CAPÍTULO V

Cómo conseguir tu objetivo *offline* utilizando la navegación del usuario

Jesús Romero Leguina

En este capítulo profundizamos en la creación de audiencias y su óptima activación dentro de nuestras estrategias omnicanales. Combinando la información de Google Analytics junto con el uso de técnicas de *machine learning* somos capaces de generar audiencias y cualificar aquellos usuarios de mayor valor para el negocio. Veremos cómo podemos hacer crecer los negocios integrando las ventas *offline* en nuestras estrategias digitales, gracias al uso de la tecnología y a los modelos matemáticos.

Palabras clave: marketing omnicanal, *machine learning,* ventas *offline*.

1. INTRODUCCIÓN

El concepto de *marketing* omnicanal resulta hoy en día un concepto lejano para muchas compañías. A pesar de que se lleva años hablando sobre ello y en el sector se entiende la importancia de que los canales *online* y *offline* estén conectados entre sí, las empresas y expertos de *marketing* se siguen encontrando con ciertas barreras que les impiden construir un diálogo de cooperación entre ambos.

Una de las grandes limitaciones que nos encontramos es la que trata los objetivos de conversión. Tradicionalmente, las campañas de *marketing* digital tienen como objetivo conseguir acciones en la web por parte de los usuarios, pero como veremos más adelante en este capítulo, estos no son los únicos objetivos de las compañías. Para que los negocios crezcan es importante que sus estrategias de *marketing online* persigan de una forma u otra los objetivos de conversión de la compañía independientemente de cuál sea el canal de venta.

En este capítulo, veremos cómo optimizar la estrategia de *marketing* digital para conseguir un aumento de nuestros objetivos de venta *offline*. Profundizaremos sobre todo en la parte más técnica y la metodología. Desarrollaremos el proceso de trabajo desde la recogida de datos de navegación de los usuarios, pasando por la creación de los modelos matemáticos que cualifican las probabilidades hasta la activación.

A lo largo del capítulo se usará a modo ilustrativo un caso de uso en el que realizaremos el proyecto de optimización sobre una campaña de *marketing* digital para una compañía inmobiliaria. En este contexto, se asumirá que esta compañía trabaja con el conjunto completo de las herramientas de *marketing* de Google. Se usará Google Analytics como fuente de datos, GMP (Google Ads, Search Ads 360, Display and Video 360 y Campaign Manager) para la activación y Google Cloud Platform como la infraestructura para todo el análisis de datos, así como la construcción y el despliegue del modelo matemático.

2. PLANTEAMIENTO DEL PROBLEMA

A pesar de que el objetivo principal de toda compañía es generar negocio, tenemos que entender que esto no significa lo mismo para todas ellas. Identificar el objetivo de conversión de las empresas es el primer paso para lograr el éxito de nuestras estrategias. El siguiente paso es entender el embudo de conversión y los diferentes canales de venta, la relación que mantienen entre ellos y la prioridad que tiene cada uno de ellos dentro del negocio.

Cuando entramos a analizar los diferentes puntos de venta que una compañía puede tener es muy habitual encontrar compañías que tienen modelos híbridos. Es decir, la venta de sus productos y sus servicios se hace tanto a través de la web como de establecimientos físicos. Algunos casos característicos de este tipo de compañías son las de gran consumo (Carrefour, Walmart, Inditex, etc) u hoteleras (Hilton Worldwide, Marriott International, Radisson Hotel Group…). Todas estas compañías tienen como objetivo empresarial vender tanto en su web

como en los puntos físicos o *call centers*. Por tanto, la optimización de toda estrategia de *marketing* (incluyendo *marketing* digital) tiene que estar dirigida a maximizar las ventas en ambos entornos.

Sin embargo, encontramos otro tipo de compañías, que es en el que nos vamos a focalizar en este capítulo, y que engloba aquellas empresas que utilizan la web como medio para iniciar los procesos de compra. Ejemplo de este tipo de compañías son las empresas inmobiliarias (Vía Célere, Hábitat, Neinor Homes) o de venta de automóviles (BMW Group, Ford, Honda) donde, por la casuística del negocio, el proceso de contratación ocurre fuera de la web. En estos casos la estrategia de las campañas de *marketing* digital suele ser la captación de tráfico y *leads* en la web. Los objetivos tradicionalmente se centran en dar visibilidad a los productos y conseguir registros para después cerrar la venta con otros canales de venta como el *call center.* Pero este sistema no es siempre óptimo. Los registros que se obtienen no siempre distinguen el nivel de interés de los usuarios y no están cualificados. Aquí es donde encontramos nuestro principal punto de mejora estratégica y es por lo tanto lo que se va a trabajar a lo largo del capítulo. Planteamos un objetivo de campañas mejorado en el que la estrategia *online* persiga que esos registros sean de calidad y se traduzcan en ventas finales *offline*.

A continuación, ilustramos con un ejemplo la realidad de una empresa inmobiliaria (la llamaremos HouseSale) y la agencia de medios que le gestiona la activación y optimización de las campañas de medios digitales.

Figura 1.

Embudo de conversión por el que pasan los usuarios para el proceso de compra en la compañía

Fuente: Elaboración propia.

