EL IMPACTO DEL *BIG DATA* EN LOS SERVICIOS FINANCIEROS

José GARCÍA MONTALVO (*)

Catedrático de Economía e investigador ICREA-Academia

Resumen

La generalización del *big data* y las nuevas técnicas asociadas al tratamiento y análisis de grandes bases de datos está revolucionando tanto el trabajo científico como la gestión empresarial. Aplicaciones como las recomendaciones personalizadas de Amazon han supuesto una mejora muy significativa de la experiencia de compra de los consumidores. En este trabajo se analizan las posibilidades del *big data* para mejorar los servicios financieros y la experiencia de los clientes y aumentar la eficiencia de las entidades en un contexto de presión sobre la rentabilidad de los bancos y la necesidad de recuperar la reputación del sector. La utilización de técnicas de *big data*, incluidos indicadores de reputación y capital social *online*, se ha extendido a la calificación crediticia de los solicitantes de crédito, la detección del fraude en tarjetas, la microsegmentación, los servicios de información a los clientes, el cumplimiento normativo y la prevención de blanqueo de capitales y operaciones de financiación de actividades terroristas, entre otras muchas actividades del sector. El trabajo también analiza algunos aspectos asociados a la utilización de *big data* como las cuestiones de privacidad de los datos y el cumplimiento de las regulaciones sobre utilización de la información.

Palabras clave: big data, *machine learning*, calificación crediticia, fraude, microsegmentación, privacidad.

Abstract

The generalization of «big data» and new techniques brought by the management and analysis of large datasets is changing drastically science and business management. Applications as the recommendation tools in Amazon, have provided consumers with a significant improvement in their shopping experience. In this paper I analyze the potentiality of «big data» to transform and improve the efficiency of the financial services industry and the business experience of costumers in the context of increasing pressure on the return of banks and the need to recover the reputation lost during the long financial crisis. The use of «big data», which currently include also indicators of online reputation and online social capital of the clients, covers credit scoring, credit card fraud detection, micro-segmentation, clients' information systems, compliance with regulatory requirements, and many other activities in the banking industry. The last section of the paper discusses issues of privacy and compliance with regulation on equal opportunity access to credit in a new business environment dominated by «big data».

Key words: big data, machine learning, credit score, credit card fraud, micro-segmentation, privacy.

JEL classification: C52, C53, C55, C80, G21, G23.

I. INTRODUCCIÓN

ESTE trabajo intenta dar una visión necesariamente parcial, por la enorme rapidez del cambio y la extensión del tema, sobre las posibilidades que ofrecen las técnicas de *big data* en el campo de los servicios financieros. Ciertamente el *big data* no es un campo nuevo. Bancos, comercios, etc. han utilizado durante mucho tiempo grandes cantidades de datos para, por ejemplo, estudios de segmentación de sus clientes. Los científicos hace mucho más tiempo que utilizan enormes cantidades de datos (1). De hecho la información que se genera en las redes sociales o en Internet no es la fuente principal de la ingente acumulación de datos que se produce cada día. Aplicaciones científicas como los radiotelescopios, los secuenciadores de ADN o los aceleradores de partículas generan mucha más información. Por ejemplo el LHC (Large Hadron Collider) del CERN produce un PetaByte de datos cada segundo (2) a partir de los sensores que captan todas las emisiones producidas por las colisiones de partículas (3). Desde un punto de vista tecnológico, en estos momentos el problema no son los datos sino que la velocidad a la que podemos almacenar la información no avanza al mismo ritmo que la información que se genera (4).

Las aportaciones de Internet (búsqueda y su aplicación a traductores, predicciones, etc.) y la expansión de las redes sociales ha tenido una gran repercusión en lo que podríamos denominar como el «*big data* social» (análisis de sentimientos, análisis textual, etc.). Pero, de hecho, la mayor parte de los datos que se generan en el planeta se producen por la interacción de máquinas (por ejemplo sensores) con otras máquinas. En estos momentos existe gran expectación sobre las posibilidades casi ilimitadas del *big data* para mejorar la productividad y el bienestar de las personas. Ante esta perspectiva es muy frecuente que en reuniones empresariales aparezcan alusiones a *big data*. La descripción de la situación actual ha sido caracterizada de forma irónica por Dan Ariely: «*Big data* es como el sexo adolescente: todos hablan de ello, nadie realmente sabe cómo hacerlo, todo el mundo piensa que todos los demás lo están haciendo y, por tanto, todo el mundo asegura que ellos también lo hacen». En el caso español se estima que solo el 3 por 100 de las empresas ha realizado algún proyecto, piloto o prueba en la dirección de aplicar técnicas de *big data*.

Este trabajo analiza cómo la utilización de *big data* está transformando los servicios bancarios y afectando a la cadena de creación de valor en el sector financiero. En el sector financiero uno de los materiales básicos es la información pues para realizar un adecuado análisis del riesgo es preciso un elevado componente informacional. De hecho esta es una de sus grandes ventajas competitivas: la gran cantidad de

datos que tienen de sus clientes. Por esto resulta paradójico, como se describe con posterioridad, que la banca lleve tanto retraso frente a otros sectores en la utilización y aprovechamiento de «*big data*». El «legacy» de unos equipos informáticos pensados para otra época dominada por *mainframes* y *software* poco flexible y cerrado no puede ser suficiente justificación de este retraso.

El objetivo no es tratar con detalle todos los cambios que Internet está produciendo en el negocio bancario (5), sino centrarse en las mejoras de eficiencia que puede producir la utilización de *big data* en el sector bancario así como la competencia por parte de productores de información masiva por la entrada en partes de la cadena de valor en los productos financieros (6). Por este motivo no se tratan negocios competidores de la banca que, aunque tienen su origen en Internet, atacan solo la parte de la cadena de valor que tiene que ver con los medios de pagos (Paypal, Square, TransferWise) ni monedas criptográficas, ni banca por móvil para usuarios sin acceso a la banca tradicional como M-Pesa y sus extensiones. En la sección II se realiza una aproximación al concepto de «*big data*», las técnicas asociadas a dichos desarrollos y las aplicaciones recientes en el campo de la economía. En la sección III se analiza el potencial del *big data* en los servicios financieros y se discuten aplicaciones a la calificación crediticia, el fraude en tarjetas de crédito, la microsegmentación, etcétera. La sección IV analiza los peligros asociados a la generalización del *big data* con especial énfasis en los temas de privacidad y consentimiento. La sección V incluye las conclusiones.

II. *BIG DATA:* CONCEPTO, TÉCNICAS Y APLICACIONES A LA ECONOMÍA

1. Una aproximación al concepto de *big data*

En la actualidad muchas líneas de investigación en ciencias (computacionales, sociales, etc.) y muchos nuevos negocios hacen referencia a la utilización de *big data*. La definición de *big data* es compleja pues muchas de las técnicas que acompañan a estas bases de datos masivos eran conocidas por otros nombres como «*data mining*». Por «*big data*» se entiende normalmente la construcción, organización y utilización de enormes cantidades de datos para extraer relaciones o crear nuevas formas de valor en mercados, organizaciones, servicios públicos, etc. (7). Esta definición debe matizarse para entender mejor la importancia del *big data*, dado que el dominio de estas técnicas no está solamente relacionado con el tamaño de las bases de datos que utilizan.

En primer lugar, cuando se piensa en *big data* se piensa en un ingente volumen de información.

En segundo lugar, una aplicación de *big data* implica la agregación de información de diversas fuentes, lo que hace particularmente importante el proceso de gestión y fusión de los datos. Los datos pueden provenir de sensores, el GPS de millones de teléfonos, clicks, *logs* de servidores, correos electrónicos, etcétera (8). Por tanto no se trata de datos numéricos organizados de manera estándar (por ejemplo en tablas). Los datos son muy heterogéneos y pueden incluir imágenes, textos, sonidos, etcétera. Asimismo su organización, por la propia heterogeneidad de los datos, impide que los datos almacenados tengan estructuras fijas como las bases relacionales clásicas. La gestión de los datos se realiza mediante sistemas NoSQL (9) (no solo SQL) por contraposición al tradicional lenguaje de consultas SQL. Esta herramienta es fundamental cuando se trabaja con muchos GigaBytes de datos o millones de observaciones con formatos heterogéneos y cuya estructura puede cambiar en el tiempo, por lo que necesita ser fácilmente escalable. Algunas de las herramientas utilizadas para manipular *big data* se están convirtiendo en el estándar del sector, como Hadoop, MapReduce, Pig, None, OpenRefine, Hive, HBase, Mahout, ZooKeeper o Impala. La gran mayoría de estas herramientas tiene como objetivo permitir el procesamiento en paralelo necesario cuando se trabaja con enormes bases de datos.

En tercer lugar, la información utilizada suele tener un nivel muy heterogéneo de ratio señal-ruido aunque, por lo general, existe bastante más ruido que en las aplicaciones habituales que utilizan datos administrativos, encuestas o sistemas de información internos de organizaciones empresariales.

En cuarto lugar, el objetivo de las técnicas de *big data* en general no es descubrir causalidades sino producir modelos predictivos. Por contraposición con la visión fundamental que se explica en los cursos de estadística y econometría tradicionales, en *big data* solo importan las correlaciones mientras la causalidad resulta irrelevante (10). Finalmente, la utilización y análisis de la información tiende a producirse a muy alta velocidad.

