¿ES EL *BIG DATA* EL SIGUIENTE PASO EN LA DIGITALIZACIÓN DE LA EMPRESA?

MARÍA TERESA BALLESTAR
ESIC Business School
DOMINGO RIBEIRO
Universidad de Valencia
JORGE SAINZ
Universidad Rey Juan Carlos

La transformación digital, entendida como el fenómeno de conexión inteligente entre productos y actividades a través del intercambio de información gracias a las tecnologías digitales, permite la modificación de los modelos de negocio impactando tanto en los proveedores de dichos bienes y servicios como en los consumidores. Como señalan Porter y Heppelmann (2015), esto permite a las compañías incrementar de forma exponencial la creación de valor para sus clientes a través de la modificación de sus cadenas de valor gracias a la aparición de nuevas funcionalidades, mayor seguridad en los procesos y mejora en la eficiencia y en la optimización de las posibilidades que ofrecen a sus clientes, tanto en ámbitos económicos, como en el análisis político, deportivo, etc. (por ejemplo, Capilla y Sainz, 2009).

La optimización de estos procesos pasa por la gestión de cantidades masivas de datos que ayuden no sólo a las compañías sino también a las administraciones públicas, al sector educativo o las organizaciones no gubernamentales a ofrecer soluciones más avanzadas a sus clientes que les permitan maximizar su creación de valor. Sin embargo, no todas las empresas están preparadas para la optimización de estos procesos, ya que la adopción de las tecnologías para el análisis masivo de datos no es el único factor a tener en cuenta, sino que también influyen la forma en que se adoptan dichas tecnologías y el uso de las mismas. Las decisiones sobre estos aspectos pueden cambiar la misma esencia del desarrollo de las actividades de las empresas, pese a que el usuario final apenas encuentre diferencias, más allá de la mejora en la eficiencia del servicio, en la calidad del producto o en la mejora de la selección.

El *Big Data* es una de estas tecnologías disruptivas que, correctamente adoptada, modifica la estructura del negocio, añade competitividad y flexibilidad a las instituciones y las dota de un mayor conocimiento propio y por lo tanto de la capacidad de decidir nuevas estrategias que beneficien su desempeño. En este artículo, además de explicar qué es el *Big Data*, el impacto que tiene su implementación en las compañías, así como sus principales y nada triviales barreras de acceso, se realiza un ejercicio empírico de las capacidades de mejora del negocio mediante distintas técnicas analíticas vinculadas a la ciencia de los datos.

Este ejercicio empírico consiste en un caso de uso sobre el mercado *online* de *cashback*, donde los negocios incentivan a los usuarios mediante *cashback* o reintegros por navegar y/o realizar actividades a través de la plataforma *online*.

¿QUÉ ES *BIG DATA* Y CÓMO SE USA?

El término «*Big Data*» surge académicamente con Francis Diebold, profesor de la Universidad de Pensilvania que utiliza el término en su artículo publicado en 2003 «*'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting*» (Diebold, 2012). Aunque el concepto es claramente anterior, desde la psicohistoria del Profesor Hari Seldon en la Saga de Fundación de Isaac Assimov, basada en la predicción del futuro a través de utilizar grandes volúmenes de datos del pasado, a las propuestas que sobre el uso masivo de datos se realizan ya en los años 50 en distintas administraciones, y que con la expansión de las tecnologías tienen su primera expresión como tal en los desarrollos de John Mashey para la empresa Silicon Graphics en los años 90. En éstos se utiliza el concepto similar al *Big Data* de un amplio grupo de desarrollo de bases de datos, gestión de los mismos, funcionamiento de algoritmos, etc. (Diebold, 2013 y Lohr, 2013b).

Si hay un consenso en la literatura sobre el origen del concepto y del nombre, no lo hay sobre una definición sobre qué es *Big Data*, su pertenencia al *Data Science*, la ciencia de los datos y su análisis (1). Siguiendo a McAfee y Brynjolfsson (2012), Pospiech y Felden (2012), Morabito (2014) o Morabito (2015), se puede escoger una definición derivada de sus características, que son las denominadas cuatro Vs: Volumen, Velocidad, Variedad y Veracidad, y que podemos caracterizar de la siguiente forma:

Volumen se refiere fundamentalmente al uso masivo de datos que se almacenan y utilizan en las bases de datos y que se utilizan. Por ejemplo, el tamaño de las bases de datos de donde se ha extraído la información para construir el ejercicio empírico de este artículo contiene un total de 3.500 billones de datos. La aparición de nuevas tecnologías, como la computación en la nube (*cloud computing*), es paralela a este desarrollo. El denominado «internet de las cosas» permite que se vaya generando una creciente cantidad de datos relativos al comportamiento de los individuos, lo que permite ofrecer servicios hasta ahora impensables para los ciudadanos. Otras tecnologías necesarias como MapReduce o los lenguajes o sistemas de programación como Hadoop, R o SPSS, que son empleados en estos ámbitos, han florecido por a la necesidad de utilizar recursos que permitan su análisis de forma sostenible y económicamente viable (Dong y Srivastava, 2013). Como ya se ha señalado antes, esta es sin duda la parte más relevante del *Data Science* (Gandomi y Haider, 2015).

- Velocidad, ya que para que los análisis resulten relevantes deben realzarse en tiempo real. El modelo desarrollado empíricamente en este artículo puede ser implementado en tiempo real por cualquier empresa para gestionar su funcionamiento e incrementar el rendimiento de su negocio. Los consumidores cada vez aprecian más la comodidad de disponer de los efectos del *Big Data* y los servicios que pueden obtener instantáneamente de los mismos, mientras que las empresas pueden obtener a través de la oferta de más servicios una mayor rentabilidad.

