UNIVERSIDAD DE OVIEDO

TRABAJO FIN DE MÁSTER

ECONOMÍA CONDUCTUAL. PUBLICIDAD ABUSIVA Y "FREQUENCY CAPPING"

BEHAVIOURAL ECONOMICS. INTRUSIVE ADVERTISING AND "FREQUENCY CAPPING"

Autor

DAVID RODRÍGUEZ LASHERAS

Directores

Amelia María Bilbao Terol

David Fernández Incio

Facultad de Ciencias de la Universidad de Oviedo

Julio, 2021

RESUMEN

Este trabajo tiene como objetivo principal estudiar la publicidad online, y más concretamente de qué forma reacciona un usuario ante repeticiones de la misma, desde el punto de vista de la Economía del Comportamiento. Para ello, se analizan las tasas de conversión publicitarias incorporando al análisis dos sesgos de esta rama de la Economía, el Efecto Marco y el Efecto Recencia. Al realizar dichos análisis, se comprobará la posible existencia de diferencias en los resultados según cada característica comportamental, con el fin de averiguar si se deben aplicar estrategias publicitarias distintas en cada caso. Finalmente se ha propuesto un modelo de árbol de decisión para conocer el número óptimo de veces que se debe enviar el mismo anuncio a un usuario concreto.

ABSTRACT

The main objective of this work is to study online advertising, and more specifically how a user reacts to repetitions of it, from the point of view of Behavioral Economics. To do this, advertising conversion rates are analyzed by incorporating two biases of this branch of the economy into the analysis, the Frame Effect and the Recency Effect. When carrying out these analyzes, the possible existence of differences in the results according to each behavioral characteristic will be checked, in order to find out if different advertising strategies should be applied in each case. Finally, a decision tree model has been proposed to know the optimal number of times that the same advertisement should be sent to a specific user.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1 Contexto

En un mundo cada vez más globalizado principalmente a causa de la revolución digital que tiene lugar desde finales del siglo pasado y principios del actual, surge una evidente necesidad, desde una perspectiva comercial, de conocer al cliente y saber cuáles son sus deseos y cómo satisfacerlos.

Actualmente, la gran mayoría de la población de países desarrollados utiliza algún dispositivo electrónico (ordenador, móvil, tablet, etc.) que les permite acceder a todo tipo de mercados, donde ver o comprar multitud de productos y servicios. Con ello, han surgido nuevas formas de obtener información del consumidor y hacérsela llegar de forma más personalizada.

En este sentido muchas empresas, independientemente del sector y actividad, contratan servicios de este tipo o incorporan un área de análisis que les permite explotar esa información para poder competir en sus mercados y alcanzar sus objetivos.

Una de las áreas más afectadas por estos cambios es la de marketing y publicidad, que ha pasado de hacerse a través de medios tradicionales que llegaban a todo el mundo (televisión, radio, periódico), a utilizar herramientas web cada vez más potentes buscando una segmentación de audiencia que facilita y permite ofrecer un producto a aquellas personas que realmente lo desean.

1.1.1 ¿Demasiada información?

En esta situación, es lógico que las empresas se pregunten si la información que ofrecen es demasiada y cómo afecta las decisiones de compra de sus potenciales clientes.

Surge por ello un concepto relativamente nuevo, *frequency capping* o limitación de frecuencia, con el fin de fijar el número de veces que un consumidor es expuesto a una publicidad concreta en un período de tiempo determinado, buscando evitar la conocida como ceguera publicitaria, que ocurre cuando una persona ignora los anuncios publicitarios que le aparecen.

Además de evitar sobreexponer al consumidor, se desea encontrar la frecuencia publicitaria óptima para que el consumidor sea consciente de ella, ya que se puede dar el caso de que, si únicamente se manda una vez, la persona que la recibe no reaccione y no tenga ningún efecto.

Otro de los objetivos que persiguen los anunciantes limitando la publicidad es optimizar costes, es decir, distribuir su presupuesto de la forma más eficiente posible pagando exclusivamente por publicidad efectiva.

Como se verá posteriormente y por la novedad de este método y las herramientas disponibles para llevarlo a cabo, este problema no ha sido muy estudiado todavía.

1.1.2 Toma de decisiones

Con la digitalización adquiriendo importancia a pasos agigantados, el desarrollo web, avances en inteligencia artificial, algoritmos que pretenden automatizar procesos, etc., parece que las personas quedan en un segundo plano, cuando la realidad es que toda esta innovación está destinada a entender nuestros comportamientos y mejorar nuestras vidas.

Estos comportamientos se han supuesto en muchos casos racionales, es decir, dando por hecho que los agentes económicos tienen un objetivo claro y con los recursos de los que disponen, toman decisiones para alcanzarlo. Para medirlo, se han desarrollado conceptos como la curva de indiferencia o la función de utilidad, que permiten saber de forma exacta lo que haría un agente racional en una situación dada.

Desde la década de los años 70 del siglo XX se ha tratado de explicar el comportamiento económico del ser humano añadiendo el componente psicológico y contradiciendo la teoría clásica anterior. En este caso, se ha comprobado cómo cada individuo toma decisiones en función de las circunstancias que le rodean, sin tener preferencias estandarizadas.

Esta corriente, basada en la teoría de la decisión propuesta por Khaneman y Tverski (1979) (conocida por sus siglas en inglés como *Prospect Theory*) se ha desarrollado principalmente sobre el comportamiento de los agentes económicos en un sector tan cambiante como el mercado de valores, donde la toma de decisiones debe ser en muchos casos rápida.

Dentro de esta teoría se definen los llamados sesgos cognitivos, que utiliza el cerebro cuando no dispone de tiempo para analizar toda la información recibida, condicionando el análisis de la realidad.

1.1.3 ¿Existe relación?

En un primer momento, la Economía del Comportamiento se centró en estudiar el comportamiento de los agentes en un entorno bursátil, pero con el paso del tiempo se está incorporando en otros campos. En este caso concreto, se intentará plantear y demostrar cómo afectan algunos de los sesgos cognitivos a la hora de recibir publicidad online.

Como se ha comentado, el *frequency capping* implica limitar el número de veces que una persona recibe cierta información, dando lugar a una situación en la que dicha persona debe decidir rápidamente si aceptarla o no, en base a una serie de factores, algunos de ellos psicológicos, que entran en juego.

1.2 Objetivos del trabajo

En este trabajo abordaremos las siguientes cuestiones:

- Revisión de los conceptos clave en cuanto al envío de impresiones publicitarias, incluyendo el *frequency capping*.
- Análisis del número de impresiones óptimas a enviar a los usuarios sobre un conjunto de datos que contenga las variables explicativas necesarias, utilizando técnicas de árboles de decisión.
- Utilización de principios de Economía del Comportamiento (Richard Thaler, Daniel Khaneman, Amos Tverski, 2018) para proponer diferentes enfoques de análisis del comportamiento del usuario.
- Comprobación estadística de las diferencias en el comportamiento del usuario aplicando a los datos los sesgos de Economía Conductual.

2. ESTADO DEL ARTE

Como se ha explicado en el apartado anterior, el núcleo del trabajo consta de dos partes; el tratamiento del número de impresiones enviadas a los usuarios, y la Economía del

Comportamiento. Antes de entrar en la explicación de los distintos análisis realizados, es conveniente dedicar este apartado para comentar los estudios relacionados que se han encontrado y la situación actual de ambos temas en el ámbito científico.

2.1 Marketing y desarrollo web

El principal elemento a estudiar y desarrollar en este trabajo, el *frequency capping*, es un concepto muy concreto incluido en otro mucho más amplio como el de "marketing", por lo que es necesario empezar hablando del término acuñado en Estados Unidos alrededor de 1912, y que para los autores Stanton, Etzel y Walker (1992), es "un sistema total de actividades de negocios ideado para planear productos satisfactores de necesidades, asignarles precios, promover y distribuirlos a los mercados meta, a fin de lograr los objetivos de la organización".

La definición que da la Asociación Americana de Marketing (2007) es que el marketing "es la actividad, conjunto de instituciones y procesos que tienen como fin crear, comunicar, entregar e intercambiar ofertas que tienen valor para los consumidores, clientes, socios y para la sociedad en general", y para Philip Kotler (1967), considerado el padre del marketing moderno, se trata de "el proceso social y administrativo por el que los grupos e individuos satisfacen sus necesidades al crear e intercambiar bienes y servicios»

Durante décadas, el marketing se llevó a cabo exclusivamente en medios denominados "tradicionales", hasta que la aparición de internet obligó a incluir un concepto nuevo en la década de 1990, el "marketing digital".

Para Kotler (2016), que ha investigado y desarrollado ampliamente el concepto y los fundamentos del marketing a lo largo de su carrera, la aparición de internet llevó a clasificar temporalmente el marketing como se explica a continuación.

- Marketing 1.0: basado en el producto y las necesidades básicas del consumidor, se lleva a cabo de forma unidireccional mediante medios tradicionales como radio, prensa o televisión.
- Marketing 2.0: se centra en el consumidor y en establecer con él un vínculo emocional, el mensaje es bidireccional y es aquí donde se incorporan medios interactivos a los tradicionales, como internet, redes sociales, etc.

- Marketing 3.0: en este punto las empresas deben tratar de mejorar la sociedad además de vender, comprender a las personas y las dificultades que atraviesan, agregando ese valor en sus productos o servicios.
- Marketing 4.0: en este penúltimo paso, el aspecto fundamental es la conectividad o interacción. Las empresas deben adaptarse a un entorno cambiante en el que los consumidores son partícipes al estar constantemente conectados. Es aquí donde las empresas deben prestar atención al *customer journey*[1] y donde surge el marketing de permisos, que implica solicitar permiso a los clientes para ofrecerles publicidad a través de la web.
- Marketing 5.0: es el más reciente, implica tener en cuenta los aspectos emocionales de las personas haciendo uso de la tecnología y los datos masivos de los que se disponen.

Figura 1: Evolución del Marketing

http://observatoriodemarketing.blogspot.com/

Actualmente cada vez hay menos empresas en el primer punto de la clasificación, ya que es muy complicado sobrevivir en un mercado sin tener en cuenta las necesidades de las personas y su entorno.

Por tanto, es evidente que el desarrollo web ha forzado, y lo sigue haciendo, a estas empresas a adaptar sus estrategias al entorno online, y dedicar una parte de su presupuesto

[1] Customer Journey: experiencia del cliente a lo largo del proceso de compra desde que se fija en el producto hasta que lo adquiere.

cada vez mayor a la promoción y publicidad a través de estos medios. Es interesante conocer la evolución de internet desde que surgió para poder entender el entorno en el que se realizan las acciones de marketing y las posibilidades que existen.

- Web 1.0: esta es la primera fase, cuando surge la web en la segunda mitad del siglo XX, permitiendo al consumidor leer información subida a los servidores pero sin mayor capacidad de interactuación con ella.
- Web 2.0: en esta fase, que tiene lugar a partir del año 2000, el usuario comienza a interactuar con la web, pudiendo colaborar con otros usuarios y creando información, apareciendo blogs, redes sociales, etc.
- Web 3.0: aparece alrededor de 2006 y permite relacionar las webs de manera semántica gracias a la estructura de la información, lo que se traduce en una mayor eficiencia y agilidad en la búsqueda. Además, ya se puede acceder a la información desde distintos tipos de dispositivos.
- Web 4.0: en esta fase el actor principal es la inteligencia artificial, ya que internet puede funcionar de forma predictiva, no sólo recibiendo órdenes del usuario. Aparece un nuevo vehículo de comunicación con los dispositivos, la voz, permitiendo dar órdenes o solicitar información hablando.

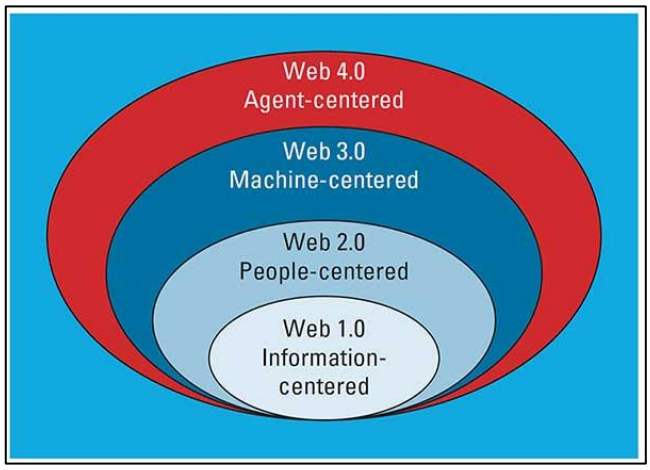

Figura 2: Evolución Desarrollo Web

https://profile.es/blog/web-4-0-el-proximo-desafio-ya-esta-aqui/

Así, el recorrido web lleva consigo una adaptación en todos los ámbitos de la sociedad que afecta, por supuesto, a la forma en que las empresas conectan con sus consumidores y de hacer marketing, que como hemos visto en la clasificación de Kotler, se desarrolla de forma similar al tener que aplicar estos conocimientos web.

Hoy en día, se plantea la web 5.0 como aquella en la que las máquinas tendrán en cuenta las emociones de los individuos para poder ofrecerles experiencias y respuestas exclusivas.

En resumen, esta continua incorporación de los avances web a la forma de hacer marketing de las empresas implica un peso cada vez mayor del marketing digital, en el que se van a focalizar los siguientes puntos.

2.1.1 Marketing digital

Para uno de los autores más reconocidos sobre el tema, David Chaffey (2012) el marketing digital "consiste en la aplicación de internet y tecnología digital relacionada junto al uso de la comunicación tradicional para alcanzar los objetivos de marketing".

En este sentido, el marketing digital presenta diferentes estrategias con las que las empresas o personas hacen llegar información sobre sus productos o servicios al consumidor utilizando la web. Entre las principales se encuentran las siguientes:

- SEO (*Search Engine Optimization*): el término comenzó a usarse en 1997 por John Audette, y se define como el conjunto de acciones orientadas a mejorar el posicionamiento de una web en un buscador, ya sea Google u otro, de forma orgánica, es decir, sin pagar.
- SEM (*Search Engine Marketing*): esta estrategia también busca posicionar una web en los primeros puestos de los buscadores, con la diferencia respecto al SEO de que en este caso no se basa en resultados orgánicos, sino que se paga por colocar la web en mejores puestos. Comúnmente, son los denominados anuncios de Google que aparecen en los primeros lugares al realizar una búsqueda.
- Redes sociales *(Social Media)*: consiste en utilizar las redes sociales y todas las posibilidades que ofrecen para alcanzar a la audiencia.
- Marketing de afiliados: este tipo de marketing está basado en la colaboración. Los afiliados son aquellas personas u otras empresas que promocionan un producto o servicio de otra a cambio de una comisión.
- Email marketing: esta estrategia consiste en el envío de emails a clientes o potenciales clientes de una empresa con distintos objetivos, ya sea fidelizar, convertir nuevos clientes, etc.

- *Display*: es un formato publicitario en el que los anunciantes enseñan sus productos a través de *banners*[2], que se muestran cada vez que el usuario accede a diferentes páginas web.

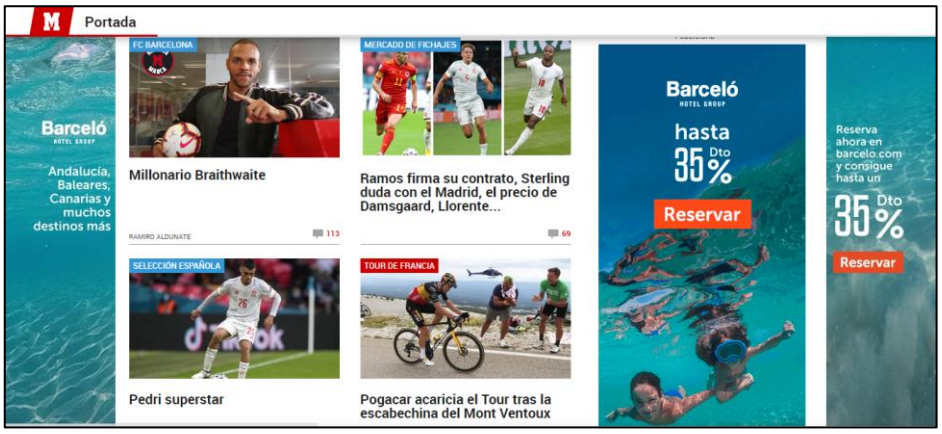

Figura 3: Ejemplo Banner Barceló

https://www.marca.com/

Con esta idea sobre las principales estrategias que se pueden seguir para llevar el contenido al posible cliente, hay que decir que lo que se pretende estudiar aquí es un aspecto fundamentalmente relacionado con la última estrategia comentada, *display*.