PAPELES DE ECONOMÍA ESPAÑOLA, N.º EXTRAORDINARIO, 2014. ISSN: 0210-9107. «NUEVOS NEGOCIOS BANCARIOS»

Mayer-Schönberger y Cukier (2013) describen *big data* como un cambio de mentalidad que supone la capacidad para analizar cantidades ingentes de datos en lugar de recurrir a muestras, la aceptación de la «suciedad» o inexactitud de los datos del mundo real como algo consustancial en lugar de buscar la exactitud en los datos, y el «creciente respecto a la correlación en lugar de la continua búsqueda de una elusiva causalidad».

Por tanto, un proyecto de *big data* utiliza principios de informática, matemáticas y estadística desarrollándose en tres etapas: 1) La primera es el proceso de captura y manipulación de datos. La enorme cantidad de información que hay que manipular implica la utilización de procesos en paralelo, programas específicos de reducción de dimensión (por ejemplo MapReduce) y gestores NoSQL. 2) La segunda etapa implica el análisis de los datos para encontrar relaciones predictivas útiles. En esta fase se utilizan técnicas estadísticas y *machine learning* (11). La diferencia entre ambos campos es difusa. Breiman (2001) señala que mientras en estadística se impone un modelo (regresión, logístico, etc.) para tratar de captar la naturaleza de la relación entre un *input* y un *output*, en *machine learning* el objetivo es encontrar una función que pueda predecir un resultado a partir de unos *inputs* sin requerirse ningún modelo sobre la naturaleza de dicha relación. 3) En una tercera etapa, tan importante como las anteriores, se utilizan técnicas de visualización para presentar los resultados y comunicarlos a los usuarios finales.

Un aspecto importante de *big data* es la llamada «reutilización» de los datos. Con *big data* el valor de los datos cambia en función de su uso (primario o uso potencial en el futuro). Esta perspectiva cambia la concepción que los negocios tienen sobre el valor de sus datos y presiona a las compañías a cambiar sus modelos de negocio y pensar constantemente en cómo usar los datos que tienen a su disposición de la forma más eficiente posible. En el pasado los datos generados por un negocio era un resultado colateral de la gestión del negocio. Cada vez más los datos son considerados valiosos por ellos mismos y su consecución y almacenamiento se convierte en parte importante del *core* del modelo de negocio.

2. *Big data* y técnicas estadísticas

> «We are drowning in information and starving for knowledge».
>
> JOHN NAISBITT

La disponibilidad de bases de datos masivas precisa, con más relevancia que en otros contextos, la utilización de técnicas analíticas para extraer conocimiento a partir de ellas. De forma similar a las dificultades que existen para seleccionar la información relevante o más precisa en el contexto de un caudal de información creciente de manera exponencial a partir de la expansión de Internet, la disponibilidad de TeraBytes de información no garantiza su utilidad a no ser que existan técnicas adecuadas para extraer conocimiento.

Aunque existe una evidente relación entre las técnicas de la estadística tradicional y los métodos de *machine learning* aplicados en el campo de *big data*, el énfasis en métodos específicos es muy diferente. Por ejemplo, en el caso de *supervised learning*, aquel que intenta aprender la relación entre un *input* y un resultado, *big data* utiliza métodos de clasificación, árboles de decisión y redes neuronales, mientras que un económetra tradicional usaría regresiones. En el campo del *unsupervised learning*, que solo utiliza *inputs*, un económetra utilizaría modelos no paramétricos para la estimación de una función de densidad, mientras en *machine learning* predominan los métodos de reducción de dimensionalidad.

La visión sobre cómo evaluar la bondad del ajuste de los procedimientos empleados también es muy diferente en la visión econométrica tradicional y la visión más orientada a *big data*. La elevada dimensionalidad de los datos generados en aplicaciones de *big data* produce algo que en la estadística clásica está ausente por los supuestos del modelo, por ejemplo, de regresión: que existan muchas más variables que observaciones disponibles ($K >> N$). Por este motivo los métodos de reducción de dimensionalidad son tan importantes, pero además la evaluación de los modelos tiene que ser necesariamente diferente. Por ejemplo, en regresión si utilizamos tantas variables como observaciones ($K = N$) conseguimos un ajuste perfecto dentro de la muestra. Este es el criterio, ajuste dentro de la muestra, que predomina en la econometría tradicional. Sin embargo el ajuste fuera de la muestra de ese modelo (suponiendo que se han generado nuevas observaciones o que se han guardado observaciones para hacer *cross-validation*) puede ser muy deficiente. Este problema se conoce en estadística como «sobreajuste» (*over-fit*) y es un tema básico en cualquier aplicación que utiliza *big data*.

Muchas de las técnicas habituales en *machine learning* y resolución de problemas de sobreajuste están teniendo cada vez mayor aplicación en el campo de la econometría. Belloni *et al.* (2014) presentan un análisis de métodos para datos de elevada dimensionalidad como es el caso de las aplicaciones de *big data*. La enorme dimensionalidad de estos problemas no solo se concentra en datos donde hay miles de variables sino también en selección de modelos entre miles de posibles especificaciones. La clave en estos métodos es el concepto de regularización o reducción de la dimensionalidad. Cuando existen potencialmente miles de variables que podrían incluirse en una regresión pero solo unas cuantas son relevantes para la predicción, el problema es seleccionar el modelo que genera mejor ajuste fuera de la muestra. En este contexto los procedimientos utilizados para realizar la selección se basan en LASSO (Least Absolute Shinkage and Selection Operator) en el que la estimación de los parámetros se produce a partir de una minimización del cuadrado de los residuos ampliada con una penalización que depende del tamaño del modelo (suma del valor absoluto de sus coeficientes). Belloni *et al.* (2014) muestran cómo se puede realizar inferencia utilizando LASSO para seleccionar variables instrumentales.

Ciertamente el énfasis de las técnicas de *machine learning* en la predicción frente al análisis de causalidad choca con la tendencia observada en los últimos veinte años, en especial en la economía, por generar situaciones experimentales, o pseudoexperimentales, para poder establecer nexos causales bien identificados. Asimismo, los proyectos de *big data* basados únicamente en aspectos predictivos están sujetos a una suerte de crítica de Lucas: la realización de un desarrollo cambia la realidad y hace el modelo inservible o poco útil. Por ejemplo, un método de detección de fraude basado en el testeo que hacen los defraudadores de la validez de la tarjeta, para ver si el fraude ha sido detectado, con pequeñas compras en tiendas *online* (por ejemplo Apps de 99 céntimos en iTunes) supone un cambio del procedimiento de los defraudadores una vez el método deja de funcionar. Por este motivo muchos de estos modelos precisan de actualizaciones frecuentes para mantener o mejorar su capacidad predictiva.

3. Aplicaciones de *big data* a la economía

Las aplicaciones de *big data* al campo de la economía son cada vez más abundantes aunque el despliegue de esta metodología en la economía sea reciente. Cada vez es más común encontrar metodologías que hacen estimaciones en tiempo real de la evolución de los precios o el gasto de los consumidores. El Billion Prices Project de MIT utiliza millones de precios de tiendas en Internet en decenas de países para obtener un índice de precios *online* que se actualiza en tiempo real (12). Esta tecnología utiliza la estabilidad o cambio de los componentes entre *tags* del lenguaje HTML utilizado para construir las páginas web para determinar los cambios en precios de los productos en el tiempo. Un programa puede, utilizando estos principios, identificar la información relevante sobre un producto y su precio. El URL de la página donde están indexados estos productos puede servir para clasificar los productos por categorías. Por ejemplo, Cavallo (2012) utiliza cientos de miles de precios de productos en Internet para comparar la evolución de la inflación oficial y la obtenida a partir de capturas de información sobre precios de tiendas *online*. Cavallo (2012) muestra que mientras en Brasil, Chile, Colombia o Venezuela la evolución de la inflación oficial y la obtenida a partir de los precios *online* siguen patrones similares, en Argentina las diferencias son muy significativas. En media la inflación en Argentina entre 2007 y 2011 definida por el índice de precios *online* fue del 20,14 por 100 frente a la inflación oficial que era tan solo del 8,38 por 100. Esto implica una diferencia acumulada del 65 por 100 en marzo de 2011.

Choi y Varian (2012) utilizan Google Trends para mejorar la capacidad predictiva de modelos sobre indicadores económicos obtenidos con muy alta frecuencia. La idea consiste en complementar la información del pasado de una serie con las búsquedas presentes en algunas categorías. Por ejemplo, el Departamento de Trabajo de Estados Unidos anuncia cada jueves el número de personas que han solicitado subsidios por desempleo. Añadiendo a un modelo AR (1) de datos históricos la información sobre búsquedas de palabras en categorías como *Jobs, Welfare, Unemployment,* se mejora un 6 por 100 la capacidad predictiva en general, y de los cambios de ciclo en particular. Utilizando el mismo sistema para el índice de confianza del consumidor se consigue una mejora del 9,3 por 100 en la capacidad predictiva.

La utilización de la información agregada sobre tarjetas de crédito y TPV es otra fuente importante de investigación económica en la actualidad. En una

serie de artículos que han resultado muy influyentes, Mian y Sufi (2009) han utilizado la información sobre tarjetas de crédito para realizar análisis económico sobre las causas de la burbuja inmobiliaria y la crisis financiera. Mian, Rao y Sufi (2013) analizan la elasticidad del consumo respecto a la riqueza inmobiliaria. El cálculo del consumo a nivel de condados se realiza utilizando las compras realizadas con tarjeta de crédito o débito intermediadas por MasterCard. Una ventaja importante de estos datos para su estudio es que el gasto en consumo está clasificado con los códigos NAICS que proporciona cada comercio donde se realiza el gasto. Mian y Sufi (2009) utilizan todas las hipotecas individualizadas producidas entre 1990 y 2007 (disponibles para cualquier ciudadano gracias a la Home Mortgage Disclosure Act) y los datos sobre crédito e impagos entre 1991 y 2007 por tipo de crédito que proporciona Analytical Services, un proveedor de datos del grupo Equifax. Agregando por condados la información sobre deuda, impagos y créditos hipotecarios (concesión o denegación entre otras variables), Mian y Sufi (2009) muestran que en los condados donde había mayor restricción de crédito antes de comenzar la burbuja inmobiliaria es donde más creció el crédito con el inicio de la expansión inmobiliaria y donde más crecieron los precios de la vivienda. También muestran cómo tras el comienzo de la crisis en esos mismos condados es donde cayeron más los precios y aumentaron más los impagos (13).

En un ejemplo muy reciente, Jiménez *et al.* (2014) utilizan información sobre 24 millones de créditos individuales, con emparejamientos entre bancos y empresas, para analizar el impacto de la política monetaria en la asunción de riesgo por parte de las entidades financieras.

III. *BIG DATA* Y SERVICIOS FINANCIEROS

1. La potencialidad de las técnicas de *big data* en el sector financiero

Hasta 2001 Amazon utilizó docenas de críticos y editores para sugerir títulos que pudieran ser de interés para sus clientes. Amazon Voice fue considerado en su tiempo como el crítico más influyente en Estados Unidos. A finales de los noventa Amazon puso en marcha un equipo para automatizar el procedimiento de recomendaciones de libros para sus clientes. Inicialmente se utilizaron muestras y se buscaron similaridades entre distintos compradores. Hasta que Linden propuso una nueva solución: el llamado filtro colaborativo *item-by-item* (14). El procedimiento utiliza algunos de los principios básicos de *big data*: se usan todos los datos (no se extraen muestras) y se busca capacidad predictiva y no explicativa o causalidad. La técnica de *machine learning* utilizada para realizar las recomendaciones no necesita saber por qué al comprador de *El Quijote* le gustaría también comprar una tostadora. Solo es necesario que exista capacidad predictiva. Cuando se compararon los dos procedimientos (críticos humanos y el algoritmo de *machine learning*) el procedimiento automatizado resultó mucho más eficiente, lo que supuso el desmantelamiento de Amazon Voice. Hoy una tercera parte de las ventas de Amazon son el resultado del sistema personalizado de recomendaciones. El sistema de Linden ha sido adoptado por muchos comercios digitales como por ejemplo Netflix, la compañía de alquiler de películas. Este procedimiento de recomendación aumenta sin duda la satisfacción de los consumidores que pueden encontrar con facilidad productos que necesitan, les interesan y que incluso no eran conscientes de que existían. ¿Podría el futuro de los servicios bancarios discurrir por estos mismos pasos? ¿Podrían los clientes bancarios beneficiarse de sistemas que acomodaran los servicios bancarios a sus necesidades específicas y pudieran ser altamente personalizados? La utilización inteligente de la tecnología y el *big data* abre la posibilidad de que la banca ponga en el centro de su estrategia futura las necesidades de cada cliente de forma singularizada al igual que Amazon realiza recomendaciones personalizadas sobre productos que pueden ser de interés para cada uno de sus clientes. El objetivo debe ser mejorar la accesibilidad de familias de renta media-baja y baja a productos financieros de bajo coste adecuados a su perfil de ingresos, capacidad de pago y nivel de aversión al riesgo. De esta forma se permite el acceso a los servicios bancarios a grupos de la población que no utilizan los mismos o tienen problemas para el acceso, y también se reducen los costes de los servicios. Por ejemplo, en un país financieramente avanzado como Estados Unidos se estima que existen 65 millones de personas que por no tener historial crediticio, o por su brevedad, no tienen calificación crediticia, lo que les impide acceder a los servicios bancarios tradicionales. Este grupo de población es susceptible de acabar suscribiendo un *payday loan* (préstamo con un alto tipo de interés, plazo muy breve y coste entre el 20 y el 30 por 100) o créditos informales.

El último informe de la European Banking Authority (EBA, 2014) vuelve a incidir en los desafíos que afronta la banca europea como consecuencia del exceso de deuda de familias y empresas y el efecto de la regulación prudencial sobre el margen neto de intereses y los beneficios (15). El deterioro de la calidad de los activos afecta no solo a los ingresos y el capital. El modelo de negocio de gran parte del sector financiero no se ha adaptado todavía a las nuevas reglas regulatorias ni a un largo periodo de bajos tipos de interés, lo que perjudica la rentabilidad del sector. De hecho, para algunos modelos de negocio del pasado las nuevas condiciones suponen un peligro inminente a su sostenibilidad. Con RoE bajos y sin claras expectativas de recuperación, la capacidad de captar nuevo capital se reduce, lo que junto al deterioro de los activos y las modestas predicciones de crecimiento económico suponen una situación preocupante para la sostenibilidad de muchos modelos de negocio bancarios. Aunque las expectativas de razonabilidad de los RoE se han moderado en el sector, no es menos cierto que incluso estas menores tasas serán difíciles de conseguir siguiendo los modelos de negocio tradicionales (16).

A los problemas asociados al propio modelo de negocio de los bancos se añade la competencia creciente de nuevos actores en la intermediación financiera que favorecerán la desintermediación bancaria y, por tanto, afectarán también las cuentas de resultados de los intermediarios bancarios tradicionales. En general, la potencial competencia en medios de pago con monedas criptográficas (Bitcoin, Ripple, etc.), pagos vía móvil, las llamadas monedas complementarias, los préstamos *peer to peer* o la financiación de la banca en la sombra a través de vehículos alternativos suponen un desafío para los modelos de negocio tradicionales que va más allá de la adopción de la multicanalidad o la omnicanalidad y amenaza a la esencia del negocio bancario. En el caso español, el desarrollo del Mercado Alternativo Bursátil (MAB) y el Mercado Alternativo de Renta Fija (MARF) también supone un paso en la dirección de la desintermediación. La desconfianza de gran parte de la ciudadanía hacia la banca como consecuencia de la pérdida de reputación del sector durante la crisis financiera y la extensión de la filosofía *low cost* al consumo, con sus implicaciones sobre una menor tendencia al endeudamiento, hacen de la recuperación de la rentabilidad en las entidades financieras una tarea compleja.

La disponibilidad creciente de datos y la utilización más intensiva de los ya existentes en las entidades financieras permite mejoras de eficiencia en la provisión de servicios financieros y aumenta la satisfacción de los clientes. Al igual que Amazon suele acertar con sus recomendaciones o la medicina se orienta a la personalización, los servicios bancarios avanzarán en la misma dirección con la utilización cada vez más intensiva de sus datos. De esta forma el sector financiero podría mejorar su competitividad y hacer frente a los nuevos competidores.

El *big data* en el sector financiero se utiliza fundamentalmente para optimizar las relaciones con los clientes, mejorar las funciones financieras, reducir el riesgo y asegurar el cumplimiento regulatorio. Las siguientes secciones describen aplicaciones de *big data* a los servicios financieros.

2. *Big data* y calificaciones crediticias

> «It's the Wild West… like the early days of FICO».
>
> PETER FADER, catedrático de Marketing de Wharton UPenn, refiriéndose a la utilización de *big data* y la incorporación de datos de redes sociales para calcular la calidad crediticia de un individuo.

Una de las tareas más importantes de una entidad financiera es evaluar adecuadamente el riesgo y la capacidad de pago de los demandantes de crédito. Esta tarea se realiza habitualmente sintetizando en un valor, el llamado *credit score*, la probabilidad de impago de un crédito. El *credit score* representa la calidad crediticia de un cliente como una estimación de la probabilidad de que dicha persona retorne el crédito. Por tanto este indicador condiciona tanto la obtención o denegación de un crédito como el tipo de interés del mismo y los límites crediticios.

Los sistemas utilizados por los bancos se basan en indicadores internos sobre relaciones previas con los clientes, transacciones pasadas, información disponible sobre renta y riqueza, etcétera. En muchos casos los modelos son propios de cada entidad. El procedimiento más conocido y estandarizado de producción de *credit scores* es el llamado FICO, generado por la Fair Isaac Corp. en Estados Unidos. El primer sistema se creó en 1958 y en 1970 se aplicó a tarjetas de crédito bancarias. Aunque la fórmula exacta para su cálculo es desconocida, los pesos aproximados de los distintos componentes del FICO son los siguientes:

a) Historial de pagos anteriores: 35 por 100. Pagos retrasados de facturas o créditos (hipotecas, tarjetas de crédito, préstamos

para la adquisición de automóviles, etc.) causan una reducción del FICO.

b) Utilización de crédito: 30 por 100. La ratio de la deuda actual sobre el límite total de crédito disponible.

c) Extensión del historial crediticio: 15 por 100. A medida que el historial crediticio aumenta, si no se producen impagos, el FICO aumenta.

d) Tipos de crédito utilizados: 10 por 100. Un historial crediticio con múltiples tipos de créditos (hipotecario, al consumo, etcétera) supone un FICO elevado.

e) Búsqueda reciente de crédito o cantidad de crédito obtenido recientemente: 10 por 100. Múltiples solicitudes de nuevo crédito pueden perjudicar el FICO.

Los componentes del FICO muestran por qué resulta complicado conseguir un crédito o una tarjeta, al menos en Estados Unidos, con un historial crediticio breve o sin historial. Asimismo en países con poco desarrollo financiero, donde no existen agencias especializadas que proporcionen informes crediticios, es complicado gestionar el riesgo y, por tanto, es difícil que muchas familias accedan al mismo.

En un plano más sustancial se podría pensar que las variables utilizadas tradicionalmente para medir la calidad crediticia de un demandante de crédito no son las más adecuadas. Jeff Steward, fundador de Lenddo Ltd., una compañía que utiliza las conexiones sociales para calcular *credit scores*, suele referirse a una anécdota para explicar la inadecuación de los criterios habituales para calcular la probabilidad de impago. En una comparecencia parlamentaria un abogado le preguntó a J.P. Morgan en 1912 si la riqueza de un hombre era el factor más importante para prestarle dinero. Morgan respondió con total seguridad: «No señor. El factor más importante es su reputación… un hombre en quien no confío no obtendría dinero de mí ni con toda la riqueza de la Cristiandad».

Esta visión de la solvencia crediticia aplicada en la era del *big data* lleva directamente al llamado «*credit score* social», que considera la reputación y el estatus social *online* y los contactos (17) como factores fundamentales, en particular para solicitantes con poco historial crediticio. Las fuentes básicas para medir dicha reputación se encuentran en Facebook, Twitter o LinkedIn (18). El modelo de negocio de las diferentes empresas que se dedican a la estimación del *credit score* social es diferente. Por ejemplo, Neo Finance (Palo Alto) se especializa en créditos para la adquisición de vehículos de solicitantes jóvenes con un breve historial crediticio y que tendrían que pagar tipos de interés muy altos en los canales habituales, aunque con elevada probabilidad tendrán altos ingresos en el futuro. Neo utiliza, con autorización del cliente, el historial laboral y el número y la calidad de las conexiones en LinkedIn con los trabajadores de su empresa para predecir la estabilidad en el empleo y los ingresos futuros. También estima la rapidez con la que volvería a conseguir un trabajo en caso de ser despedido utilizando los contactos con personal de otras empresas. El objetivo último es predecir la estabilidad en el empleo. Para mejorar la capacidad predictiva de su *credit score* Neo almacena también los datos de Facebook de sus clientes. De hecho, Neo utiliza como emblema publicitario la frase «credit based on income not 'credit score'».

Kreditech es una *start-up* alemana que también utiliza Facebook para realizar sus estimaciones (19). Inicialmente su objetivo era la concesión de pequeños créditos *online* en Alemania, Polonia y España, aunque se ha extendido a muchos otros países como México, República Dominicana, Brasil, Perú, Rusia, República Checa, Rumanía, Filipinas y Australia. En este caso los demandantes cuyos amigos parecen tener mejores trabajos o viven en mejores barrios tienen mayor probabilidad de obtener un crédito, mientras que demandantes que tienen amigos que han impagado créditos tienen menor probabilidad. La empresa señala que su algoritmo de *machine learning* tiene un tiempo medio de resolución de una solicitud de ocho segundos, menos de seis minutos hasta recibir el dinero y menos de un 10 por 100 de impagos.

Lenddo es otra empresa que utiliza las redes sociales para estimar el «capital social *online*» de los demandantes de crédito. Con este objetivo construye un *score* entre 0 y 1.000 determinado por el número de seguidores en Facebook, las características de los mismos, su nivel educativo y su empleador y el historial crediticio de los amigos (20). Lenddo utiliza 100 bases de datos y redes sociales para evaluar el número de conexiones del demandante de crédito y su localización geográfica. Una característica diferencial de Lenddo es la utilización de la presión de grupo para el pago de créditos impagados. Si un amigo deja de pagar esto afecta al *credit score* del resto de los amigos. En este sentido el mecanismo es muy similar al utilizado por los micro-

créditos y, en particular, el procedimiento habitual de presión del grupo (*peer pressure*) en los crédito del Grameen Bank.

Aunque FICO ha señalado que, en principio, no tiene intención de añadir datos sobre redes sociales en sus modelos de estimación del *credit score*, otras agencias como Experian con su Extended View o Equifax con el nuevo VantageScore están abiertas a todo tipo de información.

En cualquier caso no parece que iniciar desde cero la tarea de generar indicadores de la calidad crediticia de los individuos basándose solo en su actividad en redes sociales o información accesible en Internet sea una estrategia que pueda funcionar. Algunos lo han denominado como «reinventar la rueda de las calificaciones crediticias». Los procedimientos que utilizan las entidades financieras en la actualidad (21), basados en una experiencia de muchos años y datos de alta calidad, difícilmente podrán ser sustituidos por la información en las redes sociales que contienen mucho ruido y poca señal.

Pero es lógico complementar las herramientas habituales con *big data* para mejorar la capacidad predictiva de los modelos. Por ejemplo, BBVA utiliza el pago con tarjeta en TPV, anónimo y agregado, para mejorar la estimación de la calidad crediticia de pequeños comercios. La evolución de la facturación en TPV puede indicar una mejora/empeoramiento del consumo en un determinado sector, un área geográfica, etcétera. BBVA comercializa una herramienta para aportar valor a terceros, con los datos de TPV de distintos tipos de comercios y áreas geográficas, para ofrecer una visión sobre las posibilidades futuras de una inversión empresarial, la evolución del comercio en el entorno de la localización elegida, los patrones de compra de los potenciales clientes y las posibilidades de *crosselling* y diversificación, los resultados de la competencia, etcétera (22).

Las tarjetas de crédito también se utilizan para realizar predicciones de gasto y empleo agregado que empresas de servicios financieros venden a otras empresas. Por ejemplo, MasterCard tiene un producto denominado «Spending Pulse» que proporciona datos en tiempo real sobre consumo en diferentes categorías comerciales, y Visa genera predicciones periódicas bastante acertadas sobre resultados de encuestas económicas. Moody's Analytics predice cada mes el empleo en el sector privado utilizando unas 500.000 empresas a las que ADP proporciona el *software* para las nóminas (23).

Khandani *et al.* (2010) muestran cómo se pueden mejorar los modelos de predicción del riesgo de impago ampliando el número de variables y utilizando algoritmos de *machine learning*. En particular, combinando la información sobre las transacciones de los clientes y los *credit scores* generados por agencias de calificación de la calidad crediticia de los consumidores (*credit bureaus*) se puede obtener una mejora muy significativa de la predicción fuera de la muestra. La mejora en la predicción del impago alcanza el 85 por 100 con un ahorro de entre el 6 y el 25 por 100 de las pérdidas totales. Kallerhoff (2012) también muestra, usando algoritmos de *machine learning*, que la calificación crediticia original es el componente más importante en la predicción de impagos en una muestra de 500.000 clientes y 250 millones de transacciones de cooperativas de crédito (24).

Einav *et al.* (2012) utilizan datos de un prestamista *subprime* especializado en la financiación de coches a consumidores con rentas bajas o mal historial crediticio. Los datos disponibles incluyen no solo las condiciones de los contratos y las opciones de devolución del crédito, sino también si los demandantes rechazaron el crédito. Einav *et al.* (2012) muestran que una entrada elevada filtra a los demandantes más arriesgados y limita el tamaño de los créditos, mientras que un margen superior en la venta del vehículo produce créditos de mayor tamaño. Por tanto, la entrada genera un *trade-off* entre la calidad del crédito y su tamaño. Las estimaciones indican que utilizar la entrada como mecanismo de *screening* mejora sustancialmente la calidad media de los créditos.

3. Aplicaciones a la detección de fraude en tarjetas de crédito

Otra de las aplicaciones más frecuentes de *big data* en los servicios financieros tiene que ver con la detección de fraude en tarjetas de crédito. El motivo es claro: un procedimiento sofisticado de detección de fraude en tarjetas de crédito puede ahorrar cientos de millones de euros a un banco. En este caso es fundamental una de las uves de *big data*: la velocidad. La autorización de una tarjeta de crédito se debe producir en unos pocos milisengundos y, por tanto, es necesario optimizar los sistemas para hacer compatible la detección de operaciones sospechosas con la rapidez en la realización de la transacción o su verificación posterior. No obstante, los algoritmos de detección de fraude y la información utilizada para hacerlo cambian constantemente a

medida que los criminales alteran sus métodos para escapar de las mejoras anteriores.

La cantidad de datos utilizados para la detección de fraude es ingente: datos sobre empleados, aplicaciones, fallecidos, encarcelados, listas negras, IRS, etcétera, así como patrones que pueden extraerse de la distribución geográfica de los pagos, las características del sector del negocio, de establecimientos similares, etcétera.

Aunque este es uno de los servicios más tradicionales de aplicación de *big data*, la cantidad de información utilizada sigue siendo pequeña (25) y las aplicaciones son significativamente mejorables pues todavía generan muchos falsos positivos. En una reunión reciente que coordiné sobre la aplicación de *big data* a la gestión empresarial (26), la mayoría de los ponentes había sufrido alguna vez la cancelación de su tarjeta de crédito por la aplicación de reglas sobre límites en la extensión geográfica aceptada de utilización de una misma tarjeta de crédito en un breve espacio de tiempo. Existen cuatro aproximaciones básicas en detección de fraude en tarjetas. La primera, y la más extendida, es la basada en reglas (patrones conocidos). En este caso los algoritmos utilizan desajustes directos entre fuentes de datos para detectar el fraude. Por ejemplo, una compra en Santiago de Compostela, después de haber pagado una comida en Barcelona unas horas después de una compra en Tel Aviv puede producir la cancelación de la tarjeta si el algoritmo prioriza la simultaneidad de las compras. Otro patrón que despierta sospechas es la falta de coincidencia entre el domicilio de facturación y el domicilio de envío de la compra.

El segundo procedimiento es la detección de anomalías u *outliers* (patrones desconocidos). Una posibilidad es la inconsistencia entre la utilización corriente y la historia pasada de uso de la tarjeta de crédito. Otros métodos alertan sobre un volumen anormal de facturación en un negocio en comparación con otros proveedores del mismo servicio. De nuevo, un cambio de hábitos de consumo puede provocar una cancelación de tarjeta.

El tercer procedimiento es utilizar análisis predictivo en búsqueda de patrones complejos. Por ejemplo, cuando un empleado de un determinado establecimiento copia tarjetas para su utilización fraudulenta normalmente testa la tarjeta haciendo pequeñas compras (por ejemplo una App de 99 céntimos en la tienda Apple) para saber si los procedimientos de protección contra el fraude lo han detectado. Los algoritmos de detección pueden buscar asociaciones entre pagos con tarjeta en un determinado comercio y compras posteriores de pequeño tamaño por los mismos usuarios. Por último, las redes sociales permiten buscar patrones por conexión asociativa. En este caso se buscan relaciones con conocidos defraudadores, redes de colusión entre receptores y negocios, manipulación de identidad, etcétera.

Seguramente la aproximación más apropiada a la detección del fraude en tarjetas de crédito sea la que utiliza un modelo híbrido, donde las cuatro metodologías comentadas con anterioridad tienen un papel a jugar. Lógicamente la complejidad de los algoritmos y su efecto sobre los tiempos de respuesta suponen en estos momentos una limitación que el avance de la tecnología puede hacer cada vez menos relevante.

En cualquier caso, la utilización todavía más intensiva de *big data* en la prevención del fraude permitirá reducir el número de tarjetas canceladas por falsos positivos y el procesamiento de la mayoría de las operaciones sin necesidad de solicitar varias pruebas de identidad a los clientes.

4. *Big data* y las nuevas plataformas de servicios financieros

Al tiempo que servicios como Amazon Prime o iTunes nos permiten pagar sin tener que utilizar la tarjeta de crédito se hace más necesario el control de ingresos o gastos. El asesoramiento financiero es uno de las grandes nichos en la gestión de *big data* en los servicios financieros que son nativos digitales. Empresas como HelloWallet, Mint.com, Level, Billguard, You Need a Budget (YNAB) o Mvlopes proporcionan a sus clientes información sobre sus gastos categorizados, predicción de *cash flow*, ahorro objetivo, etcétera, permitiéndoles agregar la información en una solo web.

Otros servicios tienen como objetivo fundamental la educación financiera o la promoción del ahorro. Smartpig, que utilizada como depositario al BBVA Compass, permite a los ahorradores fijar unos objetivos de ahorro y contribuye a esos objetivos combinando un tipo de interés atractivo y premios. El objetivo es promocionar entre sus clientes una mentalidad de ahorrar para luego gastar.

Obviamente muchos de estos nuevos negocios de servicios financieros facilitados por el *big data* y los nuevos procedimientos de tratamiento de información masiva han desaparecido. Este

es el caso, por ejemplo, de PerkStreet, que comenzó su actividad en 2009 y cerró en 2013 sin haber podido consolidar su modelo de negocio. En otros casos las entidades han sido absorbidas por entidades financieras tradicionales. Este es el caso de BillShrink, cuyo modelo de negocio se basaba en la colaboración con comercios y bancos para capacitar a los primeros a ofrecer beneficios por el gasto total en dicho comercio y analizar la proximidad a un comercio de la cadena para el envío de dichos cupones. BillShrink se transformó en Truaxis y fue absorbido por Mastercard.

Simple.com es otro banco digital donde la simplicidad es la base del modelo de negocio. Simple solo ofrece cuentas corrientes asociadas a una tarjeta de débito. Además, como en algunas de las plataformas comentadas anteriormente, existe la posibilidad de fijar metas de ahorro (generales o específicas, como por ejemplo para un viaje en verano), organizar los gastos por categorías, tener en cuenta gastos que todavía no se han producido, crear presupuestos para determinadas partidas (alimentación, gasolina, etc.). Recientemente BBVA ha pasado a ser el socio financiero de esta plataforma.

La adquisición por parte de entidades financieras de algunas de estas plataformas que plantean negocios bancarios en Internet es una reacción a la amenaza que supone la entrada de competidores como la telecos, Google, Facebook, Apple o PayPal en las actividades tradicionales de la banca. Estas nuevas iniciativas tienen ventajas claras: poseen una cantidad ingente de datos, no tienen el condicionamiento de una tecnología heredada anticuada y difícilmente escalable, y no tienen el estigma de ser una entidad bancaria en su origen (en este segundo punto comparten percepción con la banca *online* nativa). Por tanto podrían beneficiarse de la tendencia a la salida de la banca tradicional que se observa en algunos sectores de la población. Parece razonable pensar que cuando alguno de estos gigantes de Internet tomen decididamente el camino de ofertar servicios bancarios (algunos ya lo hacen a pequeña escala) serán también unos competidores importantes por las nuevas *start-up* que ofrecen servicios financieros.

Por último, el proceso de desintermediación financiera está generando el incremento del recurso al mercado frente a la financiación bancaria o la obtención directa de préstamos de fondos de inversión, *hedge funds*, etc. Esta tendencia se está observando fundamentalmente en los préstamos corporativos. Esto no quiere decir que no existan también nuevas plataformas en Internet que, aprovechándose del *big data*, ofrezcan *peer-to-peer lending* a consumidores y pymes. Este es el caso de RateSetter o Zopa en Reino Unido o el Lending Club o Prosper Marketplace en Estados Unidos. En el caso de la financiación a estudiantes se pueden encontrar Social Finance, CommonBond o Upstart (27). Al igual que sucede en otros servicios financieros, la atención de los medios de comunicación a la deuda de los graduados universitarios de Estados Unidos y los problemas en su devolución ha espoleado la aparición de estos servicios, muchos de los cuales se financian con *crowdfunding*. Estas plataformas no son bancos sino que hacen de intermediarios entre inversores y solicitantes de crédito. El modelo de negocio de estas iniciativas tiene en común unos costes operativos muy bajos, escaso condicionamiento por las plataformas tecnológicas del pasado y la capacidad de usar *big data* para evaluar la calidad crediticia de los solicitantes de crédito. Al igual que en el caso de los seguros, los solicitantes de crédito aceptan que estas plataformas obtengan datos sobre ellos de sus empleadores así como información de redes sociales.

5. Otras utilidades del *big data* en los servicios financieros

Sería muy extenso intentar recorrer toda la potencialidad del *big data* en el campo de los servicios financieros. Sin duda un aspecto importante es la optimización de la función financiera (por ejemplo tesorería, negociación en mercados financieros, etc.). En estos momentos los algoritmos para hacer *trading* en los mercados financieros son la tendencia más habitual. Lo importante en estos casos es la rapidez de acceso al mercado y la capacidad de análisis de la información disponible.

En el apartado de segmentación de los clientes y las recomendaciones de productos, el sector financiero lleva bastante retraso respecto al sector comercial. Además, la creciente multicanalidad de las relaciones entre bancos y clientes genera potencialmente dificultades para agregar toda la información relevante para mejorar la provisión de servicios a un cliente particular. No son frecuentes los proyectos que intentan correlacionar las interacciones de los clientes con campañas de marketing o comportamiento de navegación del cliente. Sin embargo la capacidad para fusio-

nar datos provenientes de diferentes canales e identificar las necesidades de los consumidores puede producir grandes ventajas competitivas en forma de mayores ventas, reducción de costes, retención de los mejores clientes y aumento de los índices de satisfacción.

El sector bancario hace tiempo que realiza una segmentación de sus clientes, aunque, por lo general, poco sofisticada. La aplicación de *big data* permite hacer una segmentación más fina, en principio pudiendo llegar a la individualización, con componentes dinámicos (cambio de condiciones personales del cliente, movimiento entre segmentos, etcétera) y una segmentación en tiempo real. La multicanalidad permite una atención personalizada a los clientes. La geolocalización asociada con los canales móviles permite, por ejemplo, ofrecer a los clientes descuentos y beneficios en tiempo real en comercios cercanos a su localización corriente en función de su histórico de consumo.

Pero sin duda la disponibilidad de grandes cantidades de datos internos de las entidades permitiría proyectos de segmentación dinámica de sus clientes a lo largo de su ciclo vital en cada entidad. Por ejemplo, Kallerhoff (2012) utiliza un análisis de clúster sobre el *mix* de productos de cada cliente bancario. El número óptimo de clústeres se fija a partir de la capacidad de predicción fuera de la muestra. El objetivo es analizar no tanto la configuración estática de estos clústeres sino la probabilidad de que un cliente pase de un clúster a otro. Esta probabilidad se calcula a partir de la distancia entre los productos del cliente y el centro de cada clúster. Los cambios entre clústeres pueden predecirse a partir de las operaciones realizadas en el pasado por el cliente y por los cambios en la base de clientes. La predicción se realiza con una técnica habitual en *machine learning* conocida como «*support vector machine*» y se prueba en el comportamiento pasado de los clientes. Las variables que toman como *inputs* el algoritmo se basan en los datos transaccionales y variables externas. Estas predicciones se pueden utilizar para recomendar productos a clientes que tienen una elevada probabilidad de tránsito de un clúster a otro. Las características, en términos de productos, de cada uno de los clústeres definen los productos más adecuados en cada fase del ciclo vital del cliente en la entidad.

Otra utilidad importante de *big data* en las entidades bancarias es la mejora de la información a los clientes. Las páginas web de los bancos generan enormes cantidades de información a partir de los clicks de sus clientes. De esta forma se puede saber qué productos visitan con más frecuencia los consumidores y, en general, sus intereses. Toda esta información del *weblog* es difícil de almacenar y analizar. Para mejorar el servicio a los clientes y reducir costes es necesario un proyecto de *big data* que fusiona la información del *weblog* con la información de los servicios automatizados de respuesta telefónica IVR (Interactive Voice Response) y la grabación de la voz del servicio telefónico. Para reducir el coste y mejorar el servicio es preciso entender dónde están los problemas en la información que se proporciona en la página web y resolverlos mediante actualizaciones frecuentes y bien orientadas.

Las compañías de seguros, cuyos productos son vendidos en muchos casos por bancos en acuerdos banca-seguro, tienden a fijar las primas utilizando la edad, el sexo, el tipo de coche, su color, educación del conductor, años con permiso de conducir, etcétera. Sin embargo, en la era del *big data* se puede conseguir una personalización más eficiente. Por ejemplo en los seguros de automóvil es muy útil, si el cliente lo permite, la colocación de sensores telemáticos en su vehículo para monitorizar cómo conduce, a qué horas del día, la velocidad, si frecuentemente pega frenazos, etcétera. De esta forma se pueden personalizar de forma muy precisa las primas reduciéndolas entre un 20 y un 40 por 100 en los buenos conductores. También se reducen significativamente los problemas de información asimétrica que las variables habitualmente utilizadas no pueden resolver de manera adecuada. En todo caso, en la actualidad no llega al 3 por 100 la proporción del mercado de seguros del automóvil que utilizan esta información para fijar sus primas.

Otro aspecto importante de *big data* en el sector bancario es el cumplimiento de las normativas regulatorias, cada vez más complejas y extensas. Por ejemplo, la detección de posibles abusos de mercado en la negociación en mercados financieros (uso de información privilegiada, anomalías en la negociación, etcétera) o la prevención del blanqueo de capitales y financiación del terrorismo. En este último apartado existen, como en muchos otros, herramientas adaptadas que incorporan procedimientos de *machine learning* e inteligencia artificial a la generación de alertas en operaciones que puedan cumplir determinadas condiciones. Los sistemas actuales utilizan aún poca información de la disponible. En muchos casos se limitan a comparar listas

de PRP (personas con responsabilidad pública) o SDN (ciudadanos de otros países especialmente susceptibles) o cuentas sospechosas. En la práctica generan multitud de falsos positivos y requieren una supervisión humana muy intensa. Pocas utilizan todavía indicadores de redes sociales u otros indicios que puedan generarse en el tráfico en Internet.

También es importante el papel de *big data* en la optimización de la estructura del capital de las entidades bancarias. En un tiempo donde el aumento de la regulación hace del capital un bien muy escaso es preciso ajustar y optimizar la utilización del mismo. El *big data* puede permitir esta optimización casi en tiempo real. Asimismo, en un tiempo en el que el sector bancario es visto con suspicacia por la población y donde su reputación se ha visto muy afectada por todo tipo de escándalos y prácticas empresariales poco deseables, es muy importante recuperar la confianza de los clientes y poder mostrar que efectivamente el sector financiero tiene un papel fundamental a cubrir en una economía moderna. En este sentido algunos bancos están estructurando parte del salario variable de sus empleados en función de la satisfacción de los clientes. Desde esta perspectiva, la utilización de *big data* a partir de los comentarios en las redes sociales y el análisis del sentimiento de marca a partir de estos datos puede proporcionar métricas para avanzar en esta dirección.

IV. LOS PELIGROS DEL *BIG DATA*

Ciertamente las técnicas basadas en *big data* tienen un enorme potencial para mejorar la provisión de servicios financieros y la satisfacción de los clientes. Sin embargo, utilizadas indiscriminadamente y sin los objetivos correctos, pueden representar un peligro. En esta sección se analizan algunos de estos potenciales problemas.

En primer lugar, si bien es cierto que *big data* proporciona herramientas muy útiles en un ambiente de mayor incertidumbre, regulación y desconfianza de los consumidores en el sector financiero, no es menos cierto que la transformación de un proyecto de *big data* en un programa de éxito no está garantizada. Estos proyectos tienen riesgo y si no se orientan adecuadamente, utilizando evidencia que indique que dicha actividad puede tener éxito, se pueden convertir en ejercicios caros que se pierden en un mar de datos. Es evidente que la acumulación de información tiene rendimientos decrecientes y que analizar datos con información redundante no puede ayudar a resolver ningún problema. Los datos por sí mismos no proporcionan una ventaja competitiva a no ser que el análisis sea adecuado, por lo que es muy importante contar con un equipo profesional de *analytics* que pueda extraer conclusiones apropiadas a partir de los datos. También es conveniente realizar un estudio coste-beneficio o una estimación del ROI de un proyecto de *big data* antes de pasar de un piloto a una implantación general.

Esta tarea no es tan sencilla como pudiera parecer. Y este es el segundo punto. La existencia de grandes cantidades de datos no puede hacer olvidar los fundamentos de la ciencia estadística, la influencia de los errores de medida o la precaución contra la utilización de correlaciones espurias. Además del conocimiento técnico hace falta estar dispuestos a analizar constantemente la capacidad predictiva de los modelos y hacer ajustes a medida que el sistema pierde potencia explicativa. La experiencia de Google Trends en la predicción de la expansión de la gripe proporciona una señal de alerta. El GFT (Global Flu Trends) funcionó de manera espectacular en 2009 cuando predijo la expansión de la gripe con bastante exactitud a un nivel muy desagregado, mientras que el Centro de Control de Enfermedades (CDC) tardaba semanas en proporcionar la misma información. El procedimiento del GFT se basa fundamentalmente en la aplicación de miles de modelos a la búsqueda en Google de términos relacionados, al menos teóricamente, con la preocupación por la gripe. Varios artículos recientes (28) muestran que las predicciones de GFT no fueron tan precisas a partir de 2012, y en 2013 la estimación de Google era el doble de la estimación del CDC. Un modelo que combina Google Flu Trends, retardos de la estimación del CDC, retardos en el error de GTF y variables estacionales semanales tiene un poder predictivo muy superior.

El tercer tema tiene que ver con la privacidad en la utilización de los datos. Los casos de los informes médicos anonimizados de los funcionarios de Massachusetts, el premio de Netflix o la publicación anonimizada de una base de datos de búsquedas de AOL han puesto de manifiesto que en muchos casos se puede conseguir reidentificar a los individuos anonimizados (29). La rapidez con la que se generan nuevos datos y situaciones especiales con los mismos implica que las leyes de protección de datos vayan siempre por detrás de la realidad. El tema no es simplemente la protección de

los datos sino también que la utilización de patrones a partir del análisis de los datos puede generar situaciones discutibles desde el punto de vista de la privacidad: si un modelo puede predecir la probabilidad de embarazo de la clienta de un centro comercial, ¿es éticamente aceptable enviarles a una menor cupones de pañales y ropa de bebé?

Las nuevas cláusulas de consentimiento en las interacciones con proveedores de servicios en Internet intentan evitar estos problemas. Sin embargo, desde el Derecho se habla de la falacia del consentimiento (30), pues el cliente da su consentimiento en la gran mayoría de los casos sin leer las condiciones. Además el consentimiento debe estar ligado no solo a los datos sino también a su utilización para una finalidad determinada para la cual los datos fueron recogidos. Sin embargo en *big data* el concepto de reutilización de los datos es un elemento fundamental: datos que han sido recogidos con un objetivo al tiempo pueden encontrar otra utilización completamente diferente al objetivo original. Por último está la cuestión del valor de los datos del usuario. Este es un tema más controvertido pues, si bien es cierto que los usuarios acceden a la utilización de sus datos sin recibir ninguna compensación directa, en la mayoría de los cosas las mejoras de eficiencia conseguidas por los procedimientos de *big data* repercuten en los precios que pagan los clientes. El caso del seguro del automóvil es un claro ejemplo.