Empezamos analizando el embudo de conversión de HouseSale. En la figura 1 podemos ver el proceso por el que pasan los clientes antes de realizar la compra o alquiler de un inmueble. La agencia de medios debe encargarse de que los usuarios vayan pasando a través de los distintos estadios para conseguir que se realicen las máximas operaciones de compraventa.

De los distintos hitos del embudo, los tres primeros (visibilidad, visita web y *lead* de contacto) son eventos que ocurren en el entorno digital y por tanto la agencia tendrá visibilidad de ellos (y la traza completa de los usuarios) a través de las herramientas de medición. Por tanto, al tener visibilidad y conocer el camino que ha seguido el usuario hasta cada uno de estos objetivos, la actividad de campañas se puede optimizar sin necesidad de integrar otras herramientas. El reto lo encontramos en los dos pasos finales (5 y 6).

Se suele dar el caso en el que, a pesar de estar generando de forma correcta un gran volumen de *leads,* que estos no se estén traduciendo en ventas de inmuebles. En este escenario y desde nuestro punto de vista (el de la agencia de medios) las campañas de *marketing* están funcionando de forma correcta y están optimizadas. Se está consiguiendo dar visibilidad de los distintos inmuebles y, además, estamos también consiguiendo que los usuarios que navegan por la web rellenen formularios contacto. En este punto, la agencia de medios está trabajando de forma correcta, ya que está alcanzando todos los objetivos de los que tiene visibilidad. Además, puede incluso que las distintas ratios como son CPA, CR, CTR, etc. estén incluso optimizadas.

La agencia hablaría con el cliente para decirle que el problema lo tiene con sus oficinas o *call centers,* ya que a pesar de que nosotros llevamos volumen, no están consiguiendo cerrar la conversión. Pero el problema real viene de esta falta de visibilidad sobre el proceso completo. Al no tener la visión de lo que ocurre con el usuario más allá de la web, puede que la agencia esté derivando usuarios de poca calidad y que ya de por sí tenían desde un inicio pocas probabilidades de comprar o alquilar el inmueble. Los usuarios que rellenan los formularios no llegan a completar los siguientes pasos, como concentrar o realizar la visita a un inmueble, enviar una oferta, que la misma sea aceptada, firmar el contrato de arras y finalmente firmar la escritura de la casa.

Nuestro reto como agencia colaboradora de HouseSale es ayudarles a generar negocio, es decir que los inmuebles se vendan. Como hemos visto en el ejemplo anterior, no es suficiente con que los usuarios se registren en la web ya que esto no garantiza el cierre final de la transacción y, por lo tanto, no aporta valor de negocio para HouseSale.

3. NUEVOS OBJETIVOS Y RETOS

Con el reto identificado, lo siguiente que debemos definir es el nuevo objetivo que queremos alcanzar e identificar los obstáculos que podemos encontrar en el proceso.

El objetivo dependerá tanto de la industria en la que trabajemos como del cliente. Por ejemplo, en una empresa de gran consumo podemos estar interesados en llevar usuarios a los

supermercados y que estos compren. Estos son dos objetivos claros, pero en otras industrias pueden ser muy diferentes. Es el caso de las inmobiliarias donde, como hemos visto antes con HouseSale tenemos múltiples objetivos, queremos que los usuarios acuerden la visita de un inmueble, realicen la visita, envíen una oferta, firmen el contrato de arras y compren el inmueble. Podemos ver que en este caso elegir un único objetivo es un poco más complejo puesto que tenemos muchos más objetivos que en el caso de la venta minorista.

Además, dentro de una misma industria el objetivo a conseguir puede variar de un cliente a otro en base a sus intereses. Por ejemplo, en el caso del gran consumo podemos tener un cliente que esté interesado en los objetivos que hemos mencionado (usuarios a los supermercados y que estos compren), pero podríamos tener otro cliente de gran consumo que en lo que esté interesado es que los usuarios vayan múltiples veces al supermercado y en vez del número de ventas esté interesado en que gasten lo máximo posible. En estos dos casos los objetivos son muy dispares, ya que en unos queremos llevar el mayor número de usuarios y que estos compren, en cambio en el otro queremos usuarios que vayan a hacer compras más grandes y con mayor frecuencia.

Volviendo a nuestro ejemplo en el que estamos ayudando a HouseSale, identificar el objetivo final es sencillo: aumentar la venta de inmuebles. Pero existen objetivos intermedios que también queremos optimizar, como ya se ha mencionado estos serán: la reserva de una visita al inmueble, la visita del inmueble, el envío de una oferta y la firma del contrato de arras.

Una vez tengamos claros cuáles son nuestros nuevos objetivos tenemos que analizar qué nuevos retos nos surgen para poder optimizar la estrategia. Empezamos viendo dos retos que van a estar siempre presentes en cualquier escenario.

- El primero de estos retos es la visibilidad de los objetivos. Necesitaremos tener visibilidad de cómo evolucionan los objetivos para ser capaces de entender si nuestra estrategia está funcionando o no.
- El segundo hace referencia a la trazabilidad de los usuarios. En *marketing* digital la optimización de usuarios se basa en tener el conocimiento completo de todos los puntos por los que ha pasado el usuario. Esto nos permite entender cuáles son las acciones y puntos de contacto que funcionan de nuestra estrategia y cuáles no. El problema cuando no integramos el mundo *online* con el *offline* es que no conocemos qué ha hecho un usuario después de salir de la web y ahí es donde se complica la optimización.

Pero, además, de manera específica para nuestro caso identificamos otros dos retos que debemos considerar:

- Por la naturaleza del negocio, el tiempo que transcurre desde la última interacción digital (rellenar el formulario de interés) hasta que tiene lugar cada una de las acciones *offline* es de semanas o incluso meses. Esto dificulta el trabajo sobre las campañas que se tendrán que optimizar contra objetivos que ocurrirán a meses en el futuro, si es que estos llegan a ocurrir.

- El último reto que nos encontramos está asociado a la trazabilidad en digital. Los usuarios de HouseSale dedican mucho tiempo a investigar antes de realizar la compra, realizando muchas acciones en la web. Pero nos encontramos con que estas interacciones ocurren desde múltiples dispositivos (móvil, ordenador, *tablet,* aplicación). Es por esto por lo que ser capaces de captar toda la navegación en los distintos dispositivos es muy relevante ya que nos dará mucha información relevante si conseguimos solventar el problema de trazabilidad entre dispositivos.

4. SOLUCIONES PROPUESTAS

Ahora que ya tenemos claro qué es lo que queremos optimizar y cuáles son los distintos retos que tenemos que superar, es el momento de diseñar una solución que nos permita conseguir ambas cosas. A continuación, vamos a ver cómo podemos, de forma general, superar los dos primeros obstáculos que hemos identificado y que son comunes a todas las industrias y clientes.

- Empezaremos por el reto de la visibilidad. De forma general existen dos vías para solventar este problema:
  - La primera solución sería tener un informe compartido por parte del cliente y que esté siempre actualizado. Es decir, un informe en tiempo real o diario en el que nos informe de cuánto hemos conseguido de cada uno de los objetivos. Esto nos permitirá optimizar la estrategia de medios a alto nivel como optimizar, por ejemplo, la distribución de presupuestos entre canales y días de la semana. También permite trabajar con modelos avanzados de tipo econométrico.
  - Por otro lado, lo que podemos hacer es subir las consecuciones de cada uno de los objetivos a las plataformas de medios, esto exige que haya trazabilidad del usuario, pero nos permitirá que las herramientas optimicen la inversión de medios a un nivel más táctico de forma similar a como lo haríamos con cualquier otro objetivo digital.
- Para el segundo obstáculo, que era el problema de la trazabilidad, lo más sencillo es utilizar un evento de registro en la web que conecte los dos mundos. De esta forma, si luego asociamos las acciones que haga el usuario fuera de la web al ID de la base de datos, tendremos toda la trazabilidad del usuario. Lo único que nos queda es subir todas las acciones que nos sean relevantes de vuelta a la plataforma de *marketing* en que queramos utilizarlos.

4.1. Trazabilidad *offline*

Veamos cómo solventamos los problemas para nuestro cliente HouseSale. Empezaremos solucionando la trazabilidad de los usuarios fuera de la web. Con HouseSale esto es sencillo

ya que un usuario que muestre interés por un inmueble tendrá que rellenar un formulario con su *email* o teléfono para poder contactarte. Y usaremos esa información para identificar al usuario fuera de la web y así ser capaces de identificar su actividad *offline.* Cuando rellene el formulario se asocia al *email*/teléfono un User ID que registramos en Google Analytics y que quedará asociado con la *cookie.* Después se crea un proceso automático que de forma diaria subirá todos los eventos *offline* generados por los usuarios y los asocia al User ID correspondiente.

4.2. Medición de objetivos *offline*

Este paso, al tener solventada la trazabilidad, consiste en elegir cuáles van a ser los eventos *offline* que queremos subir para cada usuario. En nuestro caso la información de los eventos se subirá solo a Google Analytics, ya que es suficiente para la activación que llevaremos a cabo. Lo eventos que se han detectado que son relevantes en inmobiliarias y, por tanto, para HouseSale son:

1. Visita física: cuando un usuario concreta una visita a un inmueble.
2. Oferta enviada: cuando un usuario envía una oferta para la compra de un inmueble.
3. Oferta aceptada: cuando la oferta enviada por el usuario es aceptada por el propietario del inmueble.
4. Contrato de arras: cuando el usuario firma un contrato de arras.
5. Escritura: cuando un usuario firma la escritura de un inmueble.

De esta forma con la ayuda del User ID y de la *cookie* tenemos la visibilidad para cada usuario de lo que ha hecho desde que ha entrado en la web hasta que adquiere el inmueble. Toda esta información está disponible en Google Analytics para ser consultada, utilizada y explotada por las personas de la agencia.

4.3. Trazabilidad *cross-device*

Para la trazabilidad completa queremos ser capaces de entender el comportamiento de los usuarios en todos sus dispositivos. De forma adicional, intentaremos tener la trazabilidad fuera de la web sabiendo con qué anuncios ha sido impactado el usuario y cómo ha interactuado con ellos.

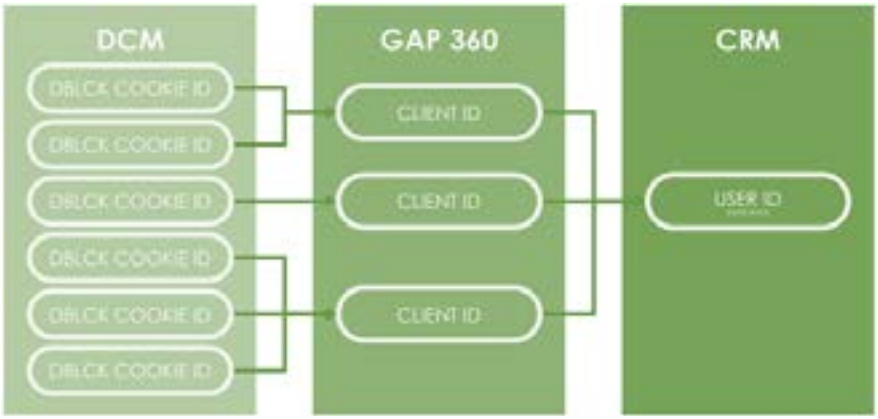
Primero veremos cómo hacemos la trazabilidad entre dispositivos, para esto usaremos el User ID que hemos mencionado previamente. Siempre que un usuario se identifique en un dispositivo sabremos cuál es su User ID y esto nos permitirá saber que es el mismo usuario (*cookie*) en otro dispositivo ya que su User ID es el mismo. Este cruce de datos no se puede

realizar de forma automática en Google Analytics, pero nosotros trabajaremos los datos en Bigquery donde hacer el cruce es sencillo.

Para tener la trazabilidad del usuario en las campañas de *marketing* usaremos un sistema parecido. Lo que haremos será disparar un *floodlight* de Campaign Manager cuando el usuario entre en la web, este evento registrará el Client ID de Google Analytics. De esta forma tenemos conectada la actividad de Google Analytics y la de Campaign Manager. Teniendo todos los datos en Bigquery y usando un esquema de relación entre los ID como el descrito en la figura 2, somos capaces de tener la traza del usuario desde las campañas que le han impactado hasta sus acciones fuera de la web.

Figura 2.

Relación entre los ID identificativos de cada uno de los usuarios en las distintas plataformas para tener una visión única del usuario



Fuente: Elaboración propia.

4.4. Conversiones lejanas en el tiempo

El problema de que el proceso de compra se alargue en el tiempo es que los algoritmos de optimización pueden interpretar que no se están consiguiendo los objetivos. Además, puede confundir al algoritmo sobre cuál de las acciones de *marketing* es la que está trayendo las conversiones. Por ejemplo, las que ocurran entre el contrato de arras y la escritura es probable que no tengan relevancia. Todo esto sumado al problema de que el número de conversiones que tendremos es muy bajo hace que la forma normal de optimización de las campañas con los algoritmos automáticos no sea la más adecuada para este negocio en concreto.

La solución que usaremos es crear un modelo matemático que, usando toda la información que tenemos gracias a la trazabilidad completa, prediga si un usuario va a realizar cada una de las acciones en las que estamos interesados. Una vez tengamos este modelo entrenado

y funcionando seremos capaces de asignar a cada usuario una probabilidad de conversión en base a su comportamiento. Teniendo esta clasificación de todos los usuarios, lo que haremos será trabajar la estrategia de *marketing* en base a audiencias, donde cada audiencia será el grupo de usuarios que tengan más o menos probabilidad de realizar una compra.

5. IMPLEMENTACIÓN

A lo largo de esta sección veremos cuáles son las distintas fases que se deben seguir para llevar a cabo la solución que se ha planteado en las secciones anteriores. Aunque las secciones son comunes para cualquier problema de este tipo, nosotros de nuevo ilustraremos cada uno de los pasos con la ayuda del ejemplo que tenemos con HouseSale. Las distintas fases y su orden es el siguiente:

1. Trazabilidad de la información
2. Importación de los datos
3. Agregación de los datos de usuario
4. Crear variables predictivas y construir tabla final
5. Dividir datos
6. Identificar objetivos
7. Construir modelo
8. Desplegar modelo
9. Disponibilizar resultados
10. Activación en medios
11. Visualización
12. Revisiones

5.1. Trazabilidad de la información

El primer paso es conseguir todos los datos que utilizaremos durante la construcción del modelo. Por lo tanto, lo primero que tenemos que hacer es solventar los problemas de trazabilidad que tenemos (trazabilidad *offline* y trazabilidad *cross-device*).

Para nuestro caso de estudio, decidimos usar Google Analytics como plataforma centralizadora que nos dará la total trazabilidad de los usuarios. Para esto, lo primero que haremos es que cuando un usuario rellene cualquier formulario en la web o se registre se enviará desde

los sistemas de HouseSale a Google Analytics un identificador único para ese usuario basado en el *email* y el teléfono. Este número nos permitirá conectar las actividades entre los distintos dispositivos en los que el usuario realice alguna de estas acciones. Además, este usuario estará presente en todos los registros que ocurran por parte de HouseSale referente a la actividad *offline* del usuario.

Como todos los eventos *offline* están asociados al ID de usuario que también tenemos en Google Analytics, podemos utilizar este ID para utilizar *measurement protocol*. De esta forma, subiremos a Google Analytics los eventos relevantes para el modelado de formas que estén disponibles en la plataforma y asociados al usuario correcto. Haremos la subida de cinco eventos distintos que son relevantes para nosotros:

- Visita física al inmueble
- Envío de oferta
- Aceptación de la oferta
- Contrato de arras
- Escritura

Para el proceso de informar los distintos eventos lo que haremos será la subida automática de forma diaria de todos los nuevos eventos que se hayan registrado de forma manual en los sistemas de la compañía.

