LAS NUEVAS OPORTUNIDADES DEL *BIG DATA* PARA LAS INSTITUCIONES FINANCIERAS

Pedro GALEANO
Daniel PEÑA
Universidad Carlos III de Madrid

Resumen

Este trabajo describe el actual entorno de información masiva, el llamado *Big Data*, en el que se mueven hoy las entidades financieras y analiza las nuevas oportunidades para bancos, cajas de ahorros y otras instituciones financieras de utilizar los nuevos datos disponibles sobre sus clientes, presentes y potenciales, para mejorar sus estrategias comerciales. Se describen las nuevas formas de información disponible, que incluyen no solo las tablas de datos numéricos tradicionales, sino imágenes, textos de la web y datos temporales y espaciales, a veces en forma de funciones. Estos nuevos datos pueden revelar patrones de comportamiento y de relaciones entre variables que permitan segmentar mejor la clientela y construir modelos con mayor capacidad predictiva que los actuales. Se revisan brevemente algunas de las nuevas herramientas desarrolladas en los últimos años bajo los nombres de «aprendizaje estadístico» (*statistical learning*), «inteligencia artificial» y «aprendizaje automático» (*machine learning*) y su potencial en distintos problemas, como predicción personalizada, análisis de redes de clientes, prevención del fraude o detección de la lealtad de los clientes, entre otros. Por último, se presenta un ejemplo de cómo la construcción de redes de clientes y su análisis puede mejorar las políticas comerciales en un gran banco internacional.

Palabras clave: aprendizaje automático, aprendizaje estadístico, *Big Data,* redes de clientes.

Abstract

This paper describes the current mass information environment, the so-called Big Data, in which financial institutions move today and analyses the new opportunities that this situation generates for banks and financial institutions to use the new data available on their present and potential clients to improve their marketing strategies. The new forms of available information are described, which include not only traditional numerical data tables, but images, web texts and temporal and spatial data, sometimes in the form of functions. These new data can reveal patterns of behaviour and relationships between variables that allow better customer segmentation and build models with greater predictive capacity than current ones. Some of the new tools developed in recent years under the names of statistical learning, artificial intelligence, and machine learning are briefly reviewed and their potential in different problems, such as personalized prediction, customer network analysis, fraud prevention or customer loyalty detection. Finally, an example is presented of how the construction of customer networks and their analysis can improve trade policies in a large international bank.

Key words: machine learning, statistical learning, Big Data, customer networks.

JEL classification: C10, C55.

I. *BIG DATA:* LOS NUEVOS DATOS MASIVOS

POR primera vez en la historia de la humanidad existen, en la mayoría de los países, datos abundantes, accesibles y con bajo coste, sobre muchos aspectos del comportamiento de los ciudadanos. Muchos de estos datos se han generado automáticamente, por el uso de dispositivos digitales conectados a Internet, especialmente teléfonos móviles, o por sensores y escáneres que los recogen. Por ejemplo, los escáneres de un supermercado guardan la información detallada de los productos adquiridos por clientes identificados con tarjetas de fidelidad o las transacciones electrónicas con tarjetas bancarias informan no solo sobre el importe de la venta, sino del tipo de comercio, día, hora y lugar, y lo mismo ocurre con muchos de los sistemas de compras *online*. Además, el desarrollo de la tecnología permite almacenar, con escaso coste, estas grandes masas de datos, analizarlos de forma rápida y utilizarlos para obtener valor comercial. Por ejemplo, personalizando los anuncios que se muestran en el teléfono móvil en función del perfil del usuario.

Una de las empresas paradigmáticas en la recogida de información de sus clientes es Google: ocho de sus servicios cuentan con más de 1.000 millones de usuarios, una proporción importante de la población mundial actual, del orden de 7.000 millones de personas, de los cuales se estima que la mitad utilizan Internet. Alphabet, la matriz de Google, dispone de varios gigas de información de cada uno de sus usuarios que, desde hace poco tiempo, pueden además descargarla: incluye las páginas web consultadas, los vídeos de YouTube vistos, las reservas de hoteles y aviones realizadas y, si no tiene desactivada la

geolocalización, las coordenadas de su posición en cada momento, sus contactos y su correo. Además de Google, empresa pionera en este campo, las redes sociales, como Facebook, Twitter o Instagram, almacenan masas de información sobre sus usuarios incluyendo sus contactos, sus fotos y vídeos. Muchas otras aplicaciones recogen a través del uso del teléfono móvil información sobre nuestro comportamiento para formar bancos de datos de valor comercial. El nombre de *Big Data* se refiere principalmente a estas nuevas masas de datos recogidos de forma automática que están ya cambiando el mundo en que vivimos y que engloban no solo los datos personales de millones de personas, sino los datos recogidos por sensores instalados para seguir el comportamiento de animales, el crecimiento de los cultivos, el funcionamiento de máquinas y procesos o la evolución de fenómenos meteorológicos o climáticos.