En cuarto lugar está la posibilidad de que los errores en la captura, fusión o limpieza de los datos generen consecuencias negativas para los ciudadanos a partir de la aplicación de técnicas de *big data* a problemas concretos. Un ejemplo es la industria de generación de *credit scores* a partir de *big data* captado en Internet. NCLC (2014) analizó la información disponible por varias agencias de generación de *credit scores* a partir de datos de Internet. En la sección anterior se comentó que el objetivo de la utilización de *big data* para el análisis de la calidad crediticia de los consumidores es superar los problemas que tienen familias sobre las que las empresas que tradicionalmente han realizado *scoring* de particulares (Equifax, TransUnion o Experian) no tienen información. NCLC (2014) seleccionó a cinco compañías de *big data* y obtuvo los *reports* sobre quince voluntarios para el estudio. Los informes recibidos tenían multitud de errores, estimaciones desmesuradas (salario doble del real del solicitante), direcciones incorrectas, multitud de información faltante (incluidas cuentas en redes sociales, etc.). Ciertamente los datos de las empresas tradicionales de generación de *credit score* de consumidores también son muy mejorables. Un estudio de 2013 de la Federal Trade Commission de Estados Unidos señalaba que el 20 por 100 de los informes crediticios de estas compañías contienen errores y un 5 por 100 de estos errores resultaron en una rebaja del *credit score* que impidió a los clientes conseguir un crédito o les supuso pagar un tipo superior. El problema de las agencias que se basan en *big data* es que los clientes no tienen forma de saber cómo se ha calculado su *credit score*. No hay forma de confirmar independientemente la capacidad predictiva del algoritmo utilizado. De esta forma el consumidor puede acabar siendo afectado negativamente por un *credit score* calculado a partir de datos erróneos, aunque, a diferencia del caso de las agencias tradicionales, es más difícil estimar si esto ha afectado a su capacidad de conseguir un crédito o el tipo.

La mala calidad de algunos datos utilizados no solo puede tener efectos negativos sobre los clientes sino que también puede incumplir alguna regulación. Por ejemplo, en Estados Unidos la Fair Credit Reporting Act (FCRA) exige que las agencias de calificación de consumidores produzcan informes veraces y ajustados para proteger la reputación y privacidad de los consumidores. Además los consumidores tienen derecho a recibir gratuitamente el informe crediticio si a causa del mismo un consumidor ha visto rechazada una solicitud de crédito (incluso aunque sea una tarjeta de crédito o una tarjeta de fidelización de un comercio). La utilización de *big data* puede producir incumplimientos de la Equal Credit Oportunity Act (ECOA) en Estados Unidos incluso si el modelo predictivo no incluye características raciales de los individuos, pues muchas otras variables están correlacionadas con estas y podrían derivar en un incumplimiento implícito de la ECOA. Muchas empresas de venta de datos incluyen desde hace algún tiempo en los contratos una limitación de responsabilidad aduciendo que sus datos no constituyen un informe sobre un consumidor ni pueden ser utilizados para establecer la elegibilidad de los individuos para créditos o seguros ni para empleos ni promociones. En cualquier caso, la FTC indica que si los informes son utilizados para la concesión de créditos o la selección de personal, entonces, con independencia de la cláusula de limitación de responsabilidad, se debe respetar la regulación.

La utilización de la localización de los demandantes de crédito (por ejemplo el código postal donde está localizada la IP desde donde se actualiza una página de Facebook o Twitter), o la calidad crediticia de clientes en establecimientos donde el demandante de crédito ha comprado recientemente también pueden afectar al *credit score* y, de nuevo, su legalidad bajo determinadas regulaciones podría ser cuestionada.

Por último los métodos de *scoring* que utilizan fundamentalmente datos de Internet además de utilizar variables con mucho ruido y poca señal se enfrentan a la posibilidad de manipulación. Si el *scoring* depende de los amigos que se tienen en Facebook no resulta complicado pensar en formas de aumentar el *score* simulando amigos.

V. CONCLUSIONES

La utilización de *big data* está transformando la gestión de muchas actividades empresariales. Las posibilidades que abre son enormes tanto para reducir los precios y mejorar los servicios que reciben los consumidores como para reducir los costes de las empresas. La disponibilidad de enormes bases de datos está también revolucionando la investigación económica tanto en metodologías como en técnicas específicas. Las aproximaciones tradicionales de regresión líneal o logística, *cluster analysis*, análisis discriminante o predicción usando series temporales se complementan con ideas de *machine learning* para poder trabajar con ingentes bases de datos. El crecimiento de *big data* como sector de actividad ligado a la actividad empresarial está generando una gran demanda de matemáticos, informáticos y estadísticos con conocimientos de economía.

En el sector financiero existen multitud de aplicaciones de *big data* que abarcan la microsegmentación, la calificación crediticia de los consumidores, la dinámica, predicción y recomendación de nuevos productos, la detección del fraude en tarjetas de crédito, la identificación de operaciones sospechosas de blanqueo de capitales o actividades terroristas, la gestión eficiente de las relaciones con clientes en un contexto de multicanalidad o la fijación de objetivos de ahorro para los clientes a partir del análisis de sus ingresos y gastos.

La creciente importancia de los datos en la gestión de cualquier actividad empresarial, y muy especialmente en el sector financiero, está atrayendo a empresas nativas digitales (Google, Amazon, etc.) hacia la industria financiera a pesar de que la rentabilidad del sector es baja y decreciente. La banca todavía tiene una importante ventaja competitiva frente a otros productores de información masiva: la calidad de sus datos (la señal sobre el ruido) es muy superior a la calidad de los datos de sus competidores. Pero esta ventaja puede irse erosionando en el tiempo con la captación de más información por parte de las grandes empresas de Internet y la utilización de técnicas de análisis cada vez más sofisticadas. La industria financiera tiene que acelerar su transformación sino quiere ver cada vez más partes de su cadena de valor atacadas por competidores externos al sector.

La utilización de la avalancha de datos que se generan en la actualidad produce cuestiones importantes relacionadas con la privacidad, la protección de datos y el cumplimiento de algunas regulaciones. Es importante tener en cuenta estas consideraciones antes de comenzar un proyecto de *big data*, especialmente en el contexto de los servicios financieros. Además hay que analizar, al igual que se haría con cualquier otra inversión, la relación coste-beneficio del proyecto así como contar con un equipo experto que sea capaz de extraer conocimiento a partir de los datos y esté atento a la actualización de los procesos ante cambios en los determinantes que condicionan los objetivos del proyecto. La parábola de los errores de Google Flu Trend debe servir de recordatorio para encontrar un equilibrio entre la búsqueda de factores con capacidad predictiva y el análisis de los motivos por los cuales dichos factores tienen esa influencia, evitando convertir el análisis *big data* en una enorme caja negra.

NOTAS

(*) El autor agradece el apoyo del proyecto ECO2011-25272 del Ministerio de Ciencia e Innovación y del programa ICREA-Academia de la Generalitat de Catalunya.

(1) Por ejemplo el campo de las finanzas siempre ha generado una gran cantidad de información, especialmente en los análisis de microestructura de los mercados financieros y datos ultrafrecuentes. García Montalvo (2003) analiza varios millones de observaciones de precios de oferta, de demanda y de transacción para determinar el impacto de la introducción de *market makers* en los contratos del MEFF.

(2) GigaByte (GB) = 10^9 bytes. TeraByte (TB) = 10^{12}. PetaByte (PT) = 10^{15}. ExaByte (EB) = 10^{18}. ZettaByte (ZB) = 10^{21}.

(3) La Bolsa de Nueva York produce un TeraByte de información sobre negociación al día. Un Airbus genera 10 TB de datos cada media hora. Twitter produce 7 TB al día y Facebook 10 TB. Se estima que a finales de 2012 la información disponible era de 2,7 ZB.

(4) Otro problema técnico tiene que ver con la energía necesaria para sostener el almacenamiento, movimiento y procesamiento de tanta información. Solamente la energía necesaria para mantener el *cloud* representa tanto como la energía consumida en Gran Bretaña en un año. De aquí que el concepto «Green» aplicado a nuevas tecnologías sea un valor en alza.

(5) KING (2013 y 2014) presenta una descripción exhaustiva del impacto de las nuevas tecnologías en el negocio financiero. Una visión más general del impacto de Internet se puede encontrar en BBVA (2013).

(6) Algunos procesos específicos, como la gamificación, facilitados por *big data* tampoco son tratados con detalle.

(7) Una versión más popular de *big data* la define por cuatro uves: volumen (escala de los datos), velocidad (de proceso y análisis), variedad (tipos de datos) y veracidad (frente a la información habitual utilizada para la toma de decisiones). Esta visión tecnológica del *big data* se deja la uve más importante: creación de valor.

(8) EINAV y LEVIN (2013) destacan, junto a la disponibilidad de grandes cantidades de datos con tipologías muy heterogéneas y menos estructura, la posibilidad de generarlos en tiempo real.

(9) Structured Query Language (SQL).

(10) VARIAN (2014) discute la relación entre econometría y *machine learning* a partir de la diferencia entre causalidad y correlación.