- Variedad, derivada de la forma de obtener los datos a través del internet de las cosas. Así pues, conviven las bases de datos estructuradas, es decir, aquellos datos que se conforman tradicionalmente con formatos claros que responden a una visión clásica de la estructura de datos, con las no estructuradas, que al contrario de las anteriores no tienen formatos transparentes hasta que no se vinculen con otros datos. Este es el caso, por ejemplo, de los datos extraídos de las opiniones y recomendaciones de internet, de la información que almacenan nuestros móviles sobre desplazamientos, etc. (Demchenko, De Laat, y Membrey 2014). Los datos utilizados en esta investigación provienen de distintos formatos, estructurados y no estructurados, combinados para obtener mejores resultados a través de la información que proporcionan.

- Veracidad, los datos deben ser creíbles, de fácil acceso y verificables. Como señalan Jin *et al*. (2015), este es un ámbito es especialmente importante por el efecto que tienen sobre el negocio y sobre las decisiones de los individuos. Sin embargo, con un mayor volumen de datos es también más difícil separar los datos «buenos» y útiles de los falsos y sin utilidad. A esto se añaden los problemas derivados del uso de los datos, su almacenamiento y la posibilidad del control del individuo al que se refieren los mismos, que conviene señalar, aunque no sean objeto de análisis en este momento, tienen un importante reflejo en las preocupaciones de la sociedad (Chatterjee, 2013 o Lohr, 2013a).

Además de las características anteriores, una de las definiciones más citadas en la literatura es la de Microsoft (2013), que define *Big Data* como «… el término cada vez más empleado para describir el proceso de utilizar una elevada capacidad de procesamiento –lo último en aprendizaje automático (*machine learning*) e inteligencia artificial– a conjuntos de información masiva generalmente muy compleja». Esta definición permite organizar el análisis sobre sus aplicaciones que es, en particular, lo que corresponde a este trabajo, sirviendo además para relacionarlo con la gestión del comercio electrónico, aportando a través del análisis de datos ventajas competitivas.

En este artículo nos centramos en un análisis empírico de un mercado en crecimiento, el de las *webs* que devuelven dinero por navegar o por realizar transacciones, conocido como *cashback*. La proliferación de la oferta de sitios web de *cashback* ha atraído un creciente interés en el mundo académico, pese a que esta área de investigación es aún muy incipiente. En este sentido, este artículo contribuye a la literatura académica existente analizando aspectos de singular relevancia en el marketing dentro del comercio electrónico como son las redes sociales, la lealtad de los consumidores, el rol que estos consumidores desempeñan dentro de la red social, su clasificación en segmentos y el *customer*

journey o evolución a través de los mismos. Todo ello permite comprobar empíricamente la eficiencia de la combinación de estrategias de marketing tradicionales con otras estrategias más innovadoras, como el *cash-back* y la recomendación mediante el boca a boca, o *word-of-mouth*.

En este estudio utilizaremos *machine learning* para diseñar una red neuronal artificial perceptrón multicapa (ANN-MLP) que permitirá validar los resultados obtenidos con una metodología de segmentación *Two-Steps Cluster Analisis*. De esta forma, se proporcionan herramientas a las compañías para conocer mejor las características de su portfolio de usuarios y mejorar la eficiencia de las campañas de recomendación *word-of-mouth*, en términos de optimización de la inversión y maximización del retorno de la inversión (ROI).

Este método, u otros, sirven para corroborar la idea de que el *Big Data* no es un método único sino una combinación de métodos que, utilizados sobre un volumen suficiente de datos, pueden ayudar a hacer más eficiente a una empresa.

BIG DATA, ANÁLISIS Y COMPETITIVIDAD

Uno de los aspectos más relevantes del análisis de *Big Data* es su capacidad para transformar los datos disponibles, a través de métodos estadísticos y computacionales, en información que es valiosa para generar una ventaja competitiva para la empresa y un valor añadido al cliente. Los distintos métodos que se utilizan, como los presentados en este artículo, sirven para realizar predicciones que permiten mejorar la toma de decisiones en un ámbito donde los clientes requieren una atención más personalizada y donde ésta sólo se puede ofrecer a través de las previsiones de su comportamiento (Lenka, 2016).

Como señalan Porte y Hepplemann (2014), el uso de los datos, el internet de las cosas, constituye una oportunidad para conseguir nuevas tecnologías que sirvan para potenciar el crecimiento empresarial y económico, centrándose no sólo en las reducciones de costes que han primado en las últimas décadas, especialmente en el ámbito salarial, sino a través de innovaciones en ganancias de productividad a través del *Big Data*. Estas innovaciones pueden modificar la tendencia de crecimiento, que en cualquier caso necesita de nuevas competencias técnicas, claras normas en la protección de datos y la mejora de la infraestructura tecnológica.

El cambio en la forma de toma de decisiones no es único ni lineal, sino que plantea multitud de combinaciones que pueden determinar la dirección de la empresa, permitiendo el ajuste de la información de la que dispone a sus necesidades, pasando a lo que Brynjolfsson y McElheran, (2016) denominan Decisiones Dirigidas por Datos (DDD). Las DDD pueden servir para obtener mejoras en el rendimiento de las empresas (Brynjolfsson, Hitt, and Kim, 2011), ya sea a través de cambios en la cadena de producción, en la estructuras de costes, o en cualquier otro ámbito donde las decisiones puedan optimizar por la existencia de flujos de información.

Dichas mejoras suponen cambios en la forma de comercializar los productos: cómo, a quién, cuándo y dónde dirigir los productos. Algunas de estas transformaciones, como la segmentación de clientes, y el aprendizaje automático o *machine learning* son las que se utilizan en esta investigación. Como señalan Adamson *et al* (2012), la forma de comercializar de forma masiva a los clientes está quedando obsoleta, ya que las nuevas estrategias de comercialización requieren conocer al cliente con suficiente profundidad para poder ofrecerle el producto mucho antes de que tenga la necesidad del mismo, ayudando a la construcción del proceso de transformación de su voluntad, no sólo recurriendo a grupos cada vez más pequeños, sino dirigiéndose directamente a los individuos como unidad de referencia objetivo.