Hasta hace poco, este tipo de campañas se realizaban incluyendo esos anuncios en espacios de webs con gran alcance para llegar al mayor público posible. Actualmente ya no es así, ya que la tecnología disponible ha dado lugar a la compra programática (o *real time bidding*).

2.1.2 Compra programática

Esta forma de hacer publicidad consiste en pujar, en tiempo real, por los espacios publicitarios que dejan libres otras webs para introducir un banner anunciando un producto o servicio.

Esto es así gracias al desarrollo web y la hiperconectividad comentada anteriormente, y permite aprovechar los conocimientos sobre las personas almacenados en registros o logs[3]

[2] Banners: anuncio o pieza publicitaria de contenido gráfico que se encuentra localizada en una página web.
[3] Log: registro de todos los acontecimientos que afectan a un usuario en particular.

para ofrecer publicidad específica a aquellos segmentos de la población realmente interesados, pudiendo segmentar audiencias en tiempo real y ajustar el precio.

Para que esto sea posible hay diferentes agentes implicados con las siguientes funciones:

- *Publishers*: se denomina de esta manera a las webs que ofrecen un espacio o soporte donde las empresas que quieren hacer publicidad pueden introducirla, a cambio de un precio. Sería el caso, por poner un ejemplo, de una web de un periódico, que contiene multitud de espacios dedicados a promocionar diferentes webs.
- *Advertisers*: son las empresas que quieren promocionar su producto y pagan por esos espacios a los *publishers*. Siguiendo con el ejemplo anterior, un *advertiser* podría ser una empresa telefónica, que compra el espacio que ofrece la web del periódico para hacer promoción de una de sus tarifas.
- DSP (*Demand Side Platform*): son las plataformas donde los anunciantes o *advertisers* compran inventario[4] de forma automatizada, eficiente y optimizada.

Por todo lo comentado, una de las cuestiones con mayor interés es saber cuándo, dónde y en qué cantidad ofrecer esos anuncios para que impacten a los usuarios pagando el menor precio posible.

2.1.3 Frequency capping

Para tratar de dar respuesta a cuánta publicidad es necesaria enviar a las personas, surge a finales de la primera década del siglo XXI el concepto de *frequency cap* o límite de frecuencia, definido como la "restricción del número de veces que se muestra un anuncio en particular a un usuario específico".

Con él encontramos también el término *frequency capping*, entendido como la "forma de evitar la quema publicitaria (o *banner burnout*[5]), el punto donde los usuarios se ven sobreexpuestos y la respuesta cae, medida sobre todo en campañas de respuesta directa a través del CTR (o tasa de clicks sobre impresiones)".

[4] Inventario publicitario: espacio disponible en los medios de comunicación e industrias de marketing para que los anunciantes inserten su publicidad, en este caso en medios digitales.
[5] *Banner burnout:* quema publicitaria, punto en el que los usuarios han visto tantas veces un mismo anuncio que se cansan de él.

Para entender mejor su funcionamiento, es necesario conocer varios elementos clave para comprender cómo lo usan las empresas actualmente y en qué situación se encuentra.

- Impresiones: cada una de las veces que un usuario se ve expuesto a un determinado contenido publicitario, le preste o no atención.
- Frecuencia: número de veces que se muestra la misma impresión a un mismo usuario en un determinado periodo de tiempo.
- Clicks: cada una de las veces que un usuario pulsa o clica sobre una impresión, siendo redirigido a la web del anunciante.
- Conversiones: son las acciones que una empresa define como exitosas una vez el usuario ha clicado en la impresión publicitaria y accedido a la web del anunciante. Normalmente hace referencia a las ventas, pero también puede tratarse de formularios completados, etc.
- CTR (*Click through Rate*): es una medida de efectividad de la campaña publicitaria. Su forma de obtenerla es la siguiente:
  CTR = número de clicks / número de impresiones
- CR (*Conversion Rate* o Tasa de Conversión): es otra medida de efectividad de una campaña, probablemente la más relevante, y se calcula como sigue:
  CR = número de conversiones / número de impresiones

Para ejemplificarlo claramente, en la Figura 4 aparece el resumen de una campaña publicitaria, con los usuarios que han recibido un número determinado de impresiones (en el primer caso por ejemplo agrupa los usuarios que reciben entre 1 y 5 impresiones) y la tasa de conversión para cada frecuencia.

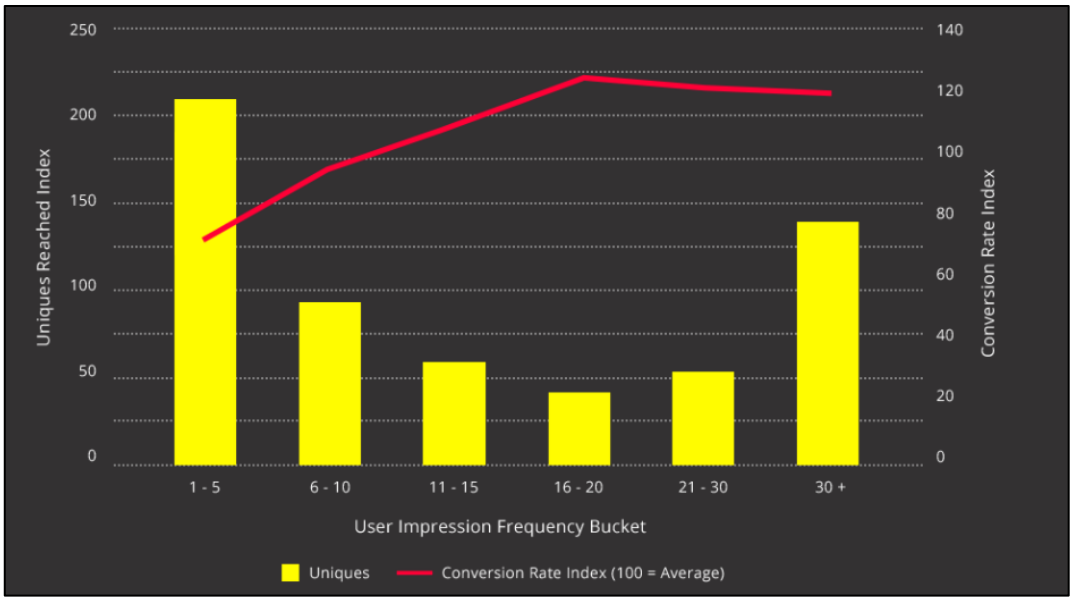

Figura 4: Ejemplo Frequency Capping

https://accordantmedia.com/case-studies/identifying-optimal-frequency-reduces-cpa-25/

Como se puede apreciar, las primeras frecuencias impresiones se envían a un mayor número de usuarios únicos, con una tasa de conversión muy baja en comparación con el resto, debido a que los usuarios aún no han sido suficientemente impactados.

Al aumentar la frecuencia, es decir, el número de repeticiones se ve como la tasa de conversión aumenta, hasta la frecuencia 16-20, donde empieza a caer, pudiendo ser ese el número óptimo de impresiones que se busca encontrar.

Hay que añadir que para tomar una decisión hay que tener en cuenta otros factores como puede ser el precio pagado por las impresiones, etc.

2.1.4 Estudios relacionados

Respecto a los estudios que se han realizado sobre el tema disponibles en la web, se han encontrado aproximaciones desde varios puntos de vista.

La mayoría de ellos tratan el *frequency capping* como un factor muy relevante a la hora de hacer publicidad online, y consideran distintos algoritmos y técnicas para maximizar el beneficio obtenido de sus anuncios.

En algunos casos se hace uso de procesos de Markov, utilizados para modelar el comportamiento humano y social, que consisten en realizar experimentos donde la

probabilidad de obtener cada resultado depende únicamente de los experimentos previos, para tener en cuenta las diferencias entre segmentos de población y maximizar la tasa de conversión (Shanahan y Van der Poel, 2010) con el beneficio que ello conlleva. Por otra parte, existen estudios que proponen resolver el problema mediante algoritmos con programación lineal (Farahat, 2009) u otros tipos de algoritmos que permiten asignar esta limitación de frecuencia para cada sitio web (Feldman et al., 2014)

También existen estudios que proponen modelos de optimización para el *frequency capping* en mercados de puja en tiempo real o *real time bidding*.

2.2 Economía del Comportamiento

El otro eje central del trabajo es la Economía del Comportamiento, que al contrario que para el *frequency capping*, sobre ella existe una extensa literatura, ya que se trata de una rama de la economía que lleva más tiempo en desarrollo.

El concepto de Economía Conductual, popularizado en la década de los 40 en la Universidad de Michigan y cuyos máximos exponentes son Khaneman, Tversky (1979) y Thaler (1987); se puede definir como la rama de la economía que trata de añadir el componente emocional a los modelos racionales, tratando de explicar la conducta humana desde el punto de vista de las preferencias sociales, heurísticas y normas a partir de las cuales se construyen nuevos modelos de conducta. Su objetivo, por tanto, es descubrir cómo funcionan los mecanismos existentes detrás de los juicios y toma de decisiones de los individuos.

Para ello, esta disciplina aglutina conceptos y herramientas económicas y de otras ciencias sociales como la psicología, sociología o las ciencias políticas. Algunos de estos conceptos, como por ejemplo el de "aversión a las pérdidas", ya fueron nombrados por economistas hace más de dos siglos (Adam Smith, 1759).

Desde el punto de vista económico, se ha comprobado en repetidas ocasiones (como ejemplo la crisis de 2008) como el modelo clásico no es consistente con la realidad, ya que los seres humanos no tenemos unas preferencias establecidas y no actuamos de forma plenamente racional y por tanto el comportamiento social no se puede modelar como la suma de comportamientos individuales.

Por este motivo se ha tratado de incorporar los campos comentados anteriormente al económico para tratar de descifrar los procesos de toma de decisiones de los individuos, dando lugar a una serie de propuestas y avances que tienen mucha relación con el estudio y aplicación de un concepto como el *frequency capping*.

2.2.1 Sistemas cognitivos

Entre los avances más importantes en este campo, se encuentra la propuesta de uno de sus principales referentes y ganador del premio Nobel (2002) de Economía Daniel Kahneman, que defiende la existencia de dos sistemas en el proceso cognitivo de toma de decisiones.

Por un lado, un Sistema I impulsivo e intuitivo, que lleva a tomar la gran mayoría de decisiones y no da lugar a la ambigüedad o duda. Ejemplos de procesos de este tipo sería tomar la decisión automática de huir cuando sentimos miedo, elegir qué marca comprar en un supermercado porque es la que escogemos habitualmente, etc.

Por el contrario, existe el Sistema II, aquel que lleva a cabo los procesos de decisión de forma consciente, reflexiva y con mayor análisis. En este caso sí que hay lugar a la duda, porque se pueden barajar distintas posibilidades al mismo tiempo, lo que provoca que este sistema sea más lento.

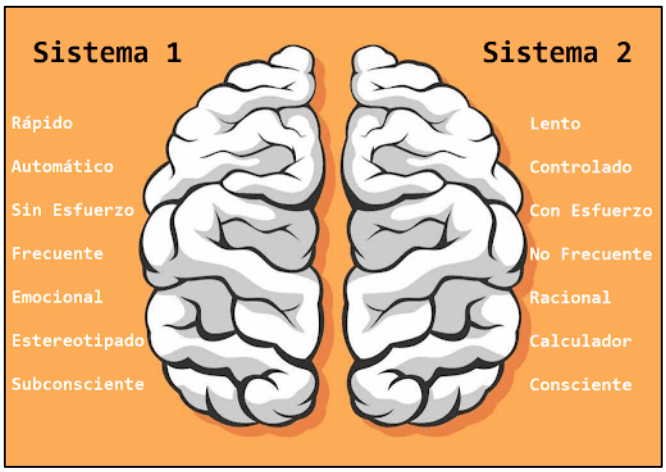

Figura 5: Sistemas Cognitivos, Kahneman

https://josedelgado.net/como-pensamos-sistema-1-y-sistema-2-de-kahneman/

En este trabajo, el proceso de toma de decisiones a estudiar está relacionado con el primer sistema, ya que la respuesta de un usuario es intuitiva, basada en el contexto y la

frecuencia pasada, es decir, el número de veces que se ha estado expuesto al mismo problema, en este caso a las impresiones recibidas.

2.2.2 Sesgos cognitivos

El concepto de sesgo cognitivo (Kahneman y Tverski, 1972), hace alusión a cómo el cerebro analiza en pocos segundos la información para tomar una decisión rápida e intuitiva, estableciendo en muchos casos juicios inexactos o llevando a interpretaciones ilógicas de la realidad.

A lo largo del tiempo se han definido multitud de sesgos que pretenden explicar qué factores influyen en la toma de decisiones.

Para esta investigación, se han definido gran cantidad de ellos con el objetivo de valorar la posible aplicación de alguno en el objeto de estudio, optando al final por incluir los siguientes:

- Efecto Marco (*Frame Effect)*: tiene lugar cuando en un proceso de toma de decisión, los individuos toman decisiones inconsistentes ante problemas idénticos, en función del marco o contexto en el que se presente el problema (Frisch, 1993).

  Se va a tratar de aplicar en esta situación debido a que, a la hora de recibir impresiones publicitarias, el contexto puede ser realmente importante. En caso de tener la información disponible, se podrían analizar diferentes situaciones. Por ejemplo, las páginas webs en las que se reciben las impresiones, el tipo de dispositivo en el que lo hacen, o la versión del mismo.
- Efecto Recencia *(Recency Effect)*: se da cuando las personas tienden mejor a recordar la última información recibida.

  Es posible incluir este efecto en esta investigación analizando las diferencias temporales entre impresiones para ver de qué forma afecta la impresión inmediatamente anterior (la última recibida) a la siguiente, valorando más concretamente cómo se comportan esas impresiones cuando preceden a los clicks.

Por último, hay que añadir que cada vez está alcanzando más importancia y se está tratando de incorporar la Economía del Comportamiento en nuevos campos, indicando algunos ejemplos interesantes en el apartado siguiente.

2.2.3 Estudios relacionados

Existen algunos ejemplos que tratan de aplicar la Economía del Comportamiento al sector de la salud (Carminati, 2020), integrando sus conceptos teóricos para dar explicación al comportamiento tanto de los profesionales de la salud como de los pacientes, estableciendo posibles vías de mejora a nivel individual y social.

Otro caso de aplicación que guarda cierta relación con este trabajo es el estudio de la tasa de clicks en links teniendo en cuenta el Efecto Primacía y el Efecto Recencia (Murphy et al., 2006). En él, se muestra como los visitantes tienen mayor tendencia a clicar en los primeros y últimos links de los sitios web, resaltando su importancia a la hora de diseñar la estructura de navegación web para los usuarios.

Encontramos ejemplos muy diferentes y más actuales de aplicación del Efecto Recencia, como uno relacionado con la electricidad y el *big data* (Wang et al., 2016), en el que se expone que la demanda de electricidad se ve afectada por la temperatura de las horas previas.

Como se ha podido comprobar, el concepto de *frequency capping* es relativamente nuevo y su desarrollo es todavía incipiente. En cuanto a la Economía del Comportamiento, es una rama de la Economía con una extensa teoría y mucha aplicación práctica, fundamentalmente en el entorno bursátil, pero en plena expansión a otros campos como la política, marketing, salud, etc.

Para el tema específico a investigar en este trabajo, se ha encontrado un estudio similar sobre el Efecto Recencia a la hora de hacer click en los links de las páginas web, pero nada que relacione el campo de la Economía Conductual con una herramienta como el *frequency capping*.

Como innovación principal, en esta investigación se va a tratar de poner de manifiesto esa relación utilizando las herramientas estadísticas convenientes.

3 METODOLOGÍA

3.1 Enfoque metodológico

En el presente trabajo se ha tratado de verificar la hipótesis de que los usuarios reaccionan de forma diferente ante la publicidad según las circunstancias y de buscar la frecuencia

óptima de impresiones a enviar a un usuario aplicando varios sesgos de Economía del Comportamiento.