5.2. Importación de los datos

Un paso sencillo de elaborar, pero crucial para la ejecución de los siguientes. Lo que haremos durante este paso será la importación en la base de datos en la que vayamos a trabajar de todos los datos de navegación y *offline* que tenemos disponibles gracias al paso previo.

Para nuestro caso de uso, donde usaremos como base de datos Bigquery y teniendo todos los datos disponibles en Google Analytics, la única tarea necesaria es la activación de Export de Google Analytics que nos enviará de forma diaria todos los eventos que hayan ocurrido en la web.

5.3. Agregación de los datos de usuario

Una vez tenemos todos los datos disponibles en la base de datos tenemos que tratarlos para que tengan un formato para que sirvan a nuestros propósitos.

Lo primero que tenemos que hacer es construir el *path* de usuario (*cross-dispositivo*). De esta forma tenemos disponible la visibilidad completa para cada usuario de los eventos que ha hecho, tanto *online* en los distintos dispositivos como *offline*.

En nuestro caso la construcción del *path* de usuario consiste en crear una tabla de correspondencia entre el ID de usuario con los ID de *cookie* de cada uno de los dispositivos en los que hayamos sido capaces de trazar al usuario. Utilizando esto, somos capaces de sustituir la *cookie* por el ID de usuario. Habiendo hecho esto estamos en condiciones de unificando toda la actividad de cada ID de usuario tener la visión de todas las interacciones de cada uno de ellos.

Una vez tenemos para cada ID de usuario toda su actividad, lo que queremos hacer es agregar la actividad para cada día, esto nos facilitará su análisis.

En el caso de HouseSale donde ya tenemos toda la actividad de cada usuario unificada lo que hacemos es utilizar el *timestamp* del evento para extraer la fecha. De esta forma, haciendo la agregación para cada ID de usuario y cada fecha y sumando el número de eventos de cada tipo tendremos para cada día y usuario todos los eventos que ha realizado el usuario en ese día.

5.4. Crear variables predictivas y construir tabla final

Una vez tenemos todos los datos de cada usuario es el momento de construir las variables que vamos a utilizar para predecir nuestros objetivos, así como las variables que queremos predecir. Como ya hemos mencionado, nuestro problema tiene un importante componente temporal y, por lo tanto, lo que haremos será construir variables basadas en el tiempo. Para cada día y usuario vamos a construir variables que nos digan si cada uno de los posibles even-

Figura 3.

Ejemplo de la tabla previa al entrenamiento del modelo en la cual están disponibles todas las variables descriptivas, así como las que queremos predecir con sus respectivas distribuciones temporales

User ID	*Fecha*	*Sesiones últimos 30 días*	*Inmuebles visitados últimos 30 días*	Clicks *últimos 30 días*	*Sesiones últimos 15 días*	*Inmuebles visitados últimos 15 días*	Clicks *últimos 15 días*
a5dc48245dg5	06/09/2021	7	10	1	7	6	1
a5dc48245dg5	07/09/2021	16	11	5	14	10	1
a5dc48245dg5	08/09/2021	23	18	6	5	13	3
f514t84h56sc	06/09/2021	10	7	2	4	7	1
f514t84h56sc	07/09/2021	7	0	1	1	0	1
f514t84h56sc	08/09/2021	1	0	0	1	0	0
f514t84h56sc	09/09/2021	0	0	0	0	0	0

Figura 3. *(continuación)*

Ejemplo de la tabla previa al entrenamiento del modelo en la cual están disponibles todas las variables descriptivas, así como las que queremos predecir con sus respectivas distribuciones temporales

User ID	*Fecha*	*Sesiones últimos 7 días*	*Inmuebles visitados últimos 7 días*	Clicks *últimos 7 días*	*Compra siguientes 30 días*	*Compra siguientes 15 días*	*Compra siguientes 7 días*
a5dc48245dg5	06/09/2021	3	5	1	1	0	0
a5dc48245dg5	07/09/2021	9	6	1	1	0	0
a5dc48245dg5	08/09/2021	5	6	2	1	1	0
f514t84h56sc	06/09/2021	4	1	1	0	0	0
f514t84h56sc	07/09/2021	1	0	1	0	0	0
f514t84h56sc	08/09/2021	1	0	0	0	0	0
f514t84h56sc	09/09/2021	0	0	0	1	0	0

Fuente: Elaboración propia.

tos ha ocurrido en un determinado periodo de tiempo. De la misma forma, construiremos para cada uno de los objetivos si este evento ha ocurrido o no en los siguientes días desde el día que se está calculando.

Cuando preparamos los datos para HouseSale sabemos que los usuarios navegan mucho antes de decidirse por una casa y, además, sabemos también que el tiempo que toma la consecución de los objetivos *offline* puede ser muy grande. Es por eso por lo que construiremos variables que nos informarán de la actividad del usuario en los últimos 7, 15 y 30 días. Y lo mismo a futuro para los objetivos *offline* que son los que estamos interesados en ser capaces de predecir. De esta forma construiremos una tabla que será la que se utilizará para predecir. Como se puede ver en el ejemplo de la figura 3 tendremos una fila para cada usuario y día. También tendremos tres columnas por cada evento (sesiones, *clicks,* inmuebles vistos, etc.) que estemos trazando para el usuario. Si el evento es un objetivo, entonces tendremos si se ha conseguido en los próximos 7, 15 o 30 días. Mientras que, para el resto de las variables, que serán las explicativas, tendremos cuántas veces ha ocurrido en los últimos correspondientes días.

5.5. Dividir datos

Como en cualquier proyecto de modelado, uno de los pasos más importantes es la división de todo el conjunto en tres distintos *dataset* que son: entrenamiento, validación y test. Donde el mayor porcentaje de los datos se dedicarán al conjunto de entrenamiento, después a validación y el resto a test.

Figura 4.

Porcentajes de división de la tabla completa de datos en cada uno de los *datasets* para el entrenamiento del modelo

70 % Entrenamiento	20 % Validación	10 % Test

Fuente: Elaboración propia.

Estos *datasets* se utilizarán para entrenar los modelos y elegir el mejor. El *dataset* de entrenamiento es el que se utiliza para construir cada uno de los modelos, que después se contrastará con los datos del *dataset* de validación. En base a los resultados del modelo, en los datos de validación se harán correcciones y se volverá a entrenar el modelo. Este proceso se repetirá hasta tener un modelo con el que estemos contentos. El modelo final se utilizará sobre los datos de test, este *dataset* no ha sido utilizado en ningún momento durante la preparación del modelo y, por tanto, son datos reales que el modelo no ha visto y simulará el desplegarlo en un entorno real.

En este paso, las particularidades del caso de uso de HouseSale no son muy grandes, los porcentajes que se utilizarán son 70 % para entrenamiento, 20 % para validación y 10 % para test tal y como se muestra en la figura 4.

5.6. Identificar objetivos

Llega el momento de analizar los datos para confirmar que son útiles y que con ellos se pueden predecir los objetivos en los que estamos interesados. Este paso va a depender de la industria y los datos que tengamos.

Para los datos de la compañía HouseSale empezaremos analizando los objetivos para tratar de ver si tienen la calidad necesaria. Lo primero con lo que nos encontramos es que la correlación entre las ofertas enviadas y las ofertas aceptadas no es muy grande, así que no será muy diferente trabajar con uno o con otro. Lo siguiente que vemos durante el análisis es el número de ocurrencias para los eventos de oferta enviada, oferta aceptada, contrato de arras y escritura. De forma que el único evento que tenemos con un gran número de ocurrencias es la visita física. Además, nos hemos encontrado con otro problema relacionado con el tiempo. El problema con el que nos encontramos es que hemos visto que los eventos posteriores a la visita física no solo tardan mucho tiempo en ocurrir, si no, que el tiempo que le lleva a cada usuario el llegar a ese objetivo difiere mucho de un usuario a otro. Esta discrepancia hace que sea muy difícil fijar cuál es la ventana de tiempo con la que debemos trabajar para predecir si alguien realizará la acción en el futuro.

Basándonos en este análisis lo que hacemos es que vamos a construir el modelo que prediga si el usuario realizará una visita. Esto nos permite fijar el número de días a futuro que

tenemos que seleccionar que en este caso serán 30 y teniendo muchos más ejemplos positivos en el caso de los otros objetivos.

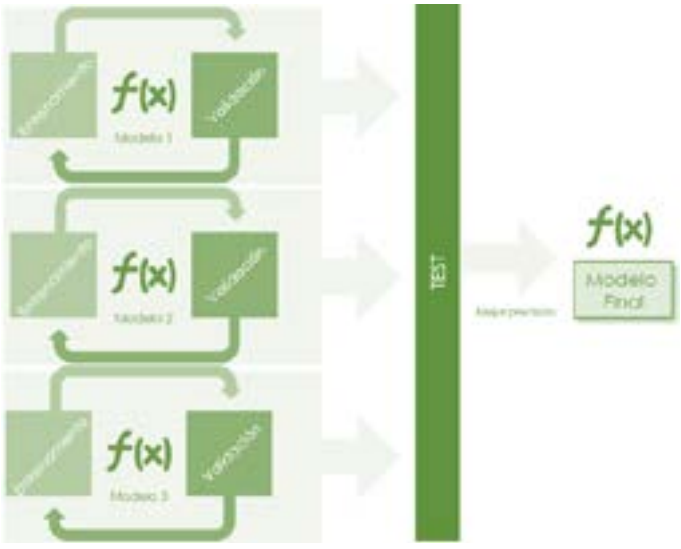
5.7. Construir modelo

Una vez definido el objetivo que queremos predecir y teniendo ya preparados los distintos *datasets*, ha llegado el momento de entrenar el modelo. Para el proceso de construcción (resumido en la figura 5) de los modelos, primero seleccionamos cuáles son las técnicas que utilizaremos para intentar predecir el objetivo. Sobre cada una de las técnicas entrenaremos un modelo que después validaremos. En base a los resultados que obtengamos en la validación cambiaremos los parámetros del modelo y lo volveremos a entrenar. Este proceso lo repetiremos hasta que tengamos un modelo con el que estemos contentos y el cual no queramos iterar más sobre él. Una vez tengamos modelos, en la fase de validación utilizaremos los datos de test para evaluar todos los modelos y de esta forma validarlos con "datos reales". En base a los resultados obtenidos de esta evaluación compararemos los distintos modelos que tenemos y el que mejor funcione en la fase de test es el que seleccionaremos como modelo a utilizar.