En este análisis, son los propios individuos los que proveen la información necesaria a través de sus opiniones y su comportamiento, haciendo obsoletas e inefectivas encuestas y otros sistemas previos de recolección de datos, construyendo mediante sus acciones información sobre sus necesidades y sus deseos presentes y futuros. A través de las distintas técnicas, análisis de búsquedas, navegación, análisis semánticos con Inteligencia Artificial, etc., aparecen los propios deseos de los consumidores, más allá de los datos que se puedan determinar en grupos generales (Morabito, 2014). Así, se obtiene información limpia en cuanto al proceso de decisión de consumo, mientras que a través de la personalización se puede ir mejorando los servicios que se ofrecen.

En esta investigación se establece una relación entre *Big Data* y *Social Data* –es decir, aquellos datos obtenidos a través de las redes sociales–, pero sus aplicaciones van mucho más allá. Como plantean Porter y Happelmann (2014), el internet de las cosas permite recopilar información de forma continuada, tanto por la actuación como por la inacción del individuo, lo que permite la transformación de todo el procedimiento de toma de conocimiento. Por supuesto, su utilidad va mucho más allá de los aspectos comerciales y se añade a los distintos servicios que una persona pueda necesitar en su relación diaria con su comunidad, su empleo o incluso en sus relaciones personales.

Muchas empresas, incluyendo los gigantes de internet como Google o Amazon, están empleando este tipo de información. Pero aún estamos lejos de que su empleo de forma masiva, y todavía no se está utilizando de forma diferencial su poder como herramienta para incrementar la productividad, especialmente como instrumento predictivo (Brynjolfsson y McElheran, 2016). Para ello, los objetivos del análisis científico de datos pasan por (Morabito, 2015):

- Mejora de la toma de decisiones, facilitando en un mejor análisis con unos mejores datos, generando previsiones sobre las decisiones de los usuarios y sus tomas de decisiones.

- Mayor rendimiento de la empresa gracias a una mejora de los datos disponibles, lo que permite que surja una inteligencia empresarial capaz de coordinar mejor los recursos y combinarlos para que se tomen las decisiones del punto anterior.

Para conseguir estos objetivos la empresa tiene que realizar transformaciones que pasan por, adquirir competencias necesarias que no se han logrado todavía en la actualidad, como reconoce el informe EPYCE 2016 (EAE BS, 2017), ya que un 10,11% de las empresas fueron incapaces de cubrir las posiciones en ese ámbito en España en el año 2016. Esta dificultad no se produce sólo por la falta de profesionales con las características técnicas, sino también por la falta de la combinación de estas mismas con conocimientos sobre la transformación del negocio y su diseño futuro. El retorno potencial si se produjeran dichas transformaciones puede, sin embargo, ser muy elevado en cuanto a ganancias en la productividad, en el entorno del 20%-30% (Tambe y Hitt, 2013).

Los profesionales se enfrentan a dos retos todavía mayores: la falta de cultura del dato en muchas empresas y las dificultades para llevar a cabo las trasformaciones necesarias dentro de la compañía. El cambio necesita el compromiso firme de la empresa, y especialmente de su dirección, algo que no resulta fácil en un ámbito donde los nuevos perfiles modifican de forma significativa la relación entre áreas a menudo aisladas entre sí, como el diseño del negocio, las tecnologías de la información o el desarrollo del negocio. Para que el resultado sea positivo se necesita flexibilidad en la toma de decisiones, un deseo de cambio de la estructura y permeabilidad al mismo, con un planteamiento similar al primer gran cambio tecnológico (Acosta *et al*., 2006). Por último, es necesaria una inversión tecnológica que potencie las ganancias en productividad (Tambe, 2014).

UNA APLICACIÓN DE *BIG DATA* EN *E-COMMERCE*

Un ejemplo sería la aplicación a un modelo de *cashback*. El *cashback*, o incentivo por reembolso es, según el diccionario de Oxford, «una forma de incentivo ofrecida a los compradores de un cierto producto mientras que recibe un reembolso en efectivo tras realizar la compra.» En el caso del mercado *online* los consumidores pueden recibir beneficio económico mediante actividades como navegar por páginas web o por realizar transacciones a través de la web de *cashback*.

En 2015, el mercado global de *cashback* se estimaba en 84.000 millones de dólares, centrado principalmente en Estados Unidos y Europa, y los datos de sus participantes son impresionantes: el 64% de los consumidores online pertenecen a sistemas de gestión de lealtad y el 71% que realiza transacciones online querría participar en este tipo de programas. Como dato adicional, sus operaciones en el Reino Unido, uno de los países donde se utiliza más intensamente, suponen un 1% del PIB (Hall y Domansky, 2017).

En esta investigación utilizamos los datos desde julio de 2007 hasta marzo de 2015 de una de las empresas líderes en la Europa continental, que opera en 14 países, con más de 2 millones de clientes y capaz de generar ventas por más de 20 millones de euros anualmente. La base de datos, con más de 400 millones de registros, contiene información relativa a la actividad comercial desarrollada por los clientes en las tiendas que ofrecen sus productos y/o servicios en el portal, su interacción con la red social interna del propio del *e-commerce*, así como sus características sociodemográficas. Las actividades que los usuarios pueden realizar en el *e-commerce*, y que son susceptibles de generar *cashback*, consisten en clics y/o visitas hacia otros sitios *web*, registros en aplicaciones o portales de otras marcas, generación de *leads*, o venta de productos y/o servicios de la oferta de marcas disponibles en la plataforma. La muestra seleccionada para realizar el análisis empírico en este artículo recoge toda la actividad realizada por los usuarios en el *e-commerce* desde enero hasta marzo de 2015, en la que 12.548 clientes realizaron 687.682 transacciones. Esta información aglutina datos procedentes tanto de fuentes de datos estructuradas como no estructuradas, lo que lo enmarca dentro de un análisis de *Big Data*.

Ballestar *et al*. (2016a, 2016b y 2017) analizan académicamente por primera vez con detalle este mercado y presentan a través de metodologías de Ciencia de los Datos/*Big Data* los resultados de la segmentación del mercado, su rentabilidad y su caracterización. El análisis que hoy presentamos utiliza una metodología de *Machine Learning* consistente en una red neuronal artificial perceptrón multicapa (ANN-MLP) que valida el modelo de segmentación realizado en Ballestar *et al*. (2017). Este modelo de segmentación de clientes utiliza una metodología de análisis clúster en dos pasos (*two-step cluster analysis*) para configurar una clasificación de los usuarios del sitio *web* de *cashback* en función de su actividad comercial y el rol que desempeñan dentro de la red social interna del propio sitio.