Para realizar este estudio se han tenido en cuenta dos variables, la llamada "tasa de conversión de usuarios", que expresa el número de impresiones que necesita recibir un usuario hasta que hace click en ella, y el CTR (*Click Through Rate*) o tasa de clicks sobre impresiones, que representa la cantidad de clicks que los usuarios han hecho al recibir cierto número de impresiones.

Esto se ha conseguido dividiendo los datos iniciales, que más tarde se describirán, según los respectivos sesgos de Economía Conductual planteados y se han evaluado las diferencias en sus ratios de forma visual, acompañada de los contrastes estadísticos necesarios.

Por último, se ha propuesto un modelo de árbol de clasificación para tratar de obtener, en los casos que sea posible, el número óptimo de impresiones a enviar a los usuarios.

3.2 Herramientas

En esta sección presentamos las herramientas utilizadas para la elaboración de este trabajo, detallando a continuación las principales características de cada una de ellas y por qué se han usado en cada caso.

3.2.1 BigQuery

BigQuery es un almacén de datos para empresas que permite realizar consultas de SQL de alta velocidad mediante el poder de procesamiento de la infraestructura de Google. Da solución al problema que supone almacenar y consultar conjuntos de datos grandes, que pueden consumir una gran cantidad de tiempo y dinero cuando no se cuenta con el hardware y la infraestructura adecuados.

No es necesario el aprovisionamiento de recursos antes de usar *BigQuery*, ya que asigna recursos de almacenamiento y consulta de forma dinámica según los patrones de uso. Esos recursos de almacenamiento se asignan a medida que los consumes, y su asignación se anula a medida que quitas datos o eliminas tablas, mientras que los de consulta se asignan de acuerdo con el tipo de consulta y la complejidad.

No se requiere un compromiso de uso mínimo para utilizarlo, sino que el servicio asigna y cobra por los recursos según el uso real. Un dato a tener en cuenta es que a la hora de realizar consultas, la aplicación refleja su coste.

Al valorar los datos con los que trabajar, se barajaron diferentes conjuntos. Uno de ellos, procedente de la plataforma *Adform*, contenía datos propicios para la investigación propuesta y se trabajó en *BigQuery* debido a que tenía 36 millones de registros de clientes.

Al hacer un análisis más exhaustivo, se desecharon por no poder obtener unos resultados concluyentes y acordes al objetivo deseado, como se detalla en el anexo.

3.2.2 Rstudio

RStudio es una interfaz o entorno de desarrollo integrado (IDE) para el lenguaje de programación R. El lenguaje R está orientado a objetos, y su principal utilidad se encuentra en el campo de la estadística. Además, R es un software libre y colaborativo donde una gran comunidad de desarrolladores proporciona multitud de librerías que añaden funcionalidades útiles a la versión base de R, explicadas detalladamente en una amplia documentación.

Aprovecha el potencial y funcionalidad de R, manteniendo un formato mucho más intuitivo, con cuatro paneles (consola, editor de código, historial de comandos y variables del entorno, y otro con diversas pestañas muy útiles) cuyo objetivo es facilitar y agilizar el trabajo del usuario.

En nuestro caso ha sido la principal herramienta de trabajo y se ha utilizado para desarrollar el código de análisis de los datos, haciendo uso de varias librerías, siendo la principal *dyplr*, para la agrupación y filtrado de datos.

También se ha trabajado con las librerías *rpart, rattle* e *imbalance* a la hora de plantear el modelo de árbol de decisión. La primera de ellas permite modelizar y dibujar árboles de decisión haciendo uso de varios parámetros interesantes directamente en el entorno de Rstudio, que se comentarán más en profundidad en el apartado correspondiente, mientras que la segunda proporciona una interfaz mucho más intuitiva para realizar las transformación y carga de datos, así como la construcción y evaluación de diferentes

modelos. Respecto a la última de ellas, proporciona funciones para construir datos nuevos en el conjunto con el objetivo de balancear una muestra.

3.2.3 Microsoft Excel

Microsoft Excel es una hoja de cálculo diseñada por Microsoft para Windows, macOS, Android e iOS, y forma parte de la suite de software Microsoft Office. Esta aplicación cuenta con cálculo, herramientas gráficas, tablas calculares y un lenguaje de programación macro llamado Visual Basic para aplicaciones.
Es uno de los programas para procesar datos más utilizados por su sencillez, permitiendo a los usuarios elaborar tablas y formatos que incluyan cálculos matemáticos mediante fórmulas; las cuales pueden usar operadores matemáticos muy útiles para agilizar los cálculos.

En el caso de este estudio también ha tenido suma importancia, utilizándose sobre todo en la fase descriptiva de los datos, agilizando la visualización y el cálculo de las medidas más orientativas sobre los mismos mediante la utilización de una de sus funcionalidades más potentes, las tablas dinámicas.

Estas tablas están basadas en dos conceptos: resumir y ordenar. Se crean a partir otras tablas y permiten analizar gran cantidad de columnas mostrando únicamente la información relevante. En este caso se han creado a partir de la tabla de datos original importada en formato csv[6].

3.3 Material

Para la realización de un trabajo como este, es básico tener unos datos que permitan realizar este tipo de estudio, lo que significa que deben tener mínimo datos de una o varias campañas publicitarias sobre los usuarios a los que se le han enviado impresiones, así como el número de impresiones y clicks de cada usuario.

Es necesario también que los datos contengan variables conductuales para poder aplicar en el análisis los sesgos comentados.

[6] Csv (*Comma-separated values*): tipo de documento de formato abierto para representar datos en forma de tabla, en el que las columnas se separan por comas y las filas por saltos de línea.

Lo idóneo sería que los datos registraran información sobre las ventas conseguidas después de esos clics, ya que es la principal medida de efectividad de una campaña, pero no ha sido posible por la dificultad de acceder a unos datos libres que registraran todas las variables deseadas. Esto ocurre porque la protección oficial de datos y el uso privado que se debe hacer de ellos no permite acceder, en el contexto de investigación para un trabajo de este tipo, a bases de datos que dispongan de gran variedad de variables, por lo que hay que adaptarse a los datos encontrados en la web.

Se explica en detalle en el apartado siguiente, pero el conjunto de datos utilizado para trabajar tiene las variables impresiones y clicks, lo que ha permitido el cálculo del CTR como medida de conversión o efectividad de la campaña publicitaria.

4 DESARROLLO DEL ANÁLISIS

Una vez definido y contextualizado el problema que se va a tratar a lo largo del trabajo, así como los objetivos que se pretenden alcanzar, se va a desarrollar paso a paso el procedimiento seguido.

Se llevará a cabo en primer lugar un análisis descriptivo de los datos con los que se trabaja para después entrar en el cálculo del *frequency capping*.

4.1 Base de datos

En primer lugar, se debe explicar el proceso de selección de datos, en los que necesariamente tenían que aparecer variables fundamentales como impresiones y clicks, así como variables de comportamiento que permitieran enfocar de esa forma el análisis, como el tipo de dispositivo móvil, la versión, etc.

Con estas premisas se obtuvieron unos datos de la web de *Kaggle*[7] sobre un sitio web de comercio electrónico que vende varios productos en su plataforma online. La empresa registra el comportamiento de los clientes y almacena sus acciones en registros individuales. En muchos casos esas acciones no corresponden con una compra, ya que

[7] Kaggle: plataforma de *Data Science* que permite a los usuarios encontrar y publicar conjuntos de datos para trabajar con ellos.

los usuarios navegan por la web visitando diferentes productos para luego compararlos con la competencia y decidir si comprar alguno o cuál de ellos adquirir.

Para mejorar las ventas, el propietario de la web ha contratado una empresa de publicidad online que, siguiendo esa cookie que deja el usuario al entrar en la web para ver los productos, le asigna unos anuncios en otras webs de sitios asociados con la empresa.

Con esto el propietario consigue impactar al usuario con los mismos o similares productos a los que ha visitado en su web, buscando que haga click en dichos anuncios, sea redirigido a la web y compre esos productos que ya visitó.

El conjunto de datos que se ha obtenido presenta información sobre las impresiones que se han enviado a los usuarios, el usuario que las recibe, en qué momento lo hace, desde qué dispositivo y si clica en ella. En la Tabla 1 se pueden ver las diferentes variables, que se detallan a continuación.

4.1.1 Análisis descriptivo

Antes de comenzar con el tratamiento de los datos, se explica en profundidad el tipo de variables que contienen y sus aspectos más importantes.

impression_id	impression_time	user_id	app_code	os_version	is_4G	is_click
5f98c2c31a8006e510448c02ec74d50f	26/11/2018 23:30	0	207	old	0	0
010ed37e44e2fdc175b4c5c6c930805a	20/11/2018 20:53	2	190	intermediate	0	0
2b12c0d47f5821a5adb3bfd973d0f708	19/11/2018 20:49	2	190	intermediate	0	0
f85e4bf0e34f3ee66add56229845e4db	20/11/2018 20:29	2	190	intermediate	0	0
577c63f9937fa0e8d4650ddf1510a03f	21/11/2018 21:47	2	190	intermediate	0	0
75f1a891e432ee31eeeb7ae145c665f1	05/12/2018 21:48	3	371	latest	1	0
de0f7e146d594e3868c1f4aae5600c2f	04/12/2018 11:56	5	127	latest	1	0
fb4cd2d75ba5915913d52e43c780c3d2	04/12/2018 18:39	5	127	latest	1	0
f178a1b8e28622bddf49657470a15cd3	30/11/2018 23:35	6	249	latest	0	0
b69052087938225e152c4942e5e81b1b	06/12/2018 14:16	8	44	intermediate	1	0
76e63faa6d1b0b44f18000e0e743ee84	03/12/2018 13:02	9	318	latest	1	0
b9071cc2d88a6bf518a6be97a88f8260	05/12/2018 8:47	9	318	latest	1	0
3812a55dce945bdb6ef283ab66ef5cf5	03/12/2018 13:13	9	318	latest	1	0
b76277cdde174c7e14d52230177b102f	05/12/2018 9:59	9	3	latest	1	0
2bd235c31c97855b7ef2dc8b414779af	15/11/2018 8:51	10	3	latest	1	0
daad8d509446c856e52d79f897232876	15/11/2018 11:24	10	320	latest	1	0
bc69c4b0159c213c43f783c7fde8f870	27/11/2018 19:02	11	386	latest	1	0
b4572f47b7c69e27b8e46646d9579e67	16/11/2018 4:06	12	465	intermediate	0	0
c4c69a272ac9ac42851a071c5ccab8fb	06/12/2018 19:25	14	386	latest	0	0
9414ea7ad8d65d96165adf1abf9cfe51	01/12/2018 20:13	15	386	intermediate	1	0
6f73a8f73c960fec79f878c25bc742cf	11/12/2018 23:11	16	386	latest	1	0

Tabla 1: Base de Datos de Kaggle

El conjunto de datos está constituido por un total de 237609 filas o impresiones y 7 columnas, siendo el periodo temporal de la campaña desde el 15/11/2018 al 13/12/2018.

Sus variables son fundamentalmente categóricas, a excepción de las que identifican las impresiones y los usuarios (Impression_id y user_id), y la que indica el momento en que recibe la impresión (Impression_time). Profundizando en ellas:

- Impression_id: índice de referencia de cada impresión enviada a un usuario. Al ser distinto para cada impresión, hay tantos como filas tiene el *dataset*. El total de impresiones es de 237609.
- Impression_time: fecha exacta en la que aparece la impresión en el dispositivo del usuario.

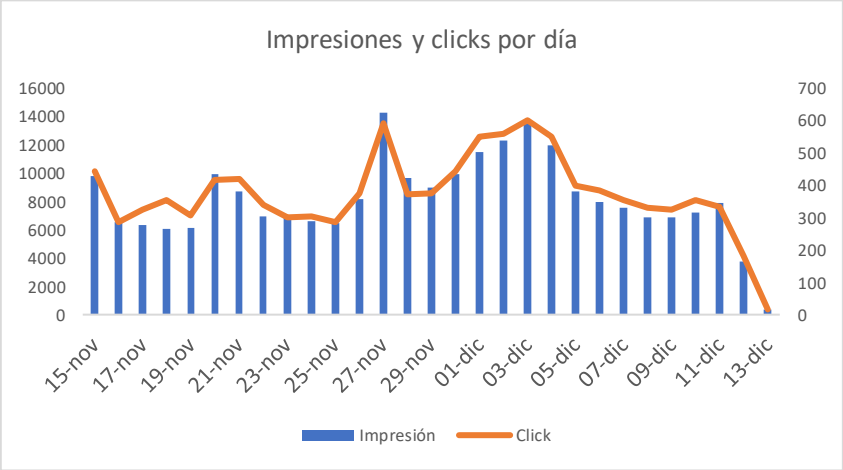

Figura 6: Distribución Temporal Impresiones y Clicks (Datos Generales)

Como se ve en la Figura 6, los clicks varían de forma muy similar a las impresiones, con el máximo el 27 de noviembre de 14243 impresiones y el de clicks el 3 de diciembre con 600.

- User_id: número de identificación de cada usuario. Hay un total de 74723 usuarios.
- App_code: número de identificación de los sitios o aplicaciones en las que aparece cada impresión.
- Os_version: versión del dispositivo que recibe la impresión.

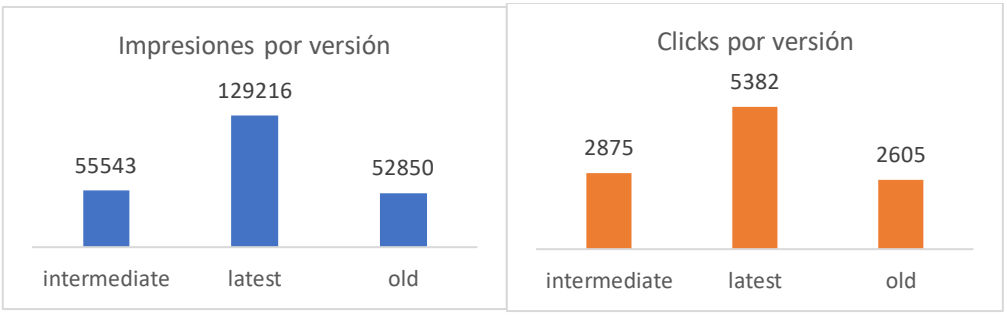

Figura 7: Impresiones y Clicks por Versión (Datos Generales)

- Is_4G: tecnología del dispositivo que recibe la impresión.

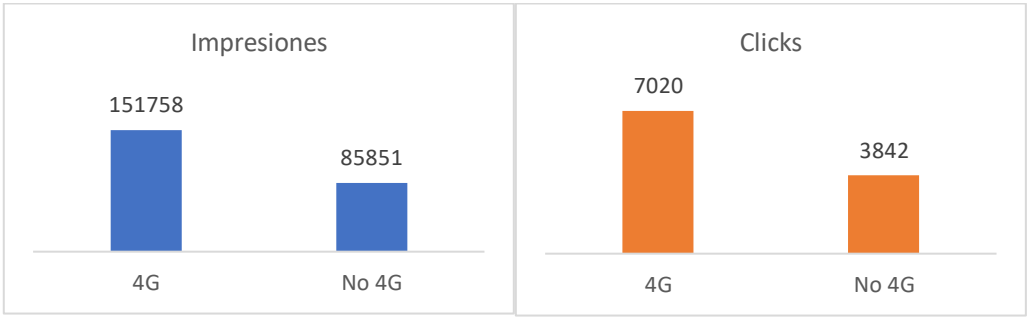

Figura 8: Impresiones y Clicks según Tecnología (Datos Generales)

- Is_click: identifica si el usuario hace click en la publicidad o no. En el periodo delimitado ha habido un total de 10862 clicks.

4.1.2 Transformación de los datos

Tal como se ha planteado al comienzo de este apartado, se busca obtener el número adecuado de impresiones para que un usuario haga un primer click, por lo que se ha eliminado del conjunto de datos todas las impresiones de cada usuario posteriores a su primer click, en caso de que lo haya hecho.

De esta forma, el *dataset* resultante presenta las mismas variables que el inicial, con la diferencia de que el número de filas se reduce considerablemente.