Para poner en perspectiva que representan varios *gigabytes* (GB) de información de millones de personas, resumiremos brevemente la evolución de la capacidad para almacenar datos digitales y procesarlos. La unidad mínima de almacenamiento digital es un *bit* (b), que representa dos posible valores para una variable (0,1). Por ejemplo, una bombilla puede estar apagada (0) o encendida (1). Uniendo 8 bits se obtiene un *byte* (B), con el que podemos formar 256, (2^8) caracteres (letras, números o símbolos). Con 4 bytes podemos representar un número de cuatro cifras o una palabra con cuatro letras, y una página de papel DIN A4 escrita a máquina con unas 500 palabras y un tamaño medio de cuatro letras por palabra requiere alrededor de 2.000 B o 2 KB (1 kB=1.000 bytes). Un libro impreso de 350 páginas ocupa unos 400 kB=400 x 10^3 bytes. Los primeros ordenadores personales (PC) tenían discos *floppy* capaces de almacenar un libro (360 kB). Los discos duros fijos iniciales de los PC tenían una capacidad de 20 *megabytes* (MB = 1.000 kB=10^6 bytes), lo que permitía almacenar unas decenas de libros, y pasaron, a comienzos de este siglo, a una capacidad de varios gigabytes (GB=1.000 MB=10^9 bytes), suficientes para almacenar fotos (entre 50 kB y 2 MB), música (una sinfonía ocupa unos 80 MB), o películas (entre 5 y 1,5 GB). Hoy, un PC puede almacenar varios *terabytes* (TB=1.000 GB =10^{12} bytes), es decir, cientos de películas, miles de canciones y fotos y cientos de miles de libros. Como referencia, la colección impresa de la Biblioteca del Congreso de los EE.UU. ocupa actualmente del orden de 15 terabytes y, seguramente el año que viene, podremos llevarla en el bolsillo en un disco duro transportable. Los servidores actuales tiene capacidad de *petabytes* (PB=1.000 TB=10^{15} bytes) y pronto lo tendrán de *exabytes* (EB=1.000 PB =10^{18} bytes). Por ejemplo, una de las bases de datos científicas mayores del mundo, el *World Data Centre for Climate (WDCC)*, almacena unos 400 terabytes de información. La base de datos de Google de varios GB de millones de usuarios supone varios petabytes. La cantidad diaria de datos generados en la actualidad de forma automática se estima de varios exabytes.

El rápido crecimiento de la capacidad de almacenar información ha venido de la mano de la facilidad de acceso gracias a lo que se denomina «la nube» y que consiste en redes de ordenadores conectados a Internet que son accesibles a través de la web. Por ejemplo, los servicios Dropbox, iCloud o Google Drive permiten a los usuarios acceder a sus archivos desde cualquier dispositivo con conexión a Internet.

Estos nuevos datos, el *Big Data*, y su facilidad de acceso y procesado, están ya cambiando todas las parcelas de nuestra actividad: cómo cuidamos nuestra salud, utilizamos nuestro ocio y nos relacionamos (Mayer-Schonberger y Cukier, 2013). También cómo aprendemos, véase Einav y Levin (2014) para su influencia en el análisis económico. Los datos masivos están teniendo una influencia decisiva en la posición relativa de las empresas en cada sector y cambiando las grandes empresas en el mundo. Por ejemplo, a finales del siglo XX las diez mayores compañías mundiales por valor en bolsa pertenecían principalmente a los sectores del petróleo y la fabricación de coches, mientras que en 2019 siete de las diez mayores empresas del mundo están basadas en la combinación de tecnología, *software* y *Big Data* y las cuatro primeras (Microsoft, Apple, Amazon y Alphabet/Google) han sido decisivas en impulsar la revolución de los datos masivos.

Es previsible que la tendencia a crear nuevos datos masivos y a procesarlos con rapidez siga acelerándose en el futuro. En particular, la posible aparición del ordenador cuántico incrementaría de forma gigantesca nuestra capacidad de procesado y crearía nuevas oportunidades en todos los campos, produciendo cambios fundamentales en el mundo tal como lo vemos hoy.

II. LA INFORMACIÓN EN LOS NUEVOS DATOS

1. Codificación para su análisis

Los nuevos datos ofrecen retos cuantitativos y cualitativos. Además, la cantidad de información útil que contienen es muy variable. Unas pocas medidas precisas realizadas en condiciones controladas en un experimento ocupan pocos bytes y pueden tener un alto valor para la predicción de la variable medida. Una foto de alta resolución del experimento, que requiere varios megabytes, puede no contener ninguna información útil para la predicción. En general, la densidad de información de los nuevos datos recogidos automáticamente es mucho menor que la de los obtenidos con un objetivo, pero, a cambio, nos proporcionan información muy desagregada de nuevas situaciones que nunca habíamos imaginado ser capaces de analizar. Por ejemplo, los datos de geolocalización indican la posición de cada persona en el espacio en cada instante, permitiendo fácilmente conocer donde duerme, si trabaja o no y dónde, si acude regularmente a un hospital o centro médico y con qué frecuencia, si frecuenta restaurantes orientales o conciertos de música electrónica, etc. Se ha comprobado que esta información es más precisa que la que cada persona recuerda de sus hábitos y movimientos, que esta filtrada por nuestra memoria selectiva: numerosos experimentos han demostrado que la mayoría de las personas se sorprenden al comprobar la regularidad de sus hábitos y la predictibilidad de sus acciones. (Kahneman, 2012).

Hasta este siglo los datos se recogían con un objetivo específico, generalmente conocer una situación descrita por un conjunto de variables (el funcionamiento de un hospital, de un aeropuerto o la economía de un país) y utilizar este conocimiento para prever valores futuros de algunas de esas variables en función de las restantes (la demora en un tratamiento, la probabilidad de retraso o el crecimiento del PIB). Muchas instituciones financieras, como los bancos comerciales, disponen de conjuntos de variables asociadas a cada cliente, como edad, profesión, saldo en cuenta a final de mes, ingresos recibidos, recibos pagados, etc., que pueden utilizar para construir un modelo explicativo que prevea en función de ellas una variable respuesta de interés, como la probabilidad de mora, de un crecimiento en sus ingresos futuros, etc. Estos datos se agrupaban habitualmente en tablas de filas y columnas que recogen valores numéricos y caracteres de los clientes. Su procesado consiste en convertir las variables cualitativas, descritas por caracteres, en variable numéricas: por ejemplo, la característica sexo se convertía en una variable con dos valores, 1 mujer y 0 hombre. De esta forma, todas las variables se agrupan en una matriz numérica donde, por ejemplo, en filas aparecen las personas y en columnas las variables.

Ahora, además de los datos tradicionales, el banco puede tener acceso a otra información sobre sus clientes, como su actividad en redes sociales, incluyendo fotos, audios y vídeos, o sus datos en otros ficheros de uso público. Estos nuevos datos son un conjunto de información no estructurada que hay que organizar para convertirla en posibles variables explicativas numéricas, que puedan ser objeto de análisis. En general, en lugar de disponer para cada persona de un conjunto pequeño de variables, tendremos un conjunto amplio de matrices numéricas relacionadas entre sí. Por ejemplo, una imagen digital es una matriz formada por celdas, que llamamos píxeles, donde en cada una de ellas se ha definido un color. Una imagen en blanco y negro contiene en cada celda un solo número, la intensidad de gris. Una en color con el sistema RGB (*red, green, blue*) contiene tres números, las intensidades de rojo, verde y azul, que al combinarse generan todas las tonalidades de colores. Por tanto, la representación de una imagen en color requiere tres matrices, cada una con tantas celdas como píxeles tenga la imagen, que se combinan con la intensidad definida en cada matriz para formar el color del píxel correspondiente. De la misma forma, un audio se codifica con números que representan intensidad de distintas frecuencias de la voz humana a lo largo del tiempo. Con los nuevos datos a cada persona podemos asignarle un gran número de esas matrices.

Cómo resumir toda esta gigantesca información en un número manejable de variables que puedan analizarse es un problema complejo. En la práctica, se seleccionan de todos los datos disponibles los rasgos o variables que parecen *a priori* prometedores, creando un número de variables explicativas que suele ser muy amplio. Además, es habitual encontrar heterogeneidad: la relación entre las variables puede no ser la misma para personas con distintos hábitos que se manifiestan en sus pautas de consumo o en su actividad en las redes sociales y debemos construir modelos distintos para clientes diferentes. Es frecuente con datos muy desagregados, como los que estamos hablando, que la relación entre dos variables puede depender de una

tercera creando interacción, o una relación no lineal entre ellas, lo que hace más complejo su análisis.

Una diferencia fundamental entre los datos clásicos y los nuevos es que, en el pasado, la distinción entre elementos a estudiar y variables era muy clara, con muchos elementos observados y unas pocas variables a estudiar en cada uno de ellos. Con los datos masivos, esa distinción se hace más ambigua e incluso, se pierde. Por ejemplo, al analizar el genoma humano nos encontramos con miles de datos de genes de pocas personas, cuando, hasta entonces, lo habitual eran unas pocas variables de muchas personas. Tiene ventajas ver el problema como una población de muchos genes, que son los elementos de la población, que observamos en ciertas variables, que son las personas, dando la vuelta al análisis convencional. En el campo financiero suponemos que analizamos muchas variables económicas y financieras de un pequeño núcleo de clientes clave en un banco. Podemos considerar que los elementos son las muchas medidas que caracterizan su comportamiento y las variables las personas estudiadas. Los nuevos tipos de datos pueden, según los objetivos del análisis, clasificarse en distintas dimensiones y seleccionar la más eficaz para analizarlos.

2. Extraer la información del *Big Data* con estadística

El objetivo de analizar los datos es convertirlos en información. Esta información aumenta nuestro conocimiento de las variables relevantes y las relaciones entre ellas y puede utilizarse para tres objetivos fundamentales. El primero, que suele denominarse *análisis descriptivo*, consiste en resumir las variables en unos pocos indicadores que nos sirvan para entender su estructura y las relaciones que hay entre ellas. Con este objetivo se intenta obtener conocimiento de las variables y datos presentes. El segundo objetivo se denomina *aprendizaje no supervisado* o métodos de clúster, y consiste en encontrar grupos de elementos o variables que tienen un comportamiento similar en los datos disponibles. Se denomina no supervisado porque se aplica a situaciones donde queremos dividir los datos en grupos homogéneos, pero no tenemos *a priori* ninguna información sobre el número de los grupos presentes ni de las variables que son importantes para definir estos grupos. El tercer objetivo, *aprendizaje supervisado*, consiste en prever una variable, o un conjunto de variables relacionadas entre sí, en función de otras muchas. Además, estos objetivos pueden ser estáticos, si los datos se refieren a un instante fijo de tiempo, o dinámicos, si disponemos de la evolución de las variables en el tiempo y/o en el espacio.

Estos tres objetivos se han estudiado tradicionalmente por la estadística, poniendo el énfasis en la comprensión de las relaciones a través de la modelización de relaciones causales. La aparición de nuevos datos no convencionales (imágenes, señales, funciones, etc.) para resolver nuevos problemas (reconocimiento de lenguaje, visión artificial, etc), donde es muy complejo entender la relación entre las variables y los nexos casuales ha generado un conjunto de métodos adicionales de *inteligencia artificial* y *aprendizaje automático* encaminados a explotar las correlaciones encontradas entre las variables para mejorar la predicción.

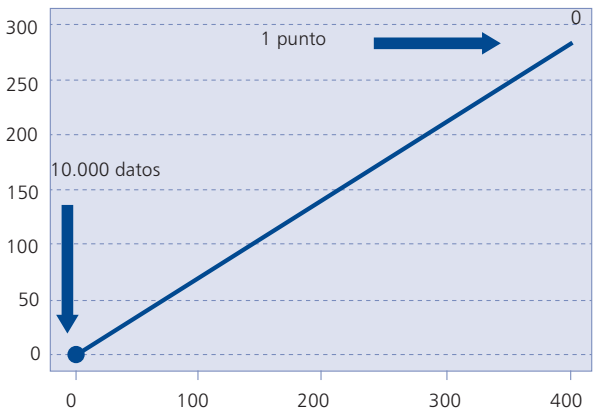
La estadística (*Statistics, ST*) ha sido el motor del avance en el conocimiento empírico en el siglo XX. Su origen es relativamente muy reciente, ya que aunque el cálculo de probabilidades aparece como parte de la matemática en el siglo XVII, el análisis de datos para inferir propiedades de una población se hace posible cuando R. A. Fisher crea, en el primer tercio del siglo pasado, las bases de la inferencia estadística. Debemos también a Fisher los principios y métodos del diseño de experimentos para aprender y contrastar hipótesis en situaciones donde, fijando las variables de control, la respuesta observada tiene variabilidad. Durante la segunda mitad del siglo XX la estadística aprovecha la llegada del ordenador para construir nuevos métodos basados en el cálculo intensivo para hacer inferencias, como el *Bootstrap*, creado por Efron en 1979 o los métodos de Montecarlo, para simular datos y estimar modelos complejos. La estadística ha desarrollado en el siglo XX modelos muy eficaces para la predicción con datos agregados y bien estructurados cuando el número de elementos observados, n, es mucho mayor que el número de variables, p. Los métodos estadísticos tratan de modelar relaciones de causalidad entre las variables, por lo que sus modelos tienen la ventaja de una fácil interpretación. Por ello, se han aplicado en todos los campos del conocimiento y de forma muy destacada en economía, con técnicas específicas para variables económicas que han creado la econometría. En el campo de las finanzas tanto la estadística como la econometría han creado herramientas de análisis cuantitativo que han hecho avanzar nuestro conocimiento de los procesos financieros y de las decisiones de inversión.

La aparición en el siglo XXI de los datos masivos asociados a nuevos problemas produjo inicialmente un «espejismo del tamaño», es decir, el error frecuente de creer que si la dimensión de los datos aumenta, con más casos y variables, podemos aplicar, con pequeños ajustes, los mismos métodos desarrollados para pocos datos a los nuevos masivos. La ciencia nos ofrece muchos ejemplos de que esta extrapolación puede no funcionar. Por ejemplo, es bien conocido que al aumentar la velocidad de un objeto y aproximarse a la de la luz, la física clásica deja de ser aplicable y tenemos que utilizar las ecuaciones de la relatividad, donde el tiempo es relativo y depende de la velocidad. En el mismo sentido, al descender a la escala microscópica aparecen las propiedades cuánticas y principios como el de superposición, donde una partícula puede estar simultáneamente en dos posiciones distintas con cierta probabilidad, que desaparecen al aumentar el tamaño. Estos principios cuánticos, únicamente observables a escala de nanómetros, hacen posible la computación cuántica que puede revolucionar los métodos de computación futura. Como ejemplo más cercano a nuestra experiencia, un medicamento tomado en pequeñas dosis nos ayuda a dormir y en dosis elevadas nos produce la muerte. Hoy es generalmente aceptado que necesitamos nuevos métodos para algunos de los nuevos problemas y que la colaboración entre científicos de distinto origen (estadística, computación e ingeniería) es fundamental en la nueva disciplina de *data science* o ciencia de los datos (véase Peña, 2014 y 2015 y Galeano y Peña, 2019).

Los nuevos datos permiten abordar nuevos problemas, pero requieren también nuevos procedimientos. En primer lugar, y como hemos comentado, el número de variables explicativas posibles que podemos crear con ellos es enorme, y resumirlos en variables utilizables plantea serios problemas estadísticos. Antes, la estadística comenzaba analizando unos datos dados; ahora, debemos utilizarla para crearlos. Por ejemplo, supongamos que queremos prever una variable, por ejemplo, si un cliente va a devolver o no un crédito solicitado, y, además de la información habitual, disponemos de toda su historia en el banco y la información pública de su actividad en redes sociales. Resumir en variables operativas su trayectoria financiera descrita en un conjunto de series temporales es ya un problema complejo, como lo es también obtener de las redes sociales variables relevantes. Por ejemplo una variable podría ser la proporción de sus amigos que aparecen en ficheros de morosos, pero hay también otras muchas variables posibles que podría construirse a partir de esta información. Además, estas variables pueden estar relacionadas entre sí, y para analizarlas conjuntamente podríamos construir una variable indicadora, combinación lineal de todas ellas. Por ejemplo, el IPC es una manera de resumir los gastos en distintos conceptos asignando a estas variables unos pesos que son su importancia en la cesta de la compra promedio. En general, no disponemos de los pesos que tenemos que dar a cada variable de un conjunto para definir una buena variable indicadora y estos pesos deben determinarse empíricamente a partir de los datos.

Un segundo problema es el de la heterogeneidad: la relación entre las variables puede no ser la misma para personas con distintos hábitos que se manifiestan en sus pautas de consumo o en su actividad en las redes sociales y debemos construir modelos distintos para clientes diferentes. Además, la relación entre dos variables puede depender de una tercera creando interacción o una relación no lineal entre ellas. Finalmente, los datos recogidos pueden tener errores de medida. Es frecuente que los sensores fallen por diversas razones (veáse, Paradis y Han, 2007, para una panorámica de este problema) y sus fallos producen valores atípicos en los datos. Podría pensarse que los datos atípicos no son importantes cuando tenemos muchos datos, pero desgraciadamente no es así: podemos tener un millón de datos de dos variables y su correlación puede depender de un solo dato con un error de medida suficientemente grande. Esta situación se ilustra en el gráfico 1, donde la relación de dependencia entre las variables

GRÁFICO 1
UN ÚNICO PUNTO CREA LA CORRELACIÓN ENTRE DOS VARIABLES AUNQUE HAYA DIEZ MILLONES DE DATOS



se genera por un único punto y se pone de manifiesto la necesidad de limpiar cuidadosamente los datos antes de cualquier análisis.

Los nuevos procedimientos para grandes masas de datos suelen agruparse bajo el nombre de *aprendizaje estadístico* (véase Hastie, Tibshirani y Friedman, 2009), y en su desarrollo la Universidad de Stanford ha tenido un papel de liderazgo. Estos nuevos métodos se diferencian de los tradicionales en: 1) utilizan criterios automáticos de selección de modelos y validación fuera de la muestra; 2) combinan muchos modelos simultáneamente para la predicción; 3) incluyen procedimientos iterativos de aprendizaje para mejorar el modelo con nuevos datos; 4) consideran relaciones no lineales entre las variables de tipo muy amplio, incluyendo estructuras en forma de árbol como los árboles de decisión (*Classification and Regression Trees, CART,* por sus siglas en inglés) y los bosques aleatorios (*random forest*). Sus resultados compiten en los últimos años con procedimientos desarrollados en *aprendizaje automático* (*machine learning*) y en *inteligencia artificial*, que comentaremos en el siguiente apartado.

3. *Aprendizaje automático* e inteligencia artificial

Se denomina *inteligencia artificial* (*artificial intelligence, AI*) la rama de las ciencias de la computación que desarrolla procedimientos para que las máquinas puedan comportarse con inteligencia humana. Por ejemplo, reconociendo el lenguaje y la visión y reaccionando en consecuencia, aprendiendo y resolviendo problemas. La robótica construye máquinas (robots) basados en estos métodos. Se denomina *aprendizaje automático* o *aprendizaje máquina* (*machine learning, ML*) a la parte de *AI* que desarrolla algoritmos para la predicción de una o varias variables utilizando una amplia base de datos observados que contiene valores de las variables a prever y de las variables explicativas que pueden utilizase para la predicción. Aunque este objetivo es similar al que buscan los modelos estadísticos hay una diferencia importante: *ML* busca la mejor predicción de acuerdo con los datos observados, es decir, utiliza las correlaciones observadas de muchas variables para obtener buenas predicciones. Sin embargo, la estadística trata de construir un modelo comprensible, que incluya relaciones causales y que funcione bien no solo en los datos observados, sino en otros que puedan generarse en el futuro. Esta diferencia indica las ventajas y limitaciones de cada enfoque.

Los algoritmos de *ML* utilizan para construir la regla de predicción una técnica estadística llamada validación cruzada (*cross-validation*), que consiste en dividir los datos disponibles en dos partes. La mayor parte de ellos, unos 2/3 del total, se utilizan en una muestra de entrenamiento para estimar distintas reglas de predicción. Los datos no utilizados forman la muestra de validación, que se utiliza para seleccionar el mejor modelo entre los estimados en la muestra de entrenamiento. Como el resultado obtenido puede depender de la partición realizada, el proceso de división de la muestra en dos partes suele repetirse de nuevo y promediar los resultados tomando la regla que funciona mejor en promedio. Naturalmente, el tiempo de cálculo para seleccionar la regla de predicción es mucho mayor que en los modelos estadísticos tradicionales, pero en algunos casos los resultados son más precisos.

La forma de la regla de predicción de la variable respuesta en función de las variables explicativas (también llamadas variables *input* o variables de entrada) puede ser muy variada (veáse Bishop, 2006). El método más utilizado habitualmente son las redes neuronales (*neural networks, NN,* por sus siglas en inglés), que son capaces de aproximar cualquier relación posible entre la respuesta y los *inputs*. Un problema frecuente de la regla obtenida es que la relación entre cada variable y la respuesta es muy compleja y no puede determinarse de forma simple. Por eso se denomina a estas reglas de predicción de «caja negra», ya que no es posible conocer fácilmente el efecto de cada variable explicativa en la respuesta. Además, el efecto de una variable puede depender de los valores de otras variables relacionadas con ella, con lo que comprender el efecto total de cada variable o grