declarativos:

Los datos declarativos nacen de aquella información que declara un usuario en el momento en que nos deja sus datos. Éstos pueden ser sociodemográficos (edad, sexo, población, idioma, estatus civil, etc.) o según sus intereses (nos indica cuáles son sus gustos y preferencias).

Datos estructurados:

Son archivos de tipo texto que se suelen mostrar en filas y columnas con títulos. Son datos que pueden ser ordenados y procesados fácilmente por todas las herramientas de minería de datos.

Función Sigmoide:

La función sigmoide transforma los valores introducidos a una escala (0,1), donde los valores altos tienen de manera asintótica a 1 y los valores muy bajos tienden de manera asintótica a 0.

Regresión logística:

Es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula.

Valores atípicos:

Un valor atípico es una observación extrañamente grande o pequeña. Los valores atípicos pueden tener un efecto desproporcionado en los resultados estadísticos, como la media, lo que puede conducir a interpretaciones engañosas.

Varianza:

La Varianza es una medida de dispersión que se utiliza para representar la variabilidad de un conjunto de datos respecto de la media aritmética de los mismo.

Agradecimientos_

Adsocy: Domingo Paillet

Datmean: Salvatore Cospito

Dentsu: Carlos Olmo

Dentsu: Sonia Casado

Konodrac: Jordi Gilabet

Labelium: Alex Masip

Rebold: Victor Templado

Wemass: Alberto Gonzalez