Para ejemplificar esta modificación se pueden ver las siguientes tablas, que reflejan los datos de un usuario concreto (usuario 98) antes y después, eliminándose las dos impresiones posteriores a su primer click.

impression_id	impression_time	user_id	app_code	os_version	is_4G	is_click
c53ac11d42c5bd5051bb953b99e01d75	20/11/2018 6:54	98	3	old	0	0
078fe7c7d854680bf010786b0806a1be	30/11/2018 6:22	98	3	old	0	0
a58af0b4534762bdc1a82ca4ad4b8148	27/11/2018 21:37	98	3	old	0	1
0645b09c4c741958c1ab33eed0adc55b	28/11/2018 10:31	98	3	old	0	0
2658c91ecbca5651a53b4b13709c62fa	29/11/2018 17:07	98	3	old	0	0

impression_id	impression_time	user_id	app_code	os_version	is_4G	is_click
c53ac11d42c5bd5051bb953b99e01d75	20/11/2018 6:54	98	3	old	0	0
078fe7c7d854680bf010786b0806a1be	30/11/2018 6:22	98	3	old	0	0
a58af0b4534762bdc1a82ca4ad4b8148	27/11/2018 21:37	98	3	old	0	1

Tabla 2: Transformación Datos a Primer Click

Una vez realizada la modificación para todos los usuarios, hay varias cosas a destacar de los nuevos datos:

- El número de impresiones a tener en cuenta se reduce a 206540, con una media diaria de 7122 impresiones enviadas y una distribución de frecuencias en la que se aprecia mayor acumulación los últimos días de noviembre y primeros de diciembre.
- Respecto a los clicks, hay un total de 7656, lo que supone un 3,7% del total de impresiones enviadas, con una media de 264 diarios. En la siguiente gráfica queda patente como las variaciones en las impresiones van acompañadas por variaciones similares en clicks.

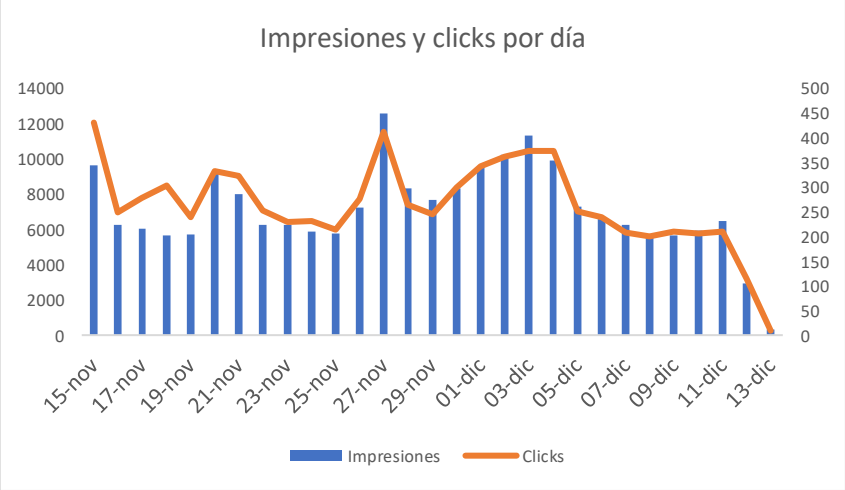

Figura 9: Distribución Temporal Impresiones y Clicks (Datos Transformados)

- El total de usuarios alcanzados no varía, siendo por tanto 74723.

- La distribución de impresiones según la versión y la tecnología pone de manifiesto que el uso de dispositivos nuevos es mayoritario, así como el hecho de no tener 4G.

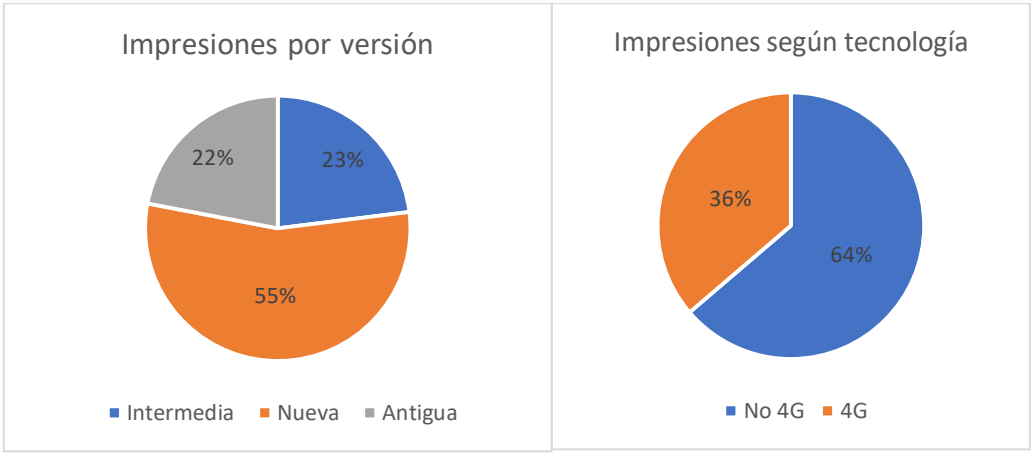

Figura 10: Impresiones por Versión y Tecnología

Entrando más en detalle, se comprueba como en los dispositivos con 4G la proporción de la versión nueva es aún más grande que en el caso global o en el de los dispositivos sin 4G. Cabe esperar que los dispositivos con mejor tecnología sean aquellos con la última versión incorporada.

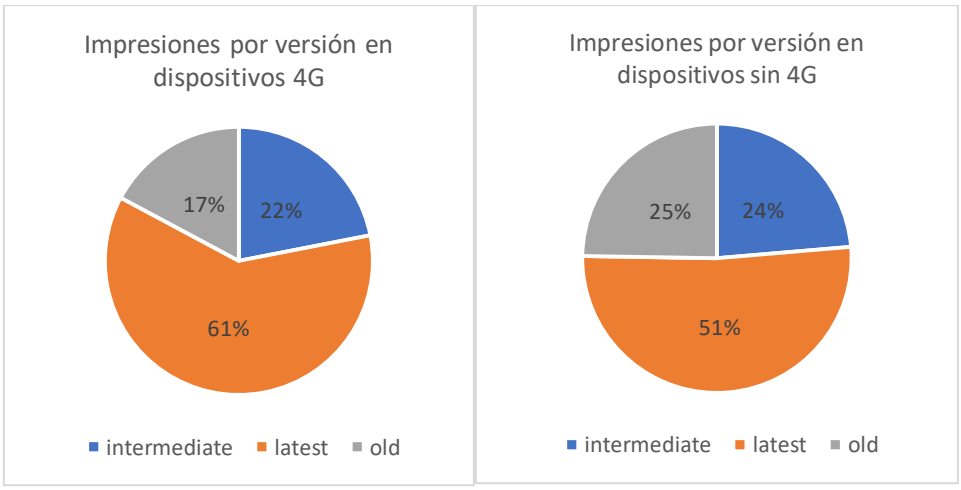

Figura 11: Impresiones por Versión según Tecnología

La información obtenida de los datos refleja en todo caso un comportamiento similar de impresiones y clicks a lo largo del tiempo en el que se tienen registros, cercano a un mes.

Asimismo, algo más de la mitad de las impresiones se han recibido en dispositivos de última versión, mientras que una amplia mayoría lo han hecho en dispositivo sin tecnología 4G.

4.2 Primeros cálculos

Después de analizar y hacer una breve descripción de los datos, se puede comenzar a desarrollar sobre ellos los cálculos de las medidas de conversión que ayuden a encontrar el número óptimo de impresiones. La idea y el objetivo que persigue está claro, pero se puede plantear desde diferentes perspectivas a la hora de buscar ese número óptimo de impresiones para que un usuario haga click.

4.2.1 Análisis por impresiones

La primera de ellas es la orientación habitual. Se trata de acumular las impresiones según el orden (o frecuencia) en el que las reciben los usuarios, es decir, agrupar todas las primeras impresiones, todas las segundas, etc., y observar el comportamiento del CTR en cada frecuencia.

Esta variante es la que se utiliza generalmente para calcular los precios que deben pagar las empresas por los inventarios publicitarios que desean contratar. Esto es así porque permite identificar fácilmente el coste por click.

De esta forma se obtiene la siguiente tabla, cuya pretensión es reflejar como actúan los usuarios ante el envío de cada frecuencia de impresiones.

Frecuencia	Impresiones	Impresiones acumuladas	Impresiones acumuladas %	Clicks	Clicks acumulados	Clicks acumulados %	CTR	Conversión acumulada
1	74723	74723	36,18%	3231	3231	42,20%	4,32%	4,32%
2	38822	113545	54,98%	1444	4675	61,06%	3,72%	4,12%
3	21352	134897	65,31%	832	5507	71,93%	3,90%	4,08%
4	14702	149599	72,43%	519	6026	78,71%	3,53%	4,03%
5	10215	159814	77,38%	352	6378	83,31%	3,45%	3,99%
6	7824	167638	81,17%	246	6624	86,52%	3,14%	3,95%
7	6041	173679	84,09%	167	6791	88,70%	2,76%	3,91%
8	4944	178623	86,48%	156	6947	90,74%	3,16%	3,89%
9	3957	182580	88,40%	117	7064	92,27%	2,96%	3,87%
10	3298	185878	90,00%	90	7154	93,44%	2,73%	3,85%
11	2751	188629	91,33%	90	7244	94,62%	3,27%	3,84%
12	2333	190962	92,46%	53	7297	95,31%	2,27%	3,82%
13	1982	192944	93,42%	52	7349	95,99%	2,62%	3,81%
14	1706	194650	94,24%	35	7384	96,45%	2,05%	3,79%
15	1473	196123	94,96%	37	7421	96,93%	2,51%	3,78%
16	1275	197398	95,57%	42	7463	97,48%	3,29%	3,78%
17	1076	198474	96,10%	33	7496	97,91%	3,07%	3,78%
18	923	199397	96,54%	21	7517	98,18%	2,28%	3,77%
19	802	200199	96,93%	16	7533	98,39%	2,00%	3,76%
20	711	200910	97,27%	17	7550	98,62%	2,39%	3,76%

Tabla 3: Resultados Análisis por Impresiones

Por ejemplo, la frecuencia 1 hace referencia a la primera impresión que reciben todos los usuarios, sin excepción. Por ello coincide el número de impresiones con el número total de usuarios comentado en el punto 4.1. En el caso de la frecuencia 2, recoge la segunda

impresión enviada a todos los usuarios, exceptuando aquellos que únicamente reciben una impresión.

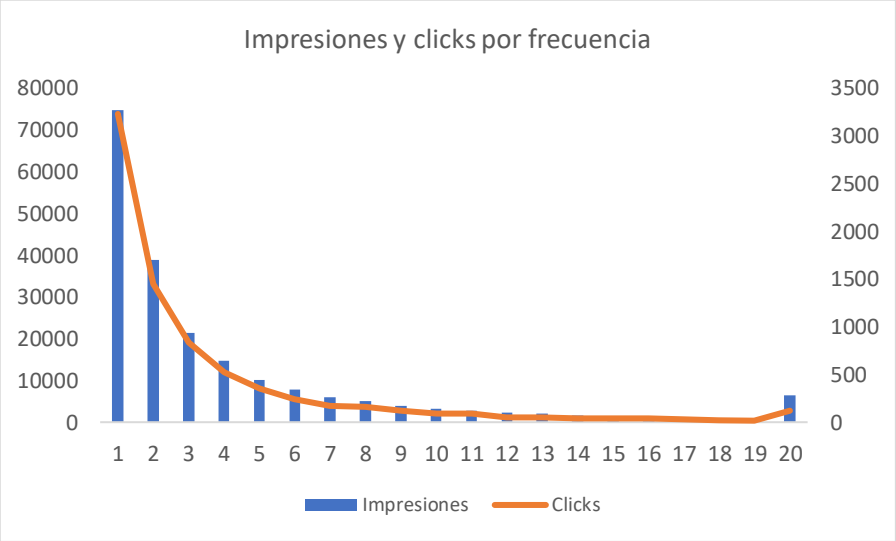

Figura 12: Distribución Impresiones y Clicks por Frecuencia

En la Figura 12 se puede ver el número total de impresiones y clicks por frecuencia, acumulando en la 20 todos los datos posteriores, que representan un porcentaje muy pequeño del total (2,73% de impresiones y 1,38% de clicks).

Lo más importante a tener en cuenta en este punto es la curva del CTR que aparece en el gráfico de línea (Figura 13), en el que se aprecia una tendencia ligeramente descendente, es decir, conforme aumenta el número de impresiones que recibe un usuario se reduce la tasa de conversión, por lo que según esto no sería conveniente enviar un número elevado de ellas.

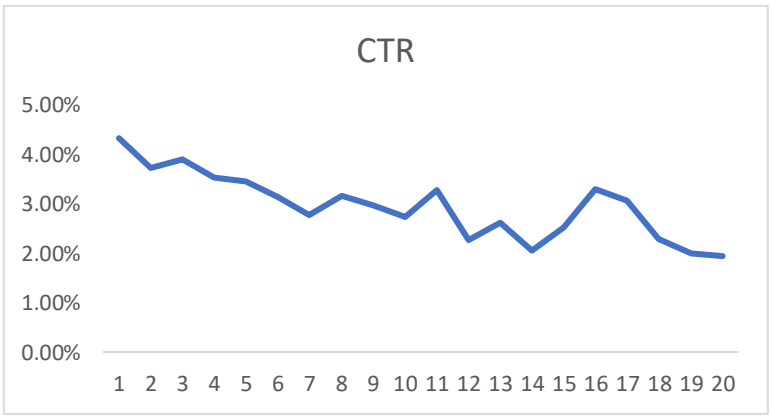

Figura 13: CTR por Impresiones Datos Globales

Es destacable el repunte de la tasa de clicks para las frecuencias entre 8 y 11 impresiones, ya que son puntos en los que hay un porcentaje muy alto de impresiones y clicks acumulados y podría identificar la zona en la que dejar de mandar impresiones.

4.2.2 Análisis por usuarios

Como se ha comentado anteriormente, no existe una sola forma de valorar el comportamiento de los usuarios frente a la publicidad recibida. En este segundo punto se plantea un acercamiento desde el punto de vista del usuario que recibe un número específico de impresiones.

Es decir, se acumulan todos los usuarios que reciben la misma frecuencia de impresiones y se comprueba cuál ha sido para ellos la tasa de conversión de usuarios y el CTR.

Para ello, primero se construye una tabla intermedia (Tabla 4) que agrupe cada usuario con su número de impresiones y si ha hecho click o no, así como la versión de dispositivo utilizada y la tecnología.

A continuación, se muestra un extracto de la misma, en la que, por ejemplo, el usuario con id 615 ha recibido un total de 5 impresiones hasta hacer click, mientras otro usuario, el id 630, ha recibido un total de 4 impresiones, pero no ha hecho click.

user_id	Impresiones	Click	version	is_4G
610	4	0	intermediate	0
611	2	0	old	0
612	1	0	latest	0
613	2	0	latest	0
614	3	0	intermediate	0
615	5	1	latest	1
618	1	0	latest	0
619	2	0	latest	0
620	1	0	old	0
621	1	0	latest	1
622	1	0	intermediate	1
623	1	0	old	0
626	1	0	intermediate	0
627	1	0	latest	0
628	1	0	latest	0
629	1	0	latest	1
630	4	0	latest	0
631	5	1	latest	1
632	1	0	latest	0

Tabla 4: Datos Agrupados por Usuario

La construcción de esta tabla de transición permite agrupar aquellos usuarios que recibieron un mismo número de impresiones y realizar un conteo de los que hicieron click tras recibir esas impresiones. Esta información viene detallada en la Tabla 5, en la que aparecen las mismas variables que en el punto 4.1.1., además de la tasa de conversión de usuarios, que como veremos al interpretar estos datos es realmente importante.

Frecuencia	Usuarios	Impresiones	Clicks	Impresiones acumuladas	Clicks acumulados	Imp acum %	Clicks acum %	Tasa conv acum	Conversión usuarios	CTR
1	35901	35901	3228	35901	3228	17,38%	42,16%	9,00%	9,00%	9,00%
2	17470	34940	1446	70841	4674	34,30%	61,05%	6,60%	8,30%	4,10%
3	6649	19947	832	90788	5506	43,96%	71,92%	6,06%	12,50%	4,20%
4	4489	17956	521	108744	6027	52,65%	78,72%	5,54%	11,60%	2,90%
5	2390	11950	351	120694	6378	58,44%	83,31%	5,28%	14,70%	2,90%
6	1783	10698	245	131392	6623	63,62%	86,51%	5,04%	13,70%	2,30%
7	1097	7679	168	139071	6791	67,33%	88,70%	4,88%	15,30%	2,20%
8	987	7896	156	146967	6947	71,16%	90,74%	4,73%	15,80%	2,00%
9	659	5931	117	152898	7064	74,03%	92,27%	4,62%	17,80%	2,00%
10	547	5470	90	158368	7154	76,68%	93,44%	4,52%	16,50%	1,60%
11	418	4598	90	162966	7244	78,90%	94,62%	4,45%	21,50%	2,00%
12	351	4212	53	167178	7297	80,94%	95,31%	4,36%	15,10%	1,30%
13	276	3588	52	170766	7349	82,68%	95,99%	4,30%	18,80%	1,40%
14	233	3262	35	174028	7384	84,26%	96,45%	4,24%	15,00%	1,10%
15	198	2970	37	176998	7421	85,70%	96,93%	4,19%	18,70%	1,20%
16	199	3184	42	180182	7463	87,24%	97,48%	4,14%	21,10%	1,30%
17	153	2601	33	182783	7496	88,50%	97,91%	4,10%	21,60%	1,30%
18	121	2178	21	184961	7517	89,55%	98,18%	4,06%	17,40%	1,00%
19	91	1729	16	186690	7533	90,39%	98,39%	4,04%	17,60%	0,90%
20	78	1560	17	188250	7550	91,14%	98,62%	4,01%	21,80%	1,10%

Tabla 5: Resultados Análisis por Usuarios

Es necesario antes de nada aclarar cómo se han calculado algunas variables y su interpretación.

- Usuarios: agrupa todos los usuarios que han recibido cada frecuencia de impresiones.
- Impresiones: se obtiene multiplicando cada frecuencia por el número de usuarios, y contabiliza todas las impresiones recibidas por cada grupo de usuarios. Por ejemplo, los usuarios que han recibido exactamente 2 impresiones (frecuencia=2) son 17470, y han recibido un total de 34940 impresiones.
- Clicks: número de clicks totales para cada grupo de usuarios, es decir, si la frecuencia es 2 como en el ejemplo anterior, hay 1446 usuarios que han clicado al recibir la segunda impresión.
- Conversión usuarios: se calcula dividiendo los clicks entre los usuarios correspondientes. Se interpreta como el número de usuarios que han reaccionado al recibir cierto número de impresiones.

  En el caso de la frecuencia igual a 2, el 8.3% de los usuarios que recibieron 2 impresiones hicieron click al recibir esa segunda impresión

De estos datos se puede extraer la siguiente información relevante, presentada en los siguientes dos gráficos (Figura 14).

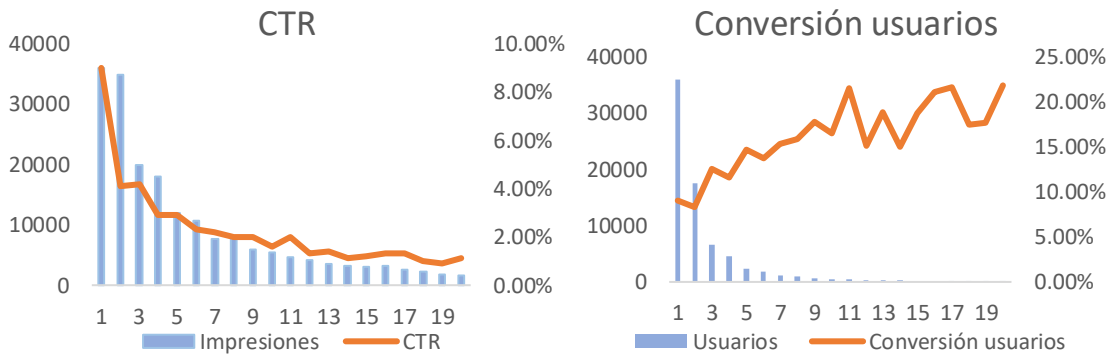

Figura 14: Ratios de Conversión por Usuarios

En ellos se puede ver que los usuarios están muy concentrados en las primeras frecuencias, lo que significa que la mayoría de ellos recibió pocas impresiones, un 90% recibió 5 o menos.

También se comprueba como los que sólo recibieron una tienen un CTR mucho más elevado que el resto, indicando que a priori y sin tener en cuenta otros factores, podría ser más efectivo enviar una sola impresión a un mayor número de usuarios.

Otra conclusión relevante que se debe mencionar es que la tasa de usuarios que clican (Figura 14) sigue una tendencia ascendente, lo que significa que las impresiones repetidas están teniendo efecto, ya que, ante un aumento de impresiones, se produce un aumento de usuarios que hacen click.

Cabe destacar además que, al igual que en el apartado anterior, hay cierto repunte en aquellos usuarios que reciben 11 impresiones.

4.3 Aplicación de la Economía Conductual

Todo lo comentado hasta el momento en este punto del trabajo permite hacerse una idea de las características globales de la campaña de marketing de este *ecommerce* y los aspectos más importantes en relación al tema que se está tratando, el *frequency capping*.

Lo planteado hasta ahora es el análisis básico y más frecuente que se realiza para determinar la frecuencia óptima de exposición a la que someter a los usuarios. Como se ha mencionado anteriormente, la novedad que se pretende incorporar a este trabajo son

los diferentes sesgos de Economía del Comportamiento que puedan tener cabida en base a las variables disponibles en el conjunto de datos, para comprobar si la probabilidad de hacer click de un usuario expuesto a diferente frecuencia de impresiones publicitarias varía según las circunstancias, y cómo lo hace.

Se ha seguido el mismo procedimiento para cada variable o tipo de análisis. En primer lugar, plantear una hipótesis de partida sin base científica, a modo de experimento.

Después dividir los datos y calcular los ratios correspondientes de conversión por frecuencias (conversión de usuarios y CTR), para más tarde comprobar mediante un test estadístico si esas diferencias resultantes son realmente significativas.

Antes de entrar en el propio análisis, se van a explicar los diferentes contrastes estadísticos utilizados para apoyar de esta forma los resultados obtenidos con el primer análisis visual.

Se ha hecho uso de tres contrastes diferentes:

- Test de normalidad (test de Shapiro-Wilks), necesario para contrastar si las variables son normales o no, y saber qué tratamientos darles posteriormente. Las hipótesis que plantea son:
  H0: $X \sim N(\mu,\sigma^2)$
  H1: $X \nsim N(\mu,\sigma^2)$
- Test de comparación de medias (prueba t de Student), es un test paramétrico empleado en caso de que las variables presenten normalidad, plantea las siguientes hipótesis:
  H0: $\mu_x = \mu_y$
  H1: $\mu_x \neq \mu_y$
- Test de comparación de medianas (test de Wilcoxon), es un test no paramétrico utilizado cuando alguna de las variables a comparar no presenta normalidad y por tanto no se puede usar la prueba anterior. Sus hipótesis son:
  H0: $P(X>Y)=0.5$
  H1: $P(X>Y)\neq 0.5$

Como ejemplo se muestra la salida de un contraste de hipótesis en RStudio (Figura 15), en este caso un contraste de normalidad. El valor a tener en cuenta en los contrastes que se verán es el p-valor, rechazando la hipótesis nula cuando éste sea inferior a un nivel de significación $\alpha = 0.1$, como en este caso.

```
> shapiro.test(CTR)

        Shapiro-Wilk normality test

data:  CTR
W = 0.72454, p-value = 3.624e-06
```

Figura 15: Ejemplo Contraste de Normalidad en RStudio

4.3.1 Efecto Marco

Este efecto, definido en la Economía del Comportamiento como el sesgo cognitivo por el que las preferencias de una persona ante un problema decisional varían según el contexto en el que se les presente, se estudia aquí por varias razones.

La primera de ellas es que el problema ante el que se encuentra el usuario es decisional, como ya se ha explicado previamente, y la segunda es que los datos recabados permiten el estudio de variables referentes al contexto o marco en el que el usuario debe decidir si hacer click, pudiendo afectar a la probabilidad de hacerlo.

Las dos variables de las que se dispone para el estudio son, por un lado, la versión de dispositivo en el que el usuario recibe la publicidad, y por otro, la tecnología del dispositivo en el que lo hace.

Para cada uno de estos análisis se han seguido los mismos pasos. El primero de ellos ha sido dividir los datos en subconjuntos según la categoría de la variable, para más tarde calcular los mismos ratios obtenidos para el conjunto de datos global, y ver así las diferencias existentes (si es que las hay) entre cada uno de ellos.

Teniendo todo esto presente, se puede proceder al estudio aplicando las variables conductuales.

4.3.1.1 Versión de dispositivo

Antes de entrar en detalle, se debe plantear la siguiente hipótesis de partida, que se tratará de comprobar con el cálculo de los diferentes ratios:

- Desde el punto de vista personal, los usuarios se sienten más cómodos utilizando la versión más antigua de móvil, que conocen mejor, por lo que se puede suponer un CTR más alto para esa versión.

- Por otro lado, desde el punto de vista de la empresa, sería factible suponer que las cookies hayan captado más información en versiones más antiguas y por tanto estas puedan enviar publicidad más cualificada al usuario, generando un mayor CTR.

Tal como se vio en el apartado descriptivo 4.1.1, la variable denominada "os_version" tiene es una variable de tipo categórico o factor, con tres categorías diferentes; antigua (*old*), intermedia (*intermedium*) y nueva (*new*).

La distribución de impresiones mostraba que la mitad de ellas eran recibidas en dispositivos con la versión nueva, mientras aproximadamente un 25% lo hacían en la intermedia y antigua, respectivamente.

Para el análisis y posterior comparación, se ha dividido el *dataset* en tres subconjuntos, uno por cada versión. En la Tabla 6 aparece un extracto del subconjunto de datos que hacen referencia a la versión antigua para ejemplificarlo. De igual manera se tienen los dos restantes.

impression_time	user_id	app_code	os_version	is_4G	is_click
26/11/2018 23:30	0	207	old	0	0
04/12/2018 12:43	31	163	old	0	0
21/11/2018 20:44	35	249	old	0	0
03/12/2018 23:06	42	386	old	0	0
15/11/2018 23:36	47	207	old	0	0
12/12/2018 1:05	47	207	old	0	0
08/12/2018 22:50	52	44	old	0	0
05/12/2018 19:29	56	386	old	0	0
15/11/2018 5:21	69	207	old	0	0
20/11/2018 2:35	69	207	old	0	0
20/11/2018 9:12	69	207	old	1	0
22/11/2018 11:00	69	207	old	0	0
25/11/2018 13:06	69	207	old	0	0
25/11/2018 14:27	69	207	old	0	0
28/11/2018 11:46	69	207	old	0	0
30/11/2018 20:36	69	207	old	0	0

Tabla 6: Ejemplo Datos Versión Antigua

En este caso ya no es necesario definir cada variable porque no varían respecto a los datos iniciales, pero sí describir brevemente cada subconjunto.

Versión	Impresiones	Clicks	Usuarios distintos
Antigua	45435	1843	16518
Intermedia	47564	1938	17128
Nueva	113541	3875	41195
Total	206540	7656	74841

Tabla 7: Resumen Datos por Versiones

Tanto el número total de impresiones como el de clicks coinciden con el total general, mientras que el total de usuarios distintos es algo mayor porque hay algunos que han recibido impresiones en más de un tipo de dispositivo.

La distribución temporal de impresiones y clicks es muy similar, como se aprecia en las siguientes gráficas, teniendo en todos los casos un pico en el CTR a finales de diciembre y cayendo hasta el día 13, el último en el que se recogieron datos.

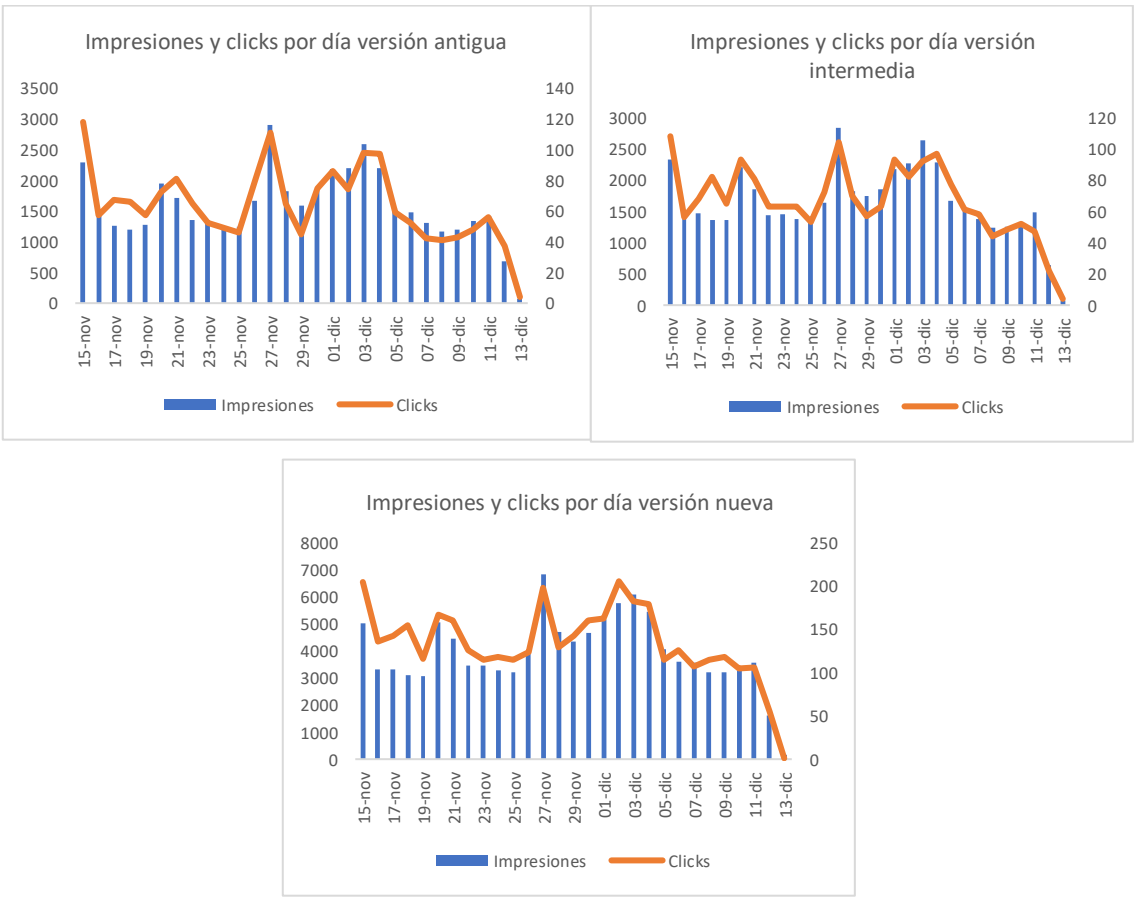

Figura 16: Distribución Impresiones y Clicks por Versiones

En cuanto a la distribución de las impresiones y clicks según la frecuencia, el estudio realizado es el mismo que en el caso general. En primer lugar, se presentan los resultados por grupos de impresiones y después por grupos de usuarios.

Para el primer caso, los CTR calculados han sido los siguientes.

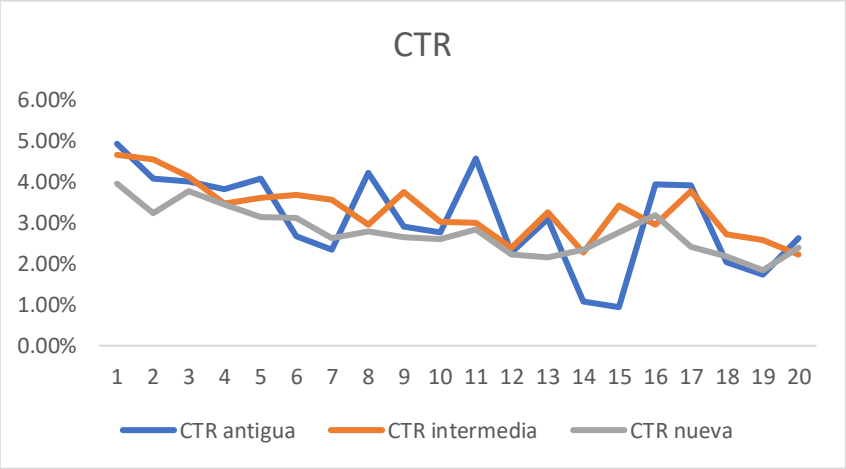

Figura 17: CTR por Impresiones según Versiones

Se puede apreciar en la Figura 17 como el CTR en las versiones antigua e intermedia está generalmente por encima del de la versión nueva. En todos los casos la primera impresión es de gran importancia al ser la que tiene un mayor ratio de clicks.