La red neuronal artificial perceptrón multicapa ha sido entrenada utilizando la misma muestra de datos utilizada para construir el modelo de *clustering* en Ballestar *et al*. (2017), seleccionando como variables independientes o entrada para la red neuronal las variables utilizadas para crear el modelo de segmentación, y como variable dependiente o de salida el segmento de pertenencia del usuario calculado por la metodología de *clustering*. Esta metodología permite, por un lado, validar mediante el uso de otras metodologías *Data Science* la calidad del modelo de segmentación presentado en Ballestar *et al*. (2017) y, por otro, generar un modelo predictivo y ejecutable en tiempo real de clasificación de clientes. Este nuevo modelo predictivo permitirá la optimización de los esfuerzos de la compañía en la captación y fidelización de clientes en el sitio *web*.

Recolección de datos

A continuación, se describen cada una de las variables, tanto de entrada como de salida, que han sido utilizadas para para entrenar la red neuronal artificial perceptrón multicapa (ANN-MLP):

Variables de entrada de la red neuronal

Se utilizan seis variables de entrada para entrenar la ANN-MLP, las mismas variables utilizadas para construir el modelo de segmentación mediante la metodología *two-step cluster analysis*. Las variables de entrada son de dos tipos distintos, por un lado, tipología de transacciones que los usuarios pueden realizar en el sitio y, por otro, el rol que estos usuarios desempeñan en la red social interna del sitio *web*. Los clientes realizan en el sitio una media de 55 transacciones durante el periodo de observación. Estas transacciones pueden ser de cinco tipos distintos como se describe a continuación:

- Transacción tipo clic o visita: Estas transacciones no requieren desembolso económico por parte del usuario. Incluyen la visualización de videos, convertirse en fan de una marca en redes sociales, rellenar encuestas, etc. El 80,90% de los usuarios han realizado al menos una transacción de este tipo. Representa el 97,76% del total de transacciones y el 13,23% del *cashback* generado en la plataforma (0,02 euros por transacción). Esta información se almacena en una variable numérica (usc_n_op_direct_c).

- Transacción de tipo registro: Estas transacciones no requieren desembolso económico por parte del usuario. La transacción de registro ocurre cuando el usuario se abre una cuenta en un comercio afiliado a la plataforma de *cashback*. En la muestra el 9,69% de los usuarios ha realizado al menos una transacción de este tipo, representando el 0,15% total de transacciones y el 3,78% del *cashback* generado en la plataforma. (12,48 euros por transacción). Esta información se almacena en una variable numérica (usc_n_op_direct_r).

- Transacción de tipo compra de producto o servicio: Estas transacciones requieren desembolso económico por parte del usuario. La transacción consiste en la compra de productos o servicios en un comercio afiliado a la plataforma. En la muestra el 7,14% de los usuarios ha realizado una operación de este tipo, representando el 0,98% del total de transacciones y el 70,21% del *cashback* generado en la plataforma (9,44 euros por transacción). Esta información se almacena en una variable numérica (usc_n_op_direct_s).

- Transacción para convertirse en *lead*: Un usuario se convierte en *lead* cuando muestra interés por comprar un producto o servicio (como un préstamo, servicio de telefonía móvil, curso, etc.) en una de las tiendas afiliadas a la plataforma, proporciona sus datos personales, pero no termina de completar el proceso de adquisición en el momento. El usuario recibe el *cashback* una vez que activa el servicio y confirma la compra, por lo que requiere un desembolso económico a posteriori. En la muestra el 20.70% de los usuarios han realizado al menos una transacción de este tipo, acumulando el 1.50% de las transacciones y el 2.43% del *cashback* generado en la plataforma. (0.21 euros por transacción). Esta información se almacena en una variable numérica (usc_n_op_direct_o).

- Transacciones procesadas manualmente: Cuando las transacciones no son procesadas de forma automática en la plataforma, son procesadas de forma manual. Este tipo de transacciones suponen el 0,005% del total y generan el 0,35% del *cashback* total (10,11 euros por transacción). Esta información se almacena en una variable numérica (usc_n_op_direct_m).

Por otra parte, se incluye en el modelo la variable relativa al rol que el usuario desempeña en la red social interna del sitio *web* de *cashback*. Los usuarios tienen la capacidad de recomendar a otros usuarios a incorporarse a la red social aplicando una estrategia de marketing de *word-of-mouth*. De esta forma, los usuarios no solo reciben un incentivo en forma de *cashback* por cada transacción que realizan (sea del tipo que sea), sino que también reciben también incentivos por cada transacción de tipo clic o visita que realiza su red de recomendados (tanto recomendados de primer nivel o «hijos», como de segundo nivel o «nietos»). El tamaño medio de la red de recomendación es de 32,8 recomendados por usuario (13 recomendados de primer nivel y 19,8 de segundo nivel). Por este motivo, también es relevante para el modelo, el rol que el usuario desempeña en la red social del sitio de *cashback*.

- Rol en la red social: Variable categórica que contiene el rol que desempeña el usuario en la red social (rec_role_subtype_in_network). Existen seis roles diferentes y el usuario puede evolucionar a lo largo de ellos conforme también evoluciona su vinculación con el sitio *web*:

    - Rol 1: Usuario que no está vinculado a la red social. Representan el 13% de los usuarios.

    - Rol 2: Usuario que se unió a la red social proactivamente, sin recomendación por parte de otros usuarios, pero que ha desarrollado una red social de recomendados hasta el segundo nivel («hijos» y «nietos»). Representan el 4,2% de los usuarios.

    - Rol 3: Usuario que se unió a la red social proactivamente, sin recomendación por parte de otros usuarios, pero que ha desarrollado una red social de recomendados hasta el primer nivel («hijos»). Representan el 7,7% de los usuarios.