(11) *Machine learning* es un conjunto de técnicas que permiten a los sistemas aprender automáticamente los programas a partir de los datos. MURPHY (2012) define *machine learning* como un «conjunto de métodos que pueden detectar automáticamente patrones en los datos para predecir futuros datos o tomar decisiones en condiciones de incertidumbre».

(12) Se almacenan 5 millones de precios de 300 tiendas en Internet en 70 países del mundo.

(13) Otras aplicaciones de *big data* al análisis económico de la crisis financiera se pueden encontrar en MIAN y SUFI (2014).

(14) LINDEN *et al.* (2003). Este algoritmo en lugar de utilizar emparejamientos con clientes similares, empareja los ítems de las compras de los clientes a otros ítems similares para combinarlos luego en un listado de recomendaciones. En el proceso se determina el emparejamiento más similar para un determinado ítem utilizando un algoritmo que construye una lista de ítems similares que el usuario tiende a comprar juntos.

(15) Véase también GARCÍA MONTALVO (2014a).

(16) GARCÍA MONTALVO (2013 y 2014b).

(17) Algunas empresas lo definen como «capital social» en consonancia con el concepto económico del mismo nombre.

(18) Estas no son las únicas fuentes de capacidad predictiva en los nuevos modelos de *scoring* basados en *big data* e información en Internet. Por ejemplo, ZestFinance asegura que su algoritmo ha identificado que los demandantes de crédito que solo utilizan mayúsculas o minúsculas tienen una probabilidad menor de devolver los créditos.

(19) Otra empresa alemana, Schufa, abandonó recientemente sus planes de utilizar las redes sociales después de una controversia en los medios de comunicación.

(20) RUSLY (2003), «Bad credit? Start Tweeting», *Wall Street Journal*.

(21) El procedimiento utilizado para la construcción del FICO se ha mostrado altamente efectivo en la predicción de la probabilidad de impago.

(22) Tanto Telefónica (Smart Steps en colaboración con GfK) como BBVA (Commerce 360 C360) han accedido a la provisión de servicios a terceros basados en su capacidad de generar y analizar *big data*. Smart Steps utiliza datos anonimizados y agregados sobre comunicaciones vía dispositivos móviles para comparar los resultados de tiendas respecto a otras tiendas en la misma área geográfica, elegir la mejor localización para un nuevo comercio, decidir horarios de apertura o cierre de comercios, etcétera.

(23) EINAV y LEVIN (2013).

(24) Para una aplicación más reciente de la potencialidad del *big data* combinado con *machine learning* en la estimación de probabilidades individuales de impago véase KRUPPA *et al.* (2013).

(25) Se estima que VISA ha utilizado tradicionalmente solo el 2 por 100 de la información transaccional. En 2005 utilizaba un único modelo con 40 variables. En la actualidad se consideran más de 500 variables en cada transacción, muchas de ellas relacionadas con geolocalización, y se prueban 16 modelos diferentes que cubren distintos mercados y regiones geográficas. Los modelos y las variables cambian constantemente pudiéndose añadir una variable nueva en una hora cuando en el pasado eran necesarios tres o cuatro días. La utilización de Hadoop hace posible la actualización rápida de los modelos y los datos así como un ahorro de coste por la utilización del paralelismo sobre ordenadores de bajo coste.

(26) «*Big data*: de la investigación científica a la gestión empresarial», Fundación Ramón Areces, 3 de julio de 2014.

(27) La compañía española PeerTransfer está especializada en hacer transferencias internacionales para el pago de las matrículas universitarias.

(28) BUTLER (2013) y LAZER *et al.* (2014).

(29) HEFFETZ y LIGETT (2014).

(30) Ponencia de Ricard Martínez en la Jornada «*Big data*: de la investigación científica a la gestión empresarial», Fundación Ramón Areces, 3 de julio de 2014.

BIBLIOGRAFÍA

BBVA (2013), *C@mbio: 19 ensayos fundamentales sobre cómo Internet está cambiando nuestras vidas*.

BELLONI, A.; CHERNOZHUKOV, V., y HANSEN, C. (2014), «High dimensional methods and inference on structural and treatment effects», *Journal of Economic Perspectives*, 28(2): 29-50.

BREIMAN, L. (2001), «Statistical modeling: the two cultures», *Statistical Science*, 16(3).

BUTLER, D. (2013), «When Google got the flu wrong», *Nature*, 494: 155.

CAVALLO, A. (2012), «Online and official price indexes: measuring Argentina's inflation», *Journal of Monetary Economics*.

CHOI, H., y VARIAN, H. (2012), «Predicting the present with Google Trends», *Economic Inquiry*, 88: 2-9.

EINAV, L.; FINKELSTEIN, A.; RYAN, S.; SCHRIMPF, P., y CULLEN, M. (2013), «Selection on Moral Hazard in Health Insurance», *American Economic Review*, 103(1): 178-219.

EINAV, L.; JENKINS, M., y LEVIN, J. (2012), «Contract pricing in consumer credit markets», *Econometrica*, 80(4): 1387-1432.

EINAV, L., y LEVIN, J. (2013), «The data revolution and economic Analysis». Technical report, NBER.

EUROPEAN BANKING AUTHORITY (2014), *Risk assessment of the European banking system*, EBA.

GOEL, S.; HOFMAN, J.; LAHAIE, S.; PENNOCK, D., y WATTS, D. (2010), «Predicting Consumer Behavior with Web Search», *Proceedings of the National Academy of Sciences*, 107(41): 17486-17490.

GOLDFARB, A., y TUCKER, C. (2012), «Privacy and Innovation», *Innovation Policy and The Economy*, 12: 65-90.

HEFFETZ, O., y LIGETT, K. (2014), «Privacy and data-based research», *Journal of Economic Perspectives*, 28(2): 75-98.

HYUNYOUNG, C., y VARIAN, H. (2012), «Predicting the present», *Economic Records*, 88(1): 2-9.

JIMÉNEZ, G.; ONGENA, S.; PEYDRO, J.L., y SAURINA, J. (2014), «Hazardous Times for Monetary Policy: What do 23 Million Loans Say About the Impact of Monetary Policy on Credit Risk-Taking?», de próxima aparición en *Econometrica*.

KALLERHOFF, P. (2012), «'Big data' and Credit Unions: 'machine learning' in member transaction». Mimeo.

KING, B. (2013), *Bank 3.0,* John Wiley and Sons, Nueva York.

— (2014), *Breaking banks: the innovators, rogues and strategists rebooting banking*, Wiley, New Jersey.

García Montalvo, J. (2003), «Liquidity and market makers: an analysis with ultra-high frequency data», *European Journal of Finance,* 2003(9): 358-378.

— (2013), «La Unión Bancaria y el modelo de negocio bancario en Europa», *Papeles de Economía Española*, 137: 57-79.

— (2014a), «Back to 'boring banking' in the age of deleveraging and new regulation», *Spanish Economic and Financial Outlook*, 3(1): 47-59.

— (2014b), «Banca aburrida: el negocio bancario tras la crisis económica», en *Las Claves del Crédito Bancario tras la Crisis*, Estudios de la Fundación Funcas (Serie Economía y Sociedad), cap. 3, pp. 101-150.

Khandani, A.; Kim, A., y Lo, A.W. (2010), «Consumer credit-risk models via machine-learning algoritms», *Journal of Banking and Finance*, 34(11): 2767-2787.

Kruppa, J.; Schwarz, A.; Arminger, G., y Ziegler, A. (2013), «Consumer credit risk: individual probabilities estimates using 'machine learning'», *Expert Systems with Applications*, 40: 5125-5131.

Lazer, D.; Kennedy, R.; King, G., y Vespignani, A. (2014), «The parable of Google Flu: Traps in 'big data' Analysis», *Science*, 343: 1203-1205.

Linden, G.; Smith, B., y York, J. (2003), «Amazon.com recomendations: ítem-to-item collaborative filtering», *IEEE Internet Computing*, 7(1): 76-80.

Mayer-Schönberger, V., y Cukier, K. (2013), *Big data*, John Murray, Londres.

McLaren, N., y Shanbhoge, R. (2011), «Using Internet search data as economic indicators», *Bank of England Quarterly Bulletin*, 51(2): 134-140.

Mian, A.; Rao, K., y Sufi, A. (2014), «Household balance sheets, consumption and economic slump», de próxima aparición en *Quarterly Journal of Economics*.

Mian, A., y Sufi, A. (2009), «The consequences of Mortgage credit expansion: evidence from the U.S. Mortgage default crisis», *Quarterly Journal of Economics*, 124: 1449-1496.

— (2014), *House of debt*, University of Chicago Press, Chicago.

Munnell, A.; Tootell, G.; Brownew, L., y McEneaney, J. (1996), «Mortgage lending in Boston: reinterpreting HMDA data», *American Economic Review*, 86(1): 25-53.

Murphy, K. (2012), «*Machine learning*»: *a probabilistic perspective*, MIT Press, Cambridge, MA.

National Consumer Law Center (2014), *Bid data: a big disappointment for scoring consumer credit risk*, NCLC, Boston, MA.

Nickerson, D., y Rogers, T. (2014), «Political campaigns and 'big data'», *Journal of Economic Perspectives*, 28(2): 51-74.

Varian, H. (2014), «'Big data': new tricks for econometrics», *Journal of Economic Perspectives*, 28(2): 3-28.