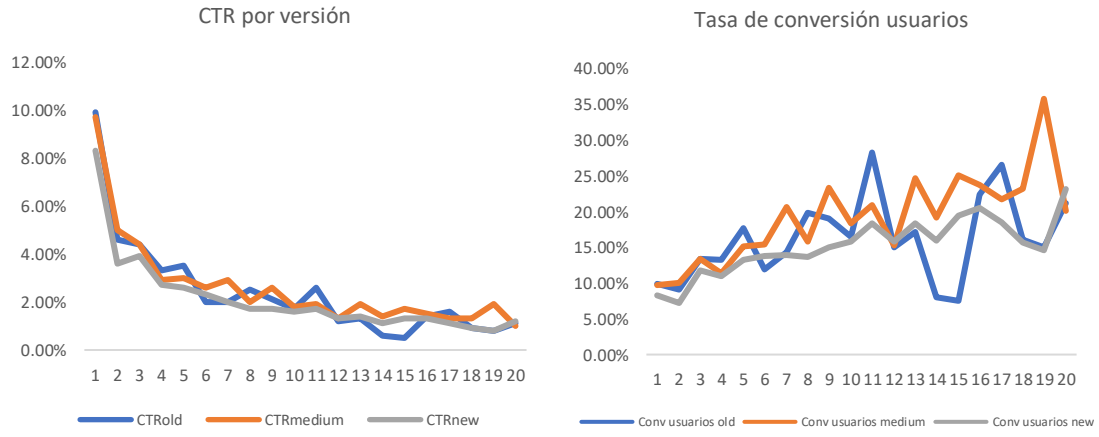

Figura 18: Ratios de Conversión por Usuarios según Versiones

Por otra parte, la información analizando la frecuencia de impresiones por grupos de usuarios permite extraer unas conclusiones similares a las generales. Aquellos usuarios que reciben sólo una impresión tienen un CTR notablemente más alto que el resto, mientras que la tasa de conversión de usuarios presenta la misma tendencia creciente hasta la frecuencia 11, haciendo ver que el hecho de reenviar impresiones tiene efecto.

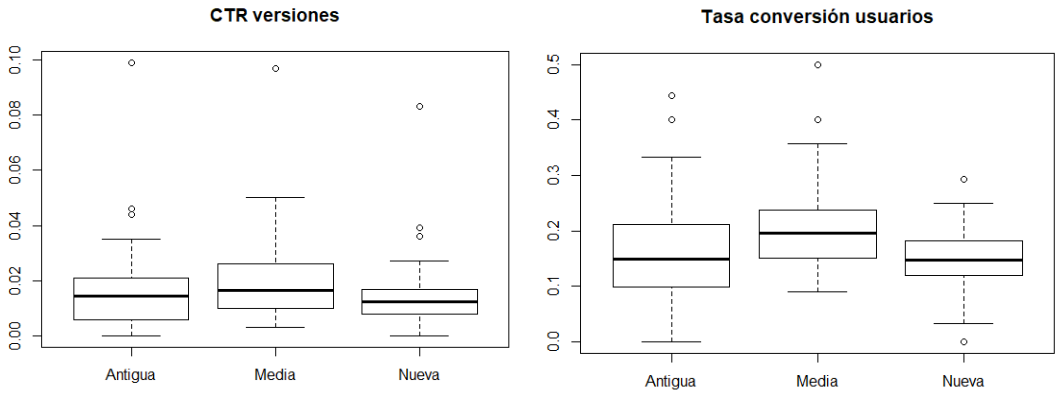

Figura 19: Diagrama de Cajas Ratios Versiones

Los gráficos de tendencia mostrados se pueden apoyar en los diagramas de cajas expuestos, en los que se intuyen más diferencias para la tasa de conversión de usuarios en aquellos con dispositivos con la versión intermedia.

En ambos casos parece que la hipótesis previa enunciada al inicio del apartado puede tener sentido, ya que la versión antigua e intermedia vuelve a presentar mejores ratios que la nueva.

Para verificar estadísticamente la hipótesis planteada, es decir, que los usuarios con versiones más antiguas de móvil van a tener mejores ratios de conversión que aquellos con la versión nueva, se ha comprobado si los ratios presentan diferencias significativas según la versión mediante los test comentados anteriormente.

Para poder hacer este contraste, se ha aplicado previamente un test de normalidad a las variables a analizar, obteniendo como resultado el indicado en la Tabla 8.

	Antigua	Media	Nueva
CTR	0.0001	0.0000	0.0000
Conversión usuarios	0.6756	0.2789	0.9367

Tabla 8: p-valores Contraste de Normalidad

Para el CTR, en todos los casos el p-valor resultante es significativo incluso al 1%, rechazando por tanto la hipótesis nula de normalidad y utilizando para contrastar las diferencias entre variables el test de Wilcoxon.

	Antigua	Media	Nueva
Antigua	1	0.4402	0.8179
Media		1	0.1322
Nueva			1

Tabla 9: p-valores Test de Wilcoxon no pareado sobre CTR

Al aplicar el contraste, en ningún caso el p-valor es inferior a 0.1, por lo que no se puede asegurar que existan diferencias significativas en el caso del CTR.

	Antigua	Media	Nueva
Antigua	1	0.1588	0.0548
Media		1	0.0002
Nueva			1

Tabla 10: p-valores Test de Wilcoxon pareado sobre CTR

Por otra parte, realizando el mismo test pareado, se rechaza la hipótesis nula al comparar la versión antigua con la nueva (p-valor=0.0548) y la intermedia con la nueva (p-valor=0.0002), pudiendo concluir que existen diferencias significativas en ambos casos.

En el caso de la tasa de conversión de usuarios el resultado del test de normalidad es en todos los casos no rechazar la hipótesis nula, considerando la variable normal y pudiendo aplicar un test de comparación de medias.

	Antigua	Media	Nueva
Antigua	1	0.1149	0.563
Media		1	0.0223
Nueva			1

Tabla 11: p-valores t.test no pareado sobre Tasa de Conversión de Usuarios

Al hacerlo, se comprueba la existencia de diferencias significativas entre la versión intermedia y la nueva, con un p-valor de 0.02.

	Antigua	Media	Nueva
Antigua	1	0.0718	0.4197
Media		1	0.0019
Nueva			1

Tabla 12: p-valores t-test pareado sobre Tasa de Conversión de Usuarios

En el caso del la prueba t de Student pareada, se obtienen unos p-valores inferiores al 10% al contrastar la versión intermedia con la nueva y con la antigua.

4.3.1.2 Tecnología del dispositivo

Al igual que en el caso anterior, se va a suponer una hipótesis de partida, para más tarde comprobar su veracidad:

- Desde el punto de vista personal, los usuarios se sienten más cómodos utilizando la versión más antigua de móvil, que conocen mejor, por lo que se puede suponer un CTR más alto para esa versión.

Esta variable, "is_4G", del dataset original es del mismo tipo que la anterior, una variable categórica que toma dos posibles valores, 4G o no.

La forma de operar ha sido la misma que en el caso de las versiones, dividiendo primero los datos en dos subconjuntos, uno con los datos cuando el dispositivo es 4G y otro cuando no lo es.

De tal forma y para ejemplificarlo, la tabla 8 representa un extracto de los datos cuando la tecnología usada es 4G.

impression_time	user_id	app_code	os_version	is_4G	is_click
16/11/2018 13:38	1150	249	latest	1	0
16/11/2018 15:13	1150	249	latest	1	0
24/11/2018 22:56	1151	296	old	1	1
29/11/2018 14:02	1154	207	latest	1	0
06/12/2018 20:12	1155	386	old	1	0
07/12/2018 17:10	1155	386	old	1	0
16/11/2018 0:57	1158	386	latest	1	0
19/11/2018 16:04	1158	386	latest	1	0
20/11/2018 0:20	1158	386	latest	1	0
26/11/2018 9:06	1158	386	latest	1	0
26/11/2018 9:36	1158	386	latest	1	1
09/12/2018 0:48	1159	207	latest	1	0
28/11/2018 10:26	1160	207	latest	1	0
28/11/2018 11:20	1160	207	latest	1	0
01/12/2018 16:43	1160	207	latest	1	0
01/12/2018 17:07	1160	207	latest	1	0

Tabla 13: Ejemplo Datos Tecnología 4G

La información presentada es idéntica a los casos anteriores, por lo que se va a hacer una breve descripción de los datos de los dos subconjuntos.

Tecnología	Impresiones	Clicks	Usuarios distintos
4G	74896	2767	32960
No 4G	131644	4889	52165
Total	206540	7656	85125

Tabla 14: Resumen Datos según Tecnología

Como refleja la Tabla 14 predominan las impresiones recibidas y los clicks en dispositivos sin 4G, es decir, con peor tecnología.

En cuanto a la distribución diaria de impresiones y clicks, en ambos casos es similar a las ya vistas, con ligeras variaciones entre ambas en los primeros días de diciembre.

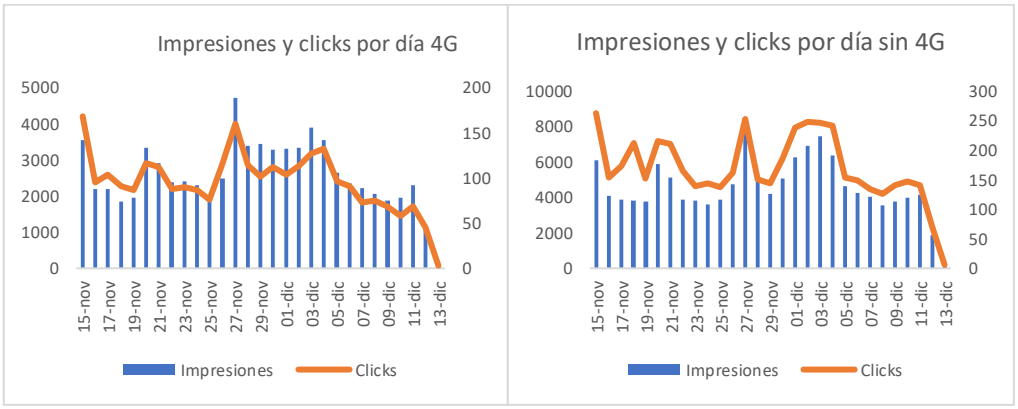

Figura 20: Distribución temporal impresiones y clicks según tecnología

Una vez descritos los datos de este punto, se ha calculado el CTR por frecuencias de impresiones para ambos conjuntos, resultando el gráfico siguiente.

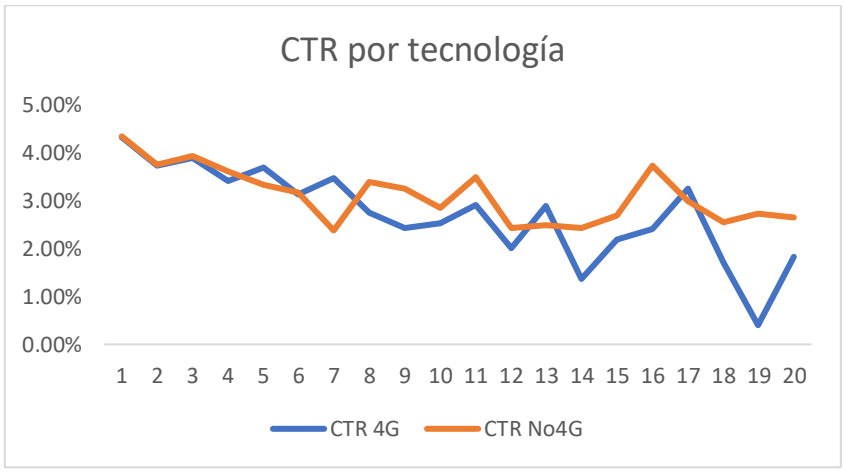

Figura 21: CTR por Impresiones según Tecnología

La Figura 21 muestra un ratio de clicks sobre impresiones casi idéntico para las impresiones enviadas a dispositivos con y sin 4G para una frecuencia de una, dos y tres repeticiones. Parece que después de la tercera impresión la tendencia es descendente en ambos casos, pero sin diferencias claras a simple vista.

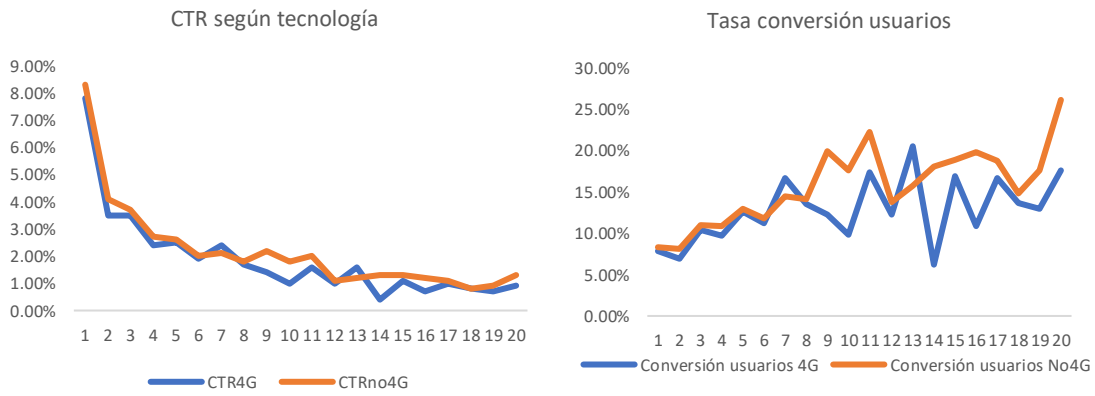

Figura 22: Ratios de Conversión por Usuarios según Tecnología

Respecto a los grupos de usuarios que reciben cada frecuencia de impresiones, el patrón es el mismo que en casos anteriores, prácticamente sin diferencias significativas entre los grupos analizados en este apartado.

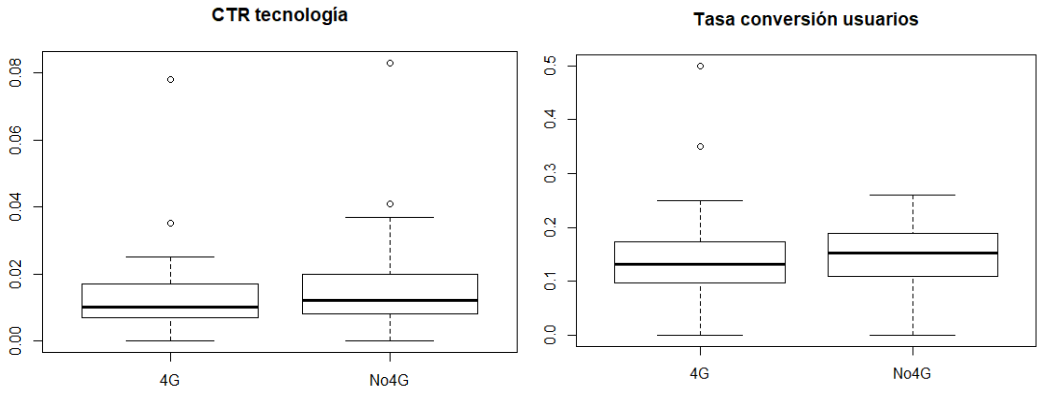

Figura 23: Diagrama de Cajas Ratios Tecnología

Se puede ver en los diagramas de cajas que el valor central de los ratios pertenecientes a los dispositivos sin 4G está ligeramente por encima. Hay mayor diferencia en la tasa de conversión de usuarios originada probablemente por aquellos usuarios que reciben 8 repeticiones o más, cuyo ratio es más alto excepto para la frecuencia 13.

Para corroborar estos resultados se han aplicado los mismos test que en el apartado 4.3.1.1., empezando por el de normalidad. El resultado ha sido rechazar la hipótesis nula de normalidad en los dos casos para el CTR (p-valores inferiores al 1%) y no rechazarla en los casos de la tasa de conversión de usuarios.

Con esta información, los test aplicados para los dos ratios han sido Wilcoxon y la prueba t de Student, respectivamente.