Para la comparación de los distintos modelos, que en este caso son modelos de clasificación, se utilizarán las técnicas tradicionales. Se utilizan métricas derivadas de la matriz de

Figura 5.

Metodología a seguir para el entrenamiento del modelo con el uso de los distintos *datasets* y cada uno de los modelos



Fuente: Elaboración propia.

confusión como son precisión, exhaustividad o ratio de acierto. Pero también utilizamos otras métricas como pueden ser el *AUC* o *Log loss* para evaluar el funcionamiento de los modelos.

En el caso específico de HouseSale tenemos todos los datos disponibles en Bigquery así que vamos a poder utilizar toda la potencia que nos ofrece Google Cloud para construir los modelos. Trabajar directamente con los datos en Bigquery quiere decir que el lugar donde hemos hecho todas las transformaciones es también el lugar donde almacenamos los datos para ser explotados por el modelo.

Las técnicas que utilizaremos en este caso para construir el modelo de predicción son:

- Regresión logística
- *Support Vector Machine (SVM)*
- *Naive Bayes*
- Redes neuronales
- *Random forest*
- Árbol de clasificación

Para la construcción de estos modelos usaremos distintas herramientas dentro de Google Cloud para ser lo más eficientes y rápidos en el entrenamiento de los modelos. Usaremos como primera herramienta Google Dataproc que nos permite poner en marcha un *cluster* con el número de núcleos y el tamaño de estos a gran velocidad. Y de la misma forma nos permite eliminar el *cluster* en unos segundos. Además, estos *clústeres* vienen de forma predefinida con todas las herramientas instaladas para poder entrenar el modelo como con Spark o Python. Todo esto nos permite de forma muy ágil levantar un modelo, entrenarlo y después destruir el *cluster,* de esa forma no tendremos ningún coste innecesario y podremos utilizar las técnicas de paralización para que los modelos se entrenen de manera rápida y así llevar a cabo el proceso de forma muy rápida.

Tras realizar todo el entrenamiento y validación se llega a la conclusión de que el modelo que mejor funciona es la regresión logística. Esto es una suerte, ya que es un modelo fácilmente interpretable, lo que es muy útil para explicar a las distintas personas involucradas en el proyecto cómo funciona el modelo y cuáles son los factores más relevantes a la hora de detectar si un usuario realizará la visita o no. Otra gran ventaja que tiene que el modelo sea una regresión logística es que podemos utilizar directamente Bigquery para construirlo. Esto hace que solamente necesitemos un único proyecto para almacenar los datos, transformarlos, entrenar los modelos y hacer las predicciones. Además, la construcción del modelo se hace directamente con sentencias SQL lo que también simplificará su mantenimiento. Tras construir el modelo en Bigquery y confirmar que funciona de la misma forma que lo hacía en Dataproc, se decide que por simplicidad se utilizará Bigquery para construir el modelo.

5.8. Desplegar modelo

Tras haber decidido el modelo que vamos a utilizar, llega el momento de asegurarnos que el modelo se ejecuta de forma que pueda ser explotado. Para esto hay varios procesos que se tienen que automatizar. El primero de todos es el que realiza todas las transformaciones hasta la creación de la tabla final. Este proceso se automatizará de forma que utilizando los datos más actuales nos dé las métricas de entrada para cada usuario en el día actual. Con esto tendremos los datos necesarios para que el modelo trabaje.

El siguiente proceso es la ejecución del modelo. Este proceso tomará la tabla de datos para todos los usuarios y le aplicará el modelo. Esto nos dará para cada uno de los usuarios de la tabla (cada una de las personas que haya navegado por la web en los últimos 30 días) un número de 0 a 1. Este número es la probabilidad que nos dice el modelo de que ese usuario vaya a realizar la acción específica en el tiempo determinado. A este número es a lo que se llamará *score* de usuario. En el caso concreto de nuestro caso de uso el *score* nos dirá la probabilidad de que un usuario realice una visita a un inmueble en los siguientes 30 días.

El último proceso es poner a disponibilidad los *scores* de los usuarios para que puedan ser consumidos por un servicio externo. Una vez estos procesos están automatizados lo que se hace es que se despliegan de forma que se ejecuten de forma periódica o en base a peticiones según se requiera.

Para HouseSale se construye el proceso de creación de la tabla de entrada para el modelo, así como la ejecución del modelo. Para disponibilizar los datos lo que se hace es que se crea una tabla en la que estarán los *scores* actualizados para cada uno de los usuarios que estén dentro del periodo de análisis. Para la casuística de HouseSale la ejecución de todos estos procesos se hará de forma diaria, así que cada día regeneraremos la tabla en la que tendremos los *scores* para cada usuario actualizados.

5.9. Disponibilizar resultados

Esta fase es la que hace que los resultados lleguen a las distintas herramientas que vayan a explotarlos. Estas herramientas pueden ser de *marketing*, para la activación de campañas, de personalización web o cualquier otra herramienta que se beneficie del uso del *score*.