    - Rol 4: Usuario que se unió a la red social por recomendación, pero que no ha desarrollado una red social de recomendados. Representan el 36% de los usuarios.

    - Rol 5: Usuario que se unió a la red social por recomendación y que ha desarrollado su red social de recomendados hasta el segundo nivel («hijos» y «nietos»). Representan el 14,2% de los usuarios.

- Rol 6: Usuario que se unió a la red social por recomendación y que ha desarrollado su red social de recomendados hasta el primer nivel («hijos»). Representan el 24,8% de los usuarios.

Variable de salida de la red neuronal

La metodología de segmentación *two-step cluster analysis* agrupa a los 12.548 clientes en ocho segmentos, en función de las variables de entrada al modelo expuestas en la sección anterior. Esta información es almacenada en una variable categórica que servirá de variable de salida/output al modelo de red neuronal artificial perceptrón multicapa (ANN-MLP).

- Segmentos de compradores por conveniencia:

Consumidores caracterizados por que se unieron de forma proactiva a la red social del sitio de *cashback*, sin recomendación por parte de otros usuarios. Inicialmente, se incorporan al segmento 5 y cuando desarrollan su propia red de recomendados (tanto de primer, como segundo nivel) pueden evolucionar hasta el segmento 7.

El segmento 5 está conformado por el 11,4% de la muestra (1.435 usuarios). Estos usuarios tienen una antigüedad en la plataforma de 2,16 años y son el segmento menos transaccional del portfolio, con una media de 20 transacciones y 6,34€ de *cashback* por cliente. Una vez que estos usuarios aumentan su vinculación con el sitio y desarrollan su red social de recomendados, pueden realizar su transición al segmento 7. Este segmento está conformado por el 10,6% de la muestra (1.328 usuarios) y tienen una antigüedad de 3,36 años en la plataforma.

- Segmentos de usuarios recomendados con un nivel de actividad medio-bajo en el sitio:

Consumidores caracterizados porque se unieron a la red social del sitio mediante la recomendación de otro usuario ya vinculado a la red social. Estos usuarios se clasifican en los segmentos 1, 6 y 8 y, aunque tienen un nivel de actividad medio-bajo en el sitio, su vinculación y nivel de actividad aumenta conforme aumenta su antigüedad como usuarios en la plataforma.

En primer lugar, el segmento 1 está constituido por los usuarios recomendados aún muy inmaduros. Representan el 29.7% de la muestra (3,722 usuarios); aún no han tenido la oportunidad de desarrollar su propia red de recomendados y su antigüedad media es de solamente 2,70 años en la plataforma, siendo el segundo grupo más joven, sólo por detrás del segmento 5. Estos usuarios presentan una baja vinculación y son los menos rentables de la plataforma, con 20,9 transacciones y 4,22€ de *cashback* por cliente.

Conforme la vinculación de estos usuarios con la red social se va incrementando, creando su propia red de recomendados hasta el primer nivel («hijos») y aumentando su actividad en la plataforma, también pueden realizar su transición al segmento 6. Este segmento en desarrollo representa el 19,5% de la muestra (2.448 usuarios); su antigüedad media es de 3,47 años y alcanza una media de 25,2 transacciones. Aun así, su rentabilidad es baja con 4,81€ de *cashback* por cliente.

Así mismo, la evolución de este segmento en términos de desarrollo de su red social hasta un segundo nivel («hijos» y «nietos») así como número de transacciones realizadas en la plataforma, da lugar al segmento 8. Este segmento representa el 11,8% de la muestra (1.485 usuarios) y su antigüedad media se ha incrementado hasta los 4,12 años. El número de transacciones por cliente se sitúa en 39,3 y el *cashback* generado por cliente en 7.06€.

Cabe destacar que, conforme los usuarios evolucionan a lo largo de estos segmentos, no sólo aumenta el número de transacciones que realizan, sino también su variedad, incluyendo aquellas transacciones que requieren desembolso económico.

- Segmentos de usuarios con un nivel de actividad alto:

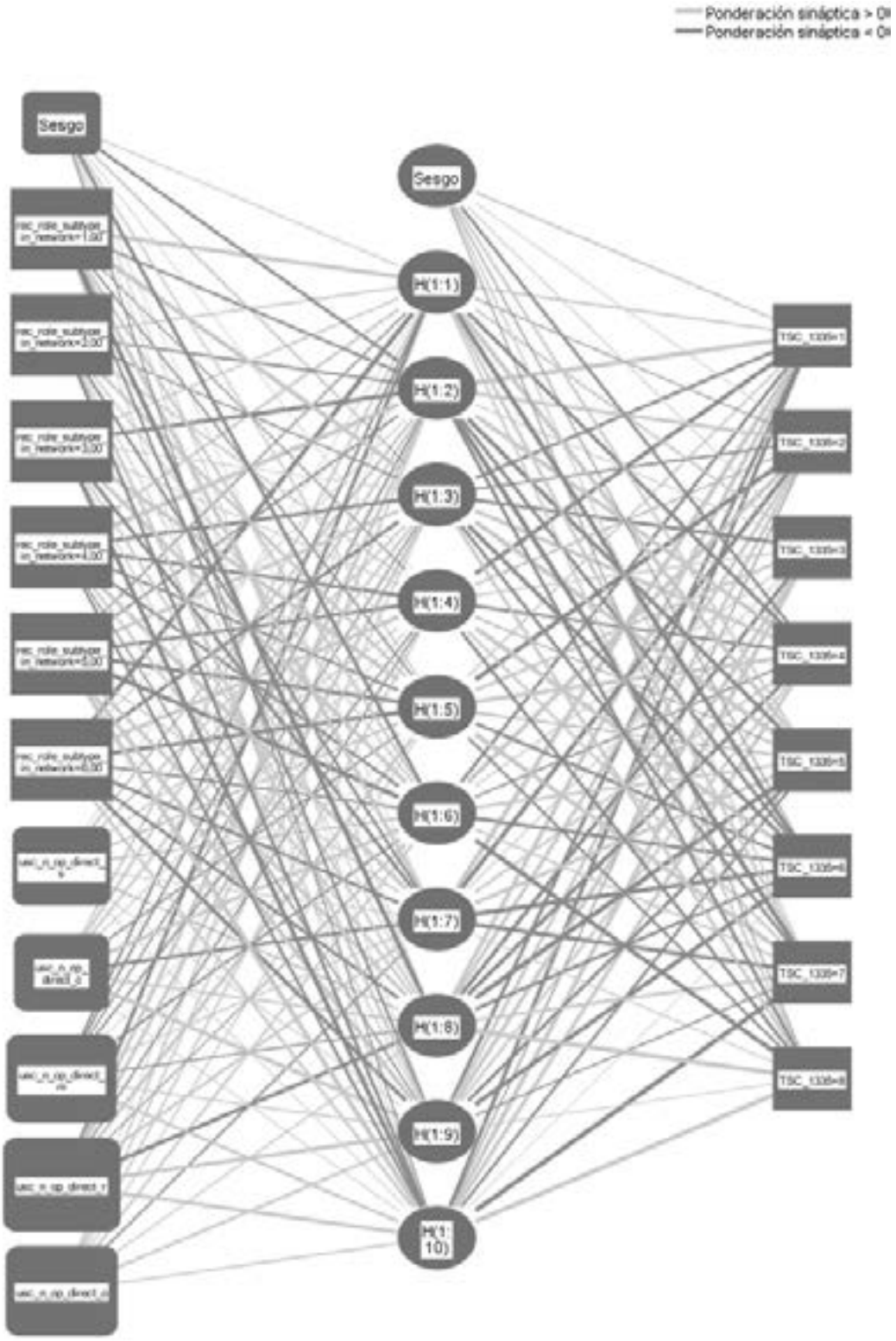
Consumidores caracterizados por tener una elevada actividad en el sitio y por ser una evolución de los segmentos 1 y 5. Por lo tanto, algunos de ellos se unieron a la red social del sitio por recomendación, mientras que otros lo hicieron de forma proactiva. Estos usuarios se clasifican en los segmentos 2, 3 y 4.

El segmento 2 corresponde a usuarios muy activos y que invierten mucho tiempo en el sitio, pero cuya rentabilidad es baja. Representan el 9,9% de la muestra (1.241 usuarios) y tienen una antigüedad media de 3,45 años. La mayor parte de estos usuarios mantiene algún tipo de relación con otros miembros de la red social (del segmento, sólo un 6% carece por completo de ellas). Es el segmento de usuarios más activo en número de transacciones, con 278,2 transacciones por usuario, pero las actividades que realizan están muy concentradas en aquellas que no requieren desembolso económico y, por tanto, el *cashback* que generan es tan solo de 7,43€ por cliente. Son usuarios que disfrutan realizando actividades en el sitio y desean obtener beneficio económico por ellas, pero sin realizar desembolsos económicos.

El segmento 3 corresponde a usuarios también muy activos (realizan una media de 182,4 transacciones por cliente); sin embargo, su actividad no está concentrada en una tipología de transacción en concreto, sino que realizan todo tipo de transacciones, incluidas aquellas que requieren un desembolso económico. Por lo tanto, este segmento es muy poco sensible al precio y resultan los más rentables del sitio, generando 127,97€ de *cashback* por cliente. Sin embargo, este segmento es el más pequeño, representando al 0,6% de la muestra (72 clientes), con una antigüedad media de 3,08 años.

Por último, el segmento 4 corresponde a usuarios que han evolucionado muy rápidamente procedentes de los segmentos 1 y 5 y tienen una antigüedad media de 2,7 años. Su nivel de actividad en la plataforma es media-alta, ocupando la posición de tercer grupo más ac-

FIGURA 1
ARQUITECTURA DE LA ANN MLP



Función de activación de capa oculta: Tangente hiperbólica

Función de activación de capa de salida: Softmax

Fuente: Elaboración propia

FIGURA 2
MATRIZ DE CONFUSIÓN

Clasificación

Muestra	Observado	Pronosticado 1	2	3	4	5	6	7	8	Porcentaje correcto
Entrenamiento	1	2595	2	0	0	0	0	0	0	99,9%
	2	15	858	0	2	1	5	0	1	97,3%
	3	0	17	26	9	0	0	0	1	49,1%
	4	0	0	0	569	0	0	0	0	100,0%
	5	0	1	0	0	1012	0	0	0	99,9%
	6	0	6	0	0	0	1704	0	0	99,6%
	7	0	1	0	0	0	0	926	0	99,9%
	8	0	7	0	0	0	0	0	1046	99,3%
	Porcentaje global	29,6%	10,1%	0,3%	6,6%	11,5%	19,4%	10,5%	11,9%	99,2%
Pruebas	1	1124	1	0	0	0	0	0	0	99,9%
	2	11	341	2	2	1	0	0	2	95,0%
	3	0	8	9	2	0	0	0	0	47,4%
	4	0	0	0	248	0	0	0	0	100,0%
	5	0	3	0	0	419	0	0	0	99,3%
	6	0	3	0	0	0	735	0	0	99,6%
	7	0	0	0	0	0	0	401	0	100,0%
	8	0	3	0	0	0	0	0	429	99,3%
	Porcentaje global	30,3%	9,6%	0,3%	6,7%	11,2%	19,6%	10,7%	11,5%	99,0%

Variable dependiente: TwoStep Cluster Number

Fuente: Elaboración propia

tivo en la plataforma, con 53,4 transacciones por usuario. Además, con una media de 23,56€ de *cashback* generado por usuario, son el segmento más eficiente a la hora de operar en la plataforma.

Modelo teórico

Las redes neuronales son modelos matemáticos que analizan las relaciones complejas, incluso las no lineales, entre las variables de entrada al modelo y las variables de salida. Las variables de entrada se corresponden con las variables independientes, mientras que las de salida, son las dependientes.