En cuanto al CTR, el test de Wilcoxon indica no rechazar la hipótesis nula, mientras que al hacerlo pareado se obtiene un p-valor inferior al 5%, pudiendo afirmar que hay diferencias significativas según la tecnología del dispositivo.

Por su parte, la tasa de conversión de usuarios da como resultado un p-valor muy cercano a 0 tanto para la muestra pareada como sin parear, por lo que se considera que el comportamiento según la tecnología utilizada también es diferente.

4.3.1.3 Conclusión Efecto Marco

Gracias a los contrastes estadísticos aplicados, se puede afirmar que el Efecto Marco es relevante para el usuario a la hora de decidir si hacer click o no en la publicidad que recibe.

Ahora bien, el supuesto planteado de inicio sólo se ha cumplido en el caso de la versión de móvil, siendo los ratios de la versión antigua e intermedia mejores que los de la nueva, mientras que en caso de la tecnología no se han encontrado diferencias significativas.

Hablando a nivel de compañía que quiere promocionar un producto, se debería recomendar llevar a cabo estrategias publicitarias diferentes en función de la versión del dispositivo que tienen los usuarios a los que se va a enviar la publicidad.

4.3.2 Efecto Recencia

El Efecto Recencia, definido como la tendencia a recordar mejor la información presentada en último lugar, se ha tratado de aplicar aquí debido a que el problema al que el decisor recibe repetidas veces la misma información, y se pretende averiguar si el tiempo que tarda en recibir la última información que se le presenta influye en su decisión.

En esta ocasión el planteamiento es diferente al expuesto anteriormente, se trata de dividir el conjunto de datos en base a las diferencias de tiempo entre una impresión y otra, y ver cómo actúa el CTR en cada conjunto.

Según el sesgo, la hipótesis de partida en este caso sería la siguiente:

- Las personas recuerdan mejor la última información presentada, por lo que aquellas impresiones que tardan menos tiempo en recibirse desde la última tienen más probabilidad de convertir en click que aquellas que tardan más tiempo.

Para evaluar esta hipótesis, se calcula primero una nueva variable llamada "diferencias" que contenga las diferencias de tiempo en segundos entre una impresión y otra para cada usuario, y se añade a los datos originales. Dicha variable no toma valores para las primeras impresiones, ya que no hay una anterior sobre la que hacer la diferencia.

Después, observando la distribución de valores mediante su histograma, se ve como es totalmente asimétrica, agrupando la mayoría de sus valores en torno al 0.

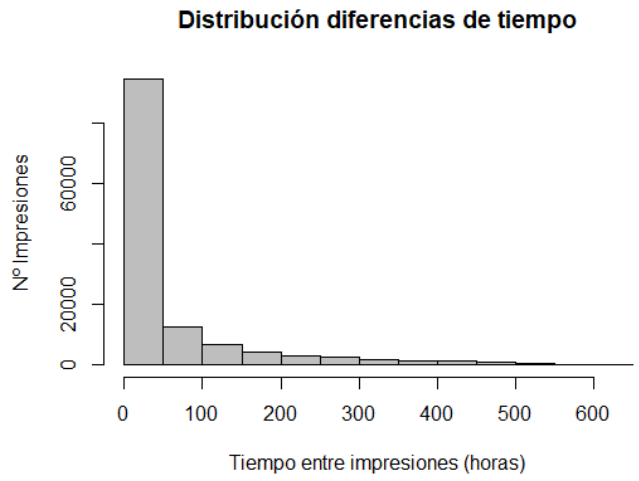

Figura 24: Histograma Variable "diferencias"

Observando las medidas de centralización y dispersión, se puede ver que la existencia de un número reducido de valores muy elevados provoca una gran distorsión en el análisis, siendo la media mucho más alta que la mediana.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0167	0.9833	16.3333	58.5865	60.9500	604.1000

Figura 25: Resumen Estadístico de la Variable "diferencias"

La media resultante es de aproximadamente 58 horas, mientras la mediana es de 16 horas. Teniendo esto en cuenta se han propuesto dos análisis para localizar diferencias útiles a nivel de negocio.

En primer lugar, se ha utilizado como punto de corte la media, estudiando la conversión para el conjunto de impresiones con valores inferiores y superiores a ella.

Se obtiene, por un lado, un subconjunto de datos donde los tiempos entre impresiones son inferiores a ese valor medio y por otro, un subconjunto donde los tiempos entre repeticiones son superiores.

Como ejemplo se ve en la tabla 15 un extracto del conjunto de datos con tiempos inferiores.

impression_time	user_id	app_code	os_version	is_4G	is_click	diferencias
23/11/2018 22:11	197	150	latest	0	0	2160
30/11/2018 14:49	197	150	latest	0	0	72900
30/11/2018 17:45	197	150	latest	0	0	10560
02/12/2018 0:45	197	150	latest	0	0	111600
04/12/2018 17:57	197	150	latest	0	0	18480
05/12/2018 17:58	197	150	latest	0	0	86460
06/12/2018 20:23	197	150	latest	0	1	95100
22/11/2018 11:28	199	422	old	0	0	960
25/11/2018 16:53	199	422	old	0	0	10620
29/11/2018 0:17	199	422	old	0	0	1080
29/11/2018 17:04	199	422	old	0	0	60420
01/12/2018 12:55	199	422	old	0	0	157860
01/12/2018 13:13	199	422	old	0	0	1080
02/12/2018 1:20	199	422	old	0	1	43620
26/11/2018 10:35	200	375	old	0	0	46500
26/11/2018 10:38	200	375	old	0	0	180
27/11/2018 5:41	200	375	old	1	0	68580
27/11/2018 9:34	200	375	old	0	0	13980

Tabla 15: Ejemplo Datos con Tiempos Inferiores a la Media (en segundos)

Respecto a los datos resultantes, es conveniente destacar que hay una clara diferencia en cuanto a tamaño de los dos conjuntos, ya que los datos superiores a la media son aproximadamente un 25% del total.

El estudio realizado es similar al del Efecto Marco, con la diferencia de que sólo se ha propuesto el análisis por grupos de impresiones, ya que tal como se ha planteado y tratado los datos, se ha agrupado por impresiones y no por usuarios.

Dando lugar a los siguientes resultados en lo que al CTR respecta.

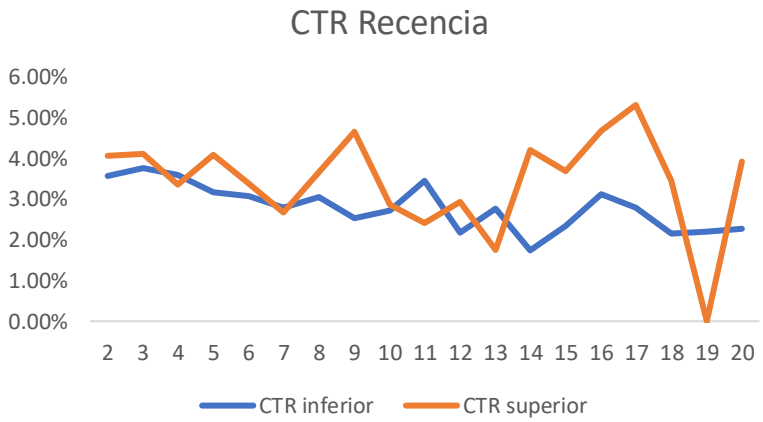

Figura 26: Ratios CTR Recencia según la Media

A simple vista, parece que para las primeras frecuencias de impresiones el CTR es más elevado en las impresiones recibidas con mayores lapsos de tiempo.

Otra observación importante es que, al haber eliminado las primeras frecuencias, la primera frecuencia sobre la que se ha calculado el CTR es la segunda.

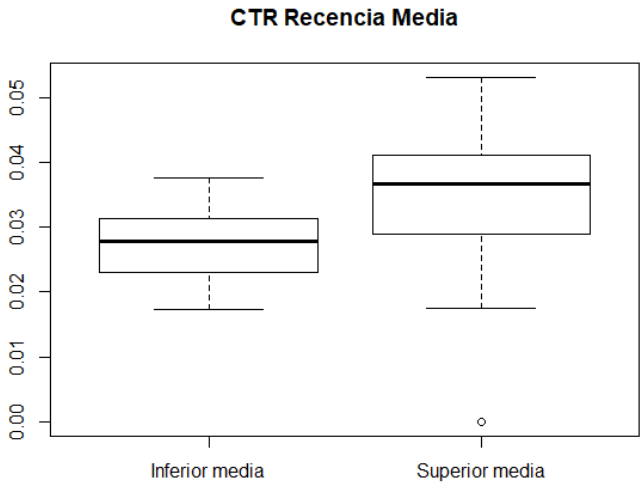

Figura 27: Diagrama de Cajas Ratios CTR según la Media

En el diagrama de cajas superior se ve como aquellas impresiones con un lapso mayor de tiempo se mueven en un rango más amplio y su valor central es considerablemente superior, apoyando así la gráfica de línea.

Para comprobar si estas diferencias observadas son realmente significativas se ha recurrido a los contrastes comentados previamente, en este caso recurriendo al test de Wilcoxon porque al contrastar la normalidad de las variables el resultado es rechazar la hipótesis nula de normalidad en el caso del CTR de los datos superiores a la media.

Al realizar el test de comparación de medianas, se obtiene un p-valor de 0.0082, rechazando la hipótesis nula de igualdad de medianas. Lo mismo ocurre al hacerlo pareado, dando como resultado un p-valor de 0.0323, concluyendo por tanto que el tiempo entre anuncios influye en el comportamiento del usuario a la hora de hacer click.

El otro análisis llevado a cabo tiene como objetivo ver si existen diferencias en la conversión tomando como puntos de corte los cuartiles de la variable "diferencias".

En este caso se divide la base de datos en cuatro conjuntos distintos y se realiza el mismo estudio que en caso anterior, calculando el CTR y comprobando mediante los contrastes si hay diferencias significativas entre ellos.

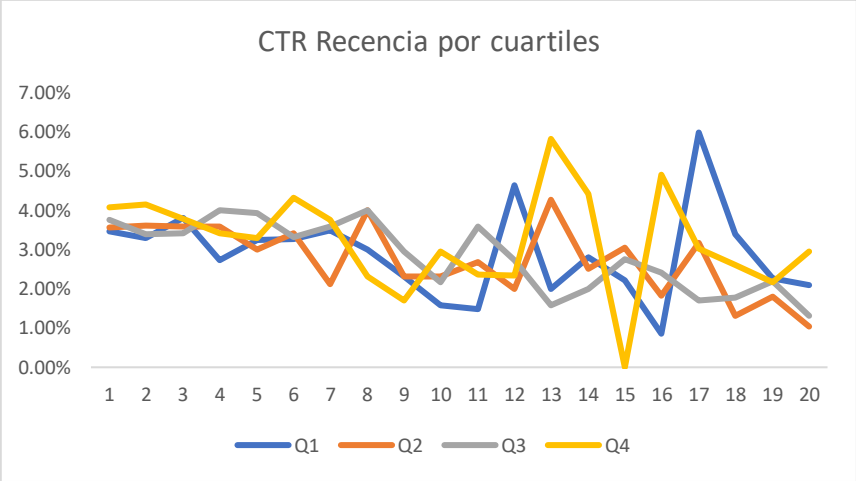

Figura 28: Ratios CTR Recencia según los Cuartiles

Como se ve en la gráfica superior, los ratios presentan cierta estabilidad para las primeras frecuencias de impresiones y mayor volatilidad conforme aumentan. En el siguiente gráfico de cajas se observa como sus valores se concentran en un rango similar entorno a una mediana de 0.03 aproximadamente.

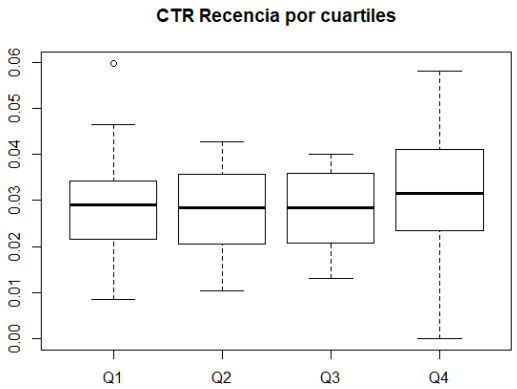

Figura 29: Diagrama de cajas Ratios CTR según cuartiles

Para comprobar estadísticamente si esas ligeras diferencias son representativas se han realizado los contrastes repetidos en casos anteriores, dando lugar a los siguientes resultados.

Los contrastes de normalidad realizados arrojan los siguientes p-valores para los CTR de cada cuartil.

Q1	Q2	Q3	Q4
0.3797	0.6935	0.1428	0.8572

Tabla 16: p-valores Contraste de Normalidad CTR por Cuartiles

En todos los casos la variable CTR presenta normalidad, por lo que se ha aplicado la prueba t de Student,

	Q1	Q2	Q3	Q4
Q1	1	0.6752	0.8327	0.4132
Q2		1	0.8075	0.2010
Q3			1	0.2726
Q4				1

Tabla 17: p-valores t.test no pareado CTR por Cuartiles

	Q1	Q2	Q3	Q4
Q1	1	0.6428	0.8275	0.4245
Q2		1	0.7354	0.1432
Q3			1	0.2928
Q4				1

Tabla 18: p-valores t.test pareado CTR por Cuartiles

Como se puede apreciar en las tablas 17 y 18, comparando cada ratio dos a dos, en ningún análisis existen diferencias significativas, tanto al hacerlo pareado como no pareado.

4.3.2.1 Conclusión Efecto Recencia

En este caso la hipótesis planteada al inicio del apartado no se verifica, ya que tras realizar los dos tipos de análisis, el basado en el rango intercuartílico no hace distinción entre las impresiones según el tiempo, mientras que el análisis basado en la media presenta diferencias significativas, pero la tasa de conversión es mayor cuando los usuarios esperan más tiempo entre impresiones.

Se puede concluir por lo tanto que a nivel empresarial, lo que le interesa a una empresa basándose en los resultados obtenidos, sería tomar la media como punto de referencia y plantear distintas estrategias a la hora de enviar impresiones a los usuarios en función de ella.

4.4 Árbol de decisión

La última parte de la investigación consiste en una propuesta de árbol de decisión, un método utilizado en muchas disciplinas como modelo de predicción, que consiste en delimitar puntos en los que tomar decisiones de acuerdo a reglas. Existen árboles de clasificación o regresión según cómo sea la variable objetivo que se quiere predecir, discreta o continua respectivamente.

En nuestro caso, el objetivo es predecir si una persona va a hacer click o no (variable respuesta discreta "Click") en base a las variables disponibles, fundamentalmente el número de impresiones, pero también la versión de móvil y su tecnología.

Se ha hecho uso de la librería "rpart" de R y se han tenido en cuenta los parámetros que se mencionan a continuación para el tratamiento y mejora del árbol:

- Minsplit: número mínimo de observaciones en el nodo previo a que el árbol haga una división.
- Minbucket: mínimo número de observaciones en el nodo posterior a una división.
- Parámetro de complejidad (CP)

Para trabajar este apartado se ha hecho uso de la tabla intermedia comentada al inicio del análisis (Tabla 4), ya que incluye las variables transformadas de la forma necesaria para la construcción del modelo.

user_id	Impresiones	Click	version	is_4G
35038	2	0	latest	0
35039	8	0	latest	0
35040	1	0	latest	1
35042	1	0	old	1
35043	7	0	intermediate	0
35044	1	1	latest	0
35045	1	0	old	0
35046	3	0	old	0
35047	16	0	intermediate	0

Tabla 4: Datos Agrupados por Usuario

En primer lugar, se han dividido estos datos aleatoriamente, tomando el 75% como datos de entrenamiento con los que construir el modelo, y el 25% restante como datos de test sobre los que aplicar el modelo propuesto, para evaluar su calidad a través de la matriz de confusión.

En un principio, la muestra de entrenamiento tenía una proporción de clicks muy desbalanceada, un 10% del total, llevando al modelo resultante a predecir siempre que el usuario no hacía click. Para tratar de solucionarlo, se ha balanceado la muestra usando la librería "imbalance" de RStudio, igualando la proporción de clicks y no clicks.

Tras un gran trabajo de mejora y optimización de los parámetros del modelo en busca de un árbol suficientemente profundo para establecer diferencias claras en base a las variables disponibles, se ha llegado al siguiente modelo.

El parámetro *minsplit* se ha reducido a 2 y el parámetro *minbucket* a 1, dando mayor flexibilidad al modelo para la creación de nodos y hojas.

Por su parte, para encontrar el parámetro de complejidad que mejor relacione la profundidad y complejidad del árbol respecto a su capacidad predictiva en los datos de test, se ha hecho crecer el árbol hasta su mayor extensión y se ha podado después hasta localizar el valor que minimice el error estándar o *xerror*.

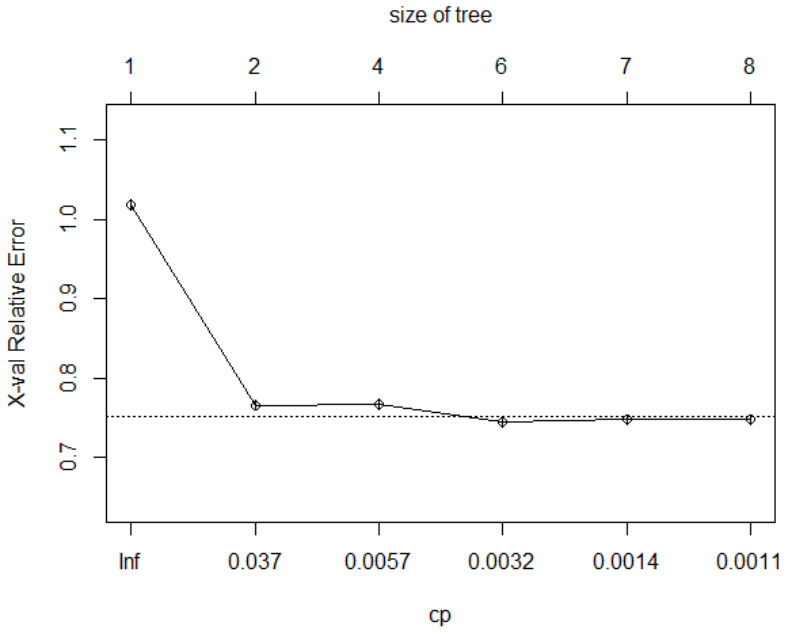

Figura 30: Parámetro de Complejidad

En la Figura 30 se muestra dicho error en relación al parámetro de complejidad. Teniéndola en cuenta, se ha escogido un parámetro de complejidad de 0.004 (situado entre 0.0032 y 0.0057), dando lugar al siguiente árbol.

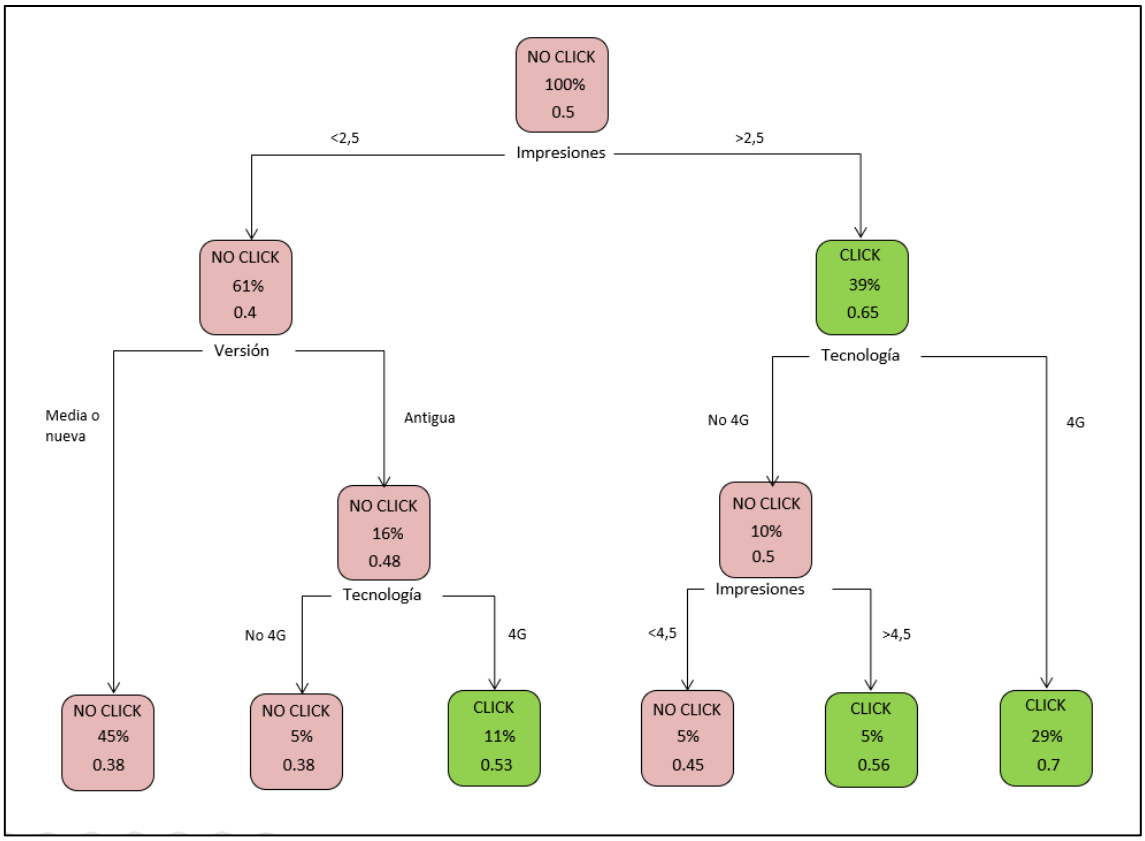

Figura 31: Árbol de Decisión

A la vista del árbol resultante, la variable principal a considerar es el número de impresiones, aumentando la probabilidad de que un usuario haga click cuando recibe más de dos impresiones hasta el 65%, y siendo más probable que no lo haga cuando recibe dos o menos.

Profundizando en el árbol a través de la primera división, encontramos otra en función de la versión, indicando que si el dispositivo tiene la versión antigua y 4G es más probable que el usuario haga click tras haberle mandado menos de tres impresiones.

Por el otro lado, queda claro que si el usuario dispone de un dispositivo con 4G su probabilidad de clicar también aumenta, habiendo recibido más de dos impresiones.

Aplicando este modelo a los datos de test, se obtiene la siguiente matriz de confusión, con las medidas de calidad que se verán después.

	Predicción	
	No click	**Click**
No click	11770	4555
Click	1514	842

Tabla 19: Matriz de Confusión

Para entender las medidas de calidad de los resultados obtenidos por el modelo al aplicarlo sobre los datos de test que se describen a continuación, es necesario especificar que los casos negativos son aquellos en los que un usuario no hace click, mientras que los positivos se dan cuando el usuario clica.

- Exactitud o *accuracy:* refleja el acierto del modelo, dividiendo los datos bien clasificados entre el total de datos.

  Exactitud = $\frac{TP+TN}{TP+FP+TN+FN} = 67.5\%$

- Precisión o *precision*: hace referencia a los casos positivos clasificados correctamente del total de positivos detectados.

  Precisión = $\frac{TP}{TP+FP} = 15.6\%$

- Sensibilidad o *recall*: indica los casos positivos que han sido clasificados correctamente por el modelo.

  Sensibilidad = $\frac{TP}{TP+FN} = 55.6\%$

- Especificidad o *specifity*: es el porcentaje los casos negativos que el modelo ha clasificado correctamente.

  Especificidad = $\frac{TN}{TN+FP} = 71.1\%$

Como se puede apreciar, el modelo da unos resultados relativamente buenos, con un porcentaje de acierto total aceptable, pero con algunas dificultades sobre todo a la hora de predecir correctamente los casos positivos. De esta forma, permite dar unas primeras aproximaciones sobre el número de impresiones que se deben mandar a un usuario para aumentar la probabilidad de que haga su primer click.

A la vista del árbol de clasificación (Figura 31), se recomendaría establecer un límite mínimo de 3 impresiones, y potenciar el envío de publicidad a aquellos dispositivos con 4G.

Es posible que la calidad del modelo presentado pueda aumentar si se amplía el número de variables predictoras añadiendo información que permita diferenciar mejor las situaciones en las que un usuario tiene mayor predisposición al click.

5 CONCLUSIONES

En este apartado final se pretende hacer un repaso por las diferentes fases de la investigación y resultados obtenidos, desde la fase inicial de recogida de información hasta la final de creación del modelo.

En primer lugar, hay que comentar que se pudo plantear la relación entre conceptos como la aceptación de la publicidad online y la Economía del Comportamiento debido a la creencia de que la decisión de clicar en determinado anuncio es un proceso de toma de decisión muy rápida que depende de muchos factores estudiados en esta rama de la economía.

Comenzando por esa primera fase, se ha buscado toda la información relacionada con el método que se iba a tratar en un principio, *frequency capping*, y la Economía del Comportamiento para valorar de qué forma se iba a incorporar al análisis. Una vez acotada la información y definidos todos los sesgos cognitivos, se redujeron a aquellos que podían tener cabida de alguna forma, a la espera de encontrar unos datos que permitieran incluirlos.

En la búsqueda de un conjunto de datos propicio para este análisis es donde se ha invertido más tiempo, por la dificultad de encontrar unos lo suficientemente grandes para poder ser tratados con las herramientas mencionadas, y que tuvieran variables que permitieran relacionar las impresiones y clicks con los sesgos definidos en la fase anterior. Tras revisar varios conjuntos de datos se ha recurrido a los datos de un *ecommerce* encontrado en la plataforma *kaggle*, que incluye variables como la versión o la tecnología del dispositivo usado por los usuarios a la hora de recibir las impresiones, así como la fecha en que lo hacen. Estas variables han permitido incorporar dos de los sesgos definidos previamente, el Efecto Marco y el Efecto Recencia.

Como se ha descrito en el estado del arte, hay que decir que la parte innovadora del trabajo ha sido esta incorporación de la Economía Conductual al estudio sobre el tratamiento de

la publicidad online, y en concreto del análisis del número de impresiones a enviar a los usuarios.

Tras un primer análisis descriptivo de los datos, donde se ha visto la distribución de las impresiones, cabe destacar que la campaña analizada dura aproximadamente un mes y se centra en dispositivos con la versión nueva.

Después de explicar el procedimiento a seguir, se ha comenzado a incorporar los diferentes sesgos, y a dividir el *dataset* general en función de ellos para calcular los ratios que han permitido comprobar que realmente existen diferencias según el contexto en el que un usuario recibe una publicidad, tanto en forma (Efecto Marco) como en tiempo (Efecto Recencia).

Es importante recalcar que las hipótesis planteadas en cada apartado son supuestos sin base científica que se han tratado de comprobar para darle un mayor dinamismo al trabajo.

Así pues, esta investigación, que en un principio parecía más enfocada al tratamiento de un método de limitación de publicidad muy específico como el *frequency capping*, ha derivado principalmente en un estudio sobre la existencia de variaciones significativas en la reacción de los individuos ante la publicidad online en función de aspectos conductuales, que como se ha podido comprobar, sí existe.

En último lugar y muy relacionado con la limitación de dicha publicidad, se ha propuesto un árbol de decisión para clasificar y predecir en qué tipo de dispositivos y con qué número de impresiones es más factible que los usuarios hagan click, utilizando el conjunto de datos global.

Pese a que el modelo no ofrece el mejor resultado en cuanto a calidad por la limitación tanto de variables como de tiempo disponibles, deja patente la necesidad de reenviar la misma publicidad a un mismo usuario si se busca aumentar su probabilidad de conversión (aumentando la probabilidad de click cuando se envían 3 impresiones o más), y abre una línea de investigación sobre la que profundizar en un futuro.

En este aspecto y en vista de los resultados, existen distintas formas de continuar desarrollando la investigación en busca de posibles mejoras.

La primera de ellas sería disponer de unos datos más completos, que contengan más información sobre la situación en la que se encuentran los usuarios al recibir las impresiones publicitarias, como por ejemplo las webs donde ven esa publicidad, a qué

producto hace referencia, etc. En este sentido, también es posible profundizar más en los datos disponibles, tratando de estudiar secciones más concretas de la muestra para ver si se pueden extraer conclusiones válidas.

Por otro lado, se pueden proponer y aplicar otros modelos que permitan explotar estos datos según nuestras necesidades. Aquí se ha utilizado un árbol de decisión con el objetivo de establecer unas reglas claras e interpretables para saber en qué casos aumenta la tasa de conversión, pero si el objetivo es únicamente predecir cuándo es más probable que un usuario va a hacer click, se podría trabajar por ejemplo con redes neuronales, ya que posiblemente darían una predicción mejor con el hándicap de no saber cómo se ha llegado a ella.

En resumen, en esta investigación se ha comprobado que las condiciones en las que las personas recibimos publicidad de diferentes empresas realmente importan, variando nuestra forma de reaccionar ante ellas, es decir, nuestro comportamiento, en función de dichas condiciones.

Por ello, para las empresas es muy importante analizar y conocer estas diferencias a la hora de preparar campañas y estrategias publicitarias si quieren llegar de forma efectiva al público.

6 BIBLIOGRAFÍA

BUCHBINDER N., FELDMAN M., GHOSH A., NAOR J. (2011). Frequency Capping in Online Advertising. *In: Dehne F., Iacono J., Sack JR. (eds) Algorithms and Data Structures. WADS 2011. Lecture Notes in Computer Science, vol 6844. Springer, Berlin, Heidelberg*. https://doi.org/10.1007/978-3-642-22300-6_13

CARMINATI, L. (2020). Behavioural Economics and Human Decision Making: Instances from the Health Care System. *Health Policy, Volume 124, Issue 6,* 659-664 doi: 10.1016/j.healthpol.2020.03.012

FARAHAT, A. (2009). Privacy Preserving Frequency Capping in Internet Banner Advertising, 1147-1148. *WWW '09: Proceedings of the 18th international conference on World wide web* https://doi.org/10.1145/1526709.1526900.

FAN, W. (2017). Education and Decision Making: An experimental study on the framing effect in China. *Nanjing University.* doi: 10.3389/fpsyg.2017.00744

HOJJAT A., TURNER J., CETINTAS S., YANG J. (2014)**.** Delivering Guaranteed Display Ads under Reach and Frequency Requirements. *University of California*, 2278-2284.

KAHNEMAN, D. (2012). Pensar rápido, pensar despacio. *Editorial: DEBATE.*

KAHNEMAN D., TVERSKY A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology. 3 (3): 430–454. doi*:*10.1016/0010-0285(72)90016-3.

KAHNEMAN, D. (2002). Mapas de racionalidad limitada. *Revista Asturiana de Economía (Traducción por Mario Piñeda). ISSN 1134-8291, Nº. 28, 2003,* 181-225.

KAHNEMAN D., TVERSKY A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, Vol. 47, No. 2*. 263-292. https://doi.org/10.2307/1914185

KOTLER, P. (1967). Principios del Marketing: Análisis, planificación y control. *Editorial: Prentice Hall.*

KOTLER P., KARTAJAYA H., SETIAWAN I. (2016). Marketing 4.0 Moving from traditional to digital. *Editorial: John Wiley & Sons, Inc.*

MURPHY J., HOFACKER C., MIZERSKI R. (2006). Primacy and Recency Effects on Clicking Behavior. *Journal of Computer-Mediated Communication, Volume 11*. 522-535.

QIN R., YUAN Y., WANG F. (2016). Exploring Optimal Frequency Caps in Real Time Bidding Advertising. *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Volume: 1*. 385-392.

SHANAHAN J., VAN DEN POEL D. (2010). Determining optimal advertisement frequency capping policy via Markov decision processes to maximize click through rates. *Ghent University, Independent Consultant (USA)*

SHASTITKO, A. (2017). Behavioral Economics: Application of the Methods of Cognitive Psychology to Economics. *Social Sciences, No.2, Vol.0048*. 142-151. DOI: 10.21557/SSC.48907824

THALER, R. (1987). Anomalies. The January effect. *Journal of economic perspectives, vol. 1, no. 1, summer 1987,* 197-201. DOI: 10.1257/jep.1.1.197

THALER, R. (2018). Economía del comportamiento: pasado, presente y futuro (Behavioral Economics: Past, Present, and Future). *Revista de Economía Institucional, Vol. 20, No. 38.*

WANG P., LIUNG B., HONG T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting, Volume 32, Issue 3*, 585-597. https://doi.org/10.1016/j.ijforecast.2015.09.006

ZINKEVICH, M. (2010). Optimal online frequency capping allocation using the weight approach. http://martin.zinkevich.org/publications/weights.pdf