En HouseSale, como se ha establecido desde el principio, el principal objetivo es la activación en medios. Para esto, lo primero en lo que estamos interesados es en enviar los *scores* para que estén en Google Analytics. Para realizar este envío utilizaremos el *data import* para subir los *scores* y asociarlos a cada ID de usuario que ya creamos previamente. Esta información estará en Google Analytics como una variable personalizada que por razones de facilitar el trabajar con ella se subirá un número de 0 a 10 (en vez del valor de 0 a 1 que arroja el modelo).

Además, los datos se dispondrán de otras formas para que puedan ser utilizados para otras explotaciones. Se utilizará Google cloud functions para que el *score* de cada uno de los usuarios esté disponible en tiempo real a través de una petición web. De esa forma el *score* podrá ser utilizado por herramientas de personalización.

La última forma a través de la que se dispondrán los datos es con el envío de un documento a un FTP que contiene la información actualizada del *score* de todos los usuarios. A través del FTP este documento llegará a los sistemas internos de HouseSale donde se convertirá en un atributo más del CRM. De esa forma esta información estará disponible para departamentos como el de *call center* que podrá priorizar sus llamadas basándose en el *score* que tiene cada uno de los usuarios.

5.10. Activación en medios

Para esta fase trataremos directamente el caso de HouseSale. En este caso usaremos Google Analytics como el centralizador de toda la activación en medios. Teniendo la variable personalizada creada se puede utilizar para crear diferentes audiencias con cada uno de los posibles valores que puede tomar el *score*. En la figura podemos ver el número de usuarios que tenemos en cada una de las audiencias.

Figura 6.

Número de usuarios, así como usuarios nuevos en cada una de las audiencias para cada valor del *score*

User score	*Users*	*New users*
0	10.032	669
1	15.867	1.327
2	23.297	466
3	22.652	227
4	13.293	798
5	9.265	371
6	7.266	291
7	6.941	69
8	4.766	95
9	3.795	0
10	1.679	0

Fuente: Elaboración propia.

Es importante conocer el número de usuarios en las audiencias creadas con vistas a la activación. Podemos ver en la figura 6 que el mayor volumen de usuarios se concentra en torno a 2-3 y que tenemos pocos usuarios en los valores altos. Basándonos en estos números tomamos las siguientes decisiones:

Para la activación en SEM, display y social se necesita un número mínimo de usuarios para poder hacer el *retargeting,* por tanto, lo que hacemos es que agrupamos a los usuarios de siete o más para poder dedicar mayor presupuesto y mejores pujas sobre los usuarios que son de alto valor. Todos los usuarios que tengan una probabilidad menor a cuatro serán excluidos de las estrategias de *retargeting* consiguiendo así un ahorro en impresiones sobre usuarios no interesados. El resto de los usuarios seguirán con la estrategia de *retargeting* que se estaba llevando a cabo hasta entonces.

5.11. Visualización

El último paso de trabajo con los resultados es presentarlos para que puedan ser consultados de forma sencilla por las distintas personas. Para esto lo que se hace es crear una visualización de datos donde se muestra toda la información asociada a los *scores.* Es importante mostrar el número de usuarios que hay en cada *score* en el momento actual. También es importante entender cómo los usuarios se mueven desde un *score* a otro a lo largo del tiempo. Y finalmente, si se tiene un modelo que sea fácilmente explicable es conveniente usar gráficos como el de la figura 7 en donde se muestre la importancia de cada una de las variables para predecir el objetivo.

Figura 7.

Importancia de cada una de las variables para la predicción de la visita física a un inmueble

Fuente: Elaboración propia.

En el caso de estudio que estamos siguiendo usaremos como herramienta para construir la visualización a Data Studio. Teniendo los datos en Bigquery es sencillo conectar los datos directamente con la plataforma. En Data Studio mostramos los datos referentes a los *scores* de usuario que se han mencionado previamente. De forma adicional, utilizando los conectores nativos de Data Studio extraemos la información de la plataforma de activación (Search Ads 360 y Campaign Manager) para conseguir también un informe actualizado de cómo es el desempeño de cada uno de los distintos *scores*.

5.12. Revisiones

Es importante notar que el trabajo de modelización es un trabajo que no se puede hacer una única vez. Los cambios en el comportamiento de los usuarios a lo largo del tiempo, en las estrategias de activación o modificaciones de la web pueden hacer que el modelo se quede obsoleto. Es por eso por lo que es importante la creación de una metodología en la que, de forma automática o manual, se revise de forma periódica el desempeño de los modelos para que sean ajustados cuando su comportamiento no sea el esperado.

En HouseSale se crea un sistema que reentrena el modelo de forma mensual para asegurar que esté actualizado. Al utilizar Bigquery para la construcción del modelo este reentrenamiento es rápido y sencillo. A pesar de tener este sistema, también se establecen sistemas de alarma que avisarán si el desempeño del modelo disminuye por debajo de los niveles deseados. Y finalmente, de forma trimestral se establece una revisión más profunda del modelo en la que se evaluará si hay cambios en el entorno que requieran de la modificación del modelo, de los objetivos o añadir / quitar variables predictivas.