Esta investigación utiliza una red neuronal artificial perceptrón multicapa (ANN-MLP) que aplica una técnica de aprendizaje de *backpropagation* (propagación hacia atrás), el cual pretende minimizar el *cross-entropy* error entre los valores reales y los predichos. Existe una gran variedad de redes neuronales artificiales, pero esta investigación utiliza perceptrón multicapa, una de las más populares. El perceptrón multicapa es un método supervisado, siendo este uno de sus principales inconvenientes, ya que la calidad y capacidad de predicción de la red neuronal estará condicionada por la calidad de la muestra (Hu *et al*., 2009; Li y Eastman, 2006).

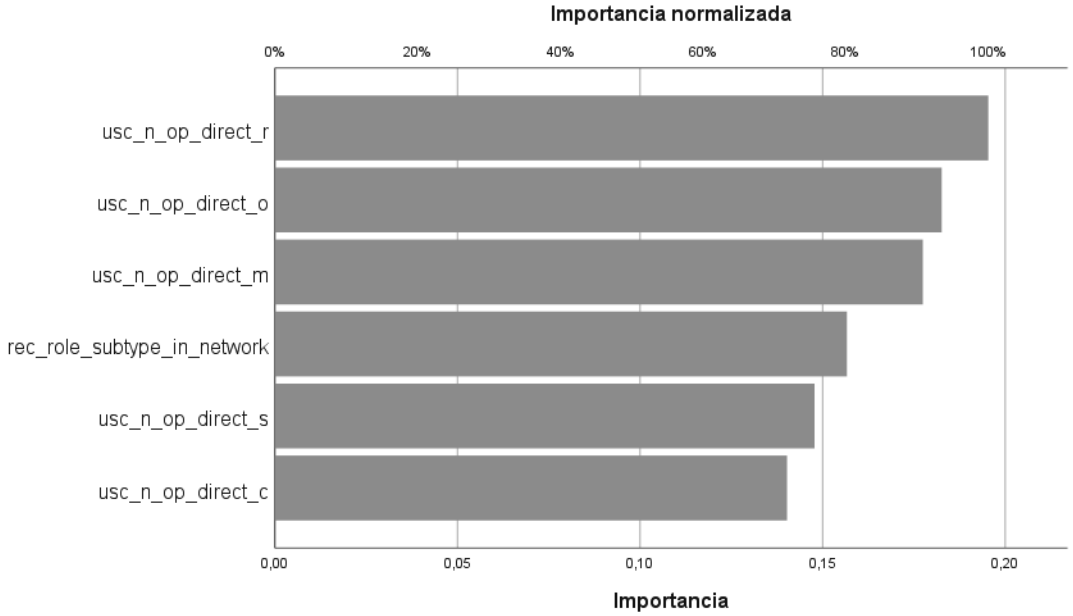
En cuanto a su estructura, esta red neuronal consta de tres capas, una capa de entrada, una oculta y otra de salida, interconectadas entre sí en una sola dirección por los pesos sinápticos. El modelo mapea las once unidades de entrada que reciben los valores de las seis variables de entrada o independientes, con las ocho unidades de salida correspondiente a la variable dependiente o de salida que contiene el segmento de pertenencia de los usuarios en el sitio. La Figura 1 muestra la arquitectura de la red neuronal e indica que la tipología de función de activación de la capa oculta y la capa de salida corresponden con una tangente hiperbólica y *softmax*, respectivamente.

Análisis empírico y resultados

El proceso de entrenamiento de la red se realiza sobre el 70,2% (8.804 usuarios) de la muestra seleccionado aleatoriamente, mientras que la validación o testeo se realiza sobre el restante 29,8% (3.444 usuarios). La precisión de la clasificación y el área debajo de la ROC curva son los indicadores más relevantes para evaluar la precisión/capacidad predictiva de la red neuronal. Estos indicadores también determinaran la medida en la que las predicciones realizadas por la ANN-MLP coinciden con la clasificación realizada por la metodología de segmentación *two-step cluster analysis*.

La precisión de la clasificación de la red neuronal es del 99,2% (una tasa de error del 0,8%). Por lo tanto, la red neuronal coincide con la clasificación realizada por la segmentación *two-step cluster* análisis en el 99,2% de los casos, implicando una elevada convergencia entre los resultados de ambas metodologías. En la Figura 2 se muestra la matriz de confusión, que contiene el porcentaje de casos clasificados correctamente sobre el total de población y también para cada uno de los segmentos de usuarios. Tanto para la muestra de entrenamiento como la de test (pruebas) estos porcentajes son muy similares, lo que descarta el sobreentrenamiento de la red neuronal.

FIGURA 3
IMPORTANCIA E IMPORTANCIA NORMALIZADA DE LAS VARIABLES DE ENTRADA EN LA ANN-MLP



Fuente: Elaboración propia

El área debajo de la ROC curva es un indicador de capacidad de clasificación de la red neuronal aún más robusto que el indicador de precisión en la clasificación expuesto anteriormente. Este indicador se ha calculado para cada uno de los ocho segmentos de usuarios en el sitio, obteniendo valores que oscilan entre 0,994 y 1. Estos valores corroboran que la capacidad predictiva del modelo es muy elevada y coincidente con el mismo output generado por el modelo de segmentación *two-step cluster analysis*. (Hosmer & Lemeshow, 2000).

Por último, la Figura 3 muestra la importancia e importancia normalizada de cada una de las seis variables de entrada o independientes en la estimación del modelo ANN-MLP. La suma de la importancia relativa de las variables es 1, pero estos valores no tienen ninguna relación con la precisión del modelo. Esto significa que la importancia relativa tan sólo proporciona información sobre la relevancia que cada variable tiene cuando la red neuronal realiza una predicción, sea esta precisa o no. En este caso, las tres variables con más relevancia son aquellas relativas a las transacciones de registro, conversión del usuario a Lead y las transacciones procesadas manualmente.

CONCLUSIONES

Este artículo presenta la importancia que, más allá de la burbuja informativa, tiene el denominado *Big Data* para los negocios digitales. El uso masivo de información de los clientes por parte de las empresas para ofrecer un mejor servicio a sus usuarios. Esto debe, a su vez, servir para incrementar la rentabilidad del mismo a través de una mayor lealtad, más transacciones y mayores márgenes.

La ciencia de los datos permite a las empresas tener a su disposición la suficiente información para mejorar su posición en el mercado. A cambio, necesitan dominar una tecnología que se está desarrollando a marchas forzadas y que combina no sólo conocimientos de estadística y de matemáticas, sino también de negocio y de estrategia empresarial.

Este objetivo se ilustra con un ejemplo que, a través del uso de redes neuronales dentro del marco de *machine learning*, permite verificar los resultados de una «clusterización» donde los datos obtenidos sirven para determinar cuál es el momento de madurez del cliente y por lo tanto cómo hay que enfocar el trato de la empresa en cuánto a su relación con el cliente para así optimizar esa relación.

No se contempla en este artículo aspectos que han empezado a llamar la atención de la sociedad, como la propiedad y el uso de los datos, que representan un fuerte componente ético del *Big Data* pero que, como tantas veces, va detrás de su uso y que representa una limitación clara que sin embargo no empaña la utilidad de la tecnología para el desarrollo de la empresa y la mejora de su gestión.

NOTAS

[1] Como señalan Gandomi y Heider (2015), la diferencia primordial entre *Big Data* y *Data Science* se centra, fundamentalmente en el volumen de los datos, que es lo que caracteriza al proceso y lo que puede complicarlo frente a los métodos tradicionales. En aras de la sencillez, aquí se utilizarán de forma indistinta.

BIBLIOGRAFÍA

Acosta, M., Sainz, J. y Salvador, B. (2006). «Hago click y opero a tu lado: Estrategia de la banca online en España». *Cuadernos de Gestión* 6 (1): 101-110.

Adamson, B., Dixon, M., y Toman, N.: (2012) The end of solution sales. Harvard Bussines Review, 90, 60–70.

Ballestar, M. T., Grau-Carles, P., y Sainz, J. (2016). Consumer behavior on cashback websites: Network strategies. Journal of Business Research, 69(6), 2101-2107.

Ballestar, M. T., Grau-Carles, P., y Sainz, J. (2017). Customer segmentation in e-commerce: Applications to the cashback business model. Journal of Business Research. https://doi.org/10.1016/j.jbusres.2017.11.047

Ballestar, M. T., Sainz, J., y Torrent-Sellens, J. (2016). Social networks on cashback websites. Psychology & Marketing, 33(12), 1039-1045.

Brynjolfsson, E. Hitt, L. y Kim H. (2011). «Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance?» SSRN working paper. Available at SSRN: http://ssrn.com/abstract = 1819486.

Brynjolfsson, E., y McElheran, K. (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5), 133-39.

Capilla, A. y Jorge Sainz, J. (2009) Cuadernos de Pensamiento Político No. 22, pp. 139-156.

Chatterjee, P. (2013). Big data: the greater good or invasion of privacy? http://www.guardian.co.uk/commentisfree/2013/mar/12/bigdata-greater-good-privacy-invasion.

Diebold, F. X. (2003). Big data dynamic factor models for macroeconomic measurement and forecasting. In Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,» (edited by M. Dewatripont, LP Hansen and S. Turnovsky) (pp. 115-122).

Diebold, F. X. (2012). On the Origin (s) and Development of the Term 'Big Data'. PIER Working Paper 12-037, University of Pennsylvania.

Dong, X. L., y Srivastava, D. (2013, April). Big data integration. In Data Engineering (ICDE), 2013 IEEE 29th International Conference on (pp. 1245-1248). IEEE.

Demchenko, Y., De Laat, C., y Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In Collaboration Technologies and Systems (CTS), 2014 International Conference on (pp. 104-112). IEEE.

EAE Business School, (2017) «Informe sobre Posiciones y Competencias más Demandas», Observatorio Permanente de Perfiles Profesionales Multisectoriales, EPYCE.

Gandomi, A., y Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

Hall, D. y Domansky, J. (2017) 2015 Cashback Industry Report; A Global Industry comes of age, CII.

Hu, X. and Weng, Q. (2009) Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks, Remote Sensing of Environment, Volume 113, Issue 10, 2009, Pages 2089-2102.

Jin, X., Wah, B. W., Cheng, X., y Wang, Y. (2015). Significance and challenges of big data research. Big Data Research, 2(2), 59-64.

Lenka, S., Parida, V., Rönnberg Sjödin, D., y Wincent, J. (2016). Digitalization and advanced service innovation: How digitalization capabilities enable companies to co-create value with customers. Management of Innovation and Technology, (3), 3-5.

Li, Z., and Eastman, J. (2006). Commitment and typicality measurements for the selforganizing map. Proceedings of SPIE The International Society for Optical Engineering, Bellingham.

Lohr. S. (2013 a) Big Data Is Opening Doors, but Maybe Too Many. https://www.nytimes.com/2013/03/24/technology/bigdata-and-a-renewed-debate-overprivacy.html?pagewanted=all& r=0.

Lohr, S. (2013 b). The origins of 'Big Data': An etymological detective story. https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

Microsoft (2013) The Big Bang: How the Big Data Explosion Is Changing the World - Microsoft UK Enterprise Insights Blog - Site Home - MSDN Blogs. http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/04/15/ big-bang-how-the-big-data-explosion-is-changing-theworld.aspx.

McAfee, A., y Brynjolfsson, E.: (2012) Big data: The management revolution. *Harvard Business Review*, 61–68.

Morabito, V. (2014) Big data. Trends and Challenges in Digital Business Innovation, Springer, London.

Morabito, V. (2015) Big Data and Analytics, Strategic and Organizational Impacts, Springer, London.

Porter, M. E., y Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard Business Review*, 64-88.

Pospiech, M. y Felden, C.: (2012) Big data—A State-of-the-Art. AMCIS 2012.

Provost, F., y Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60(6), 1452-1469.

Tambe, P.,y Hitt, L. M. (2013). Job hopping, information technology spillovers, and productivity growth. Management Science, 60(2), 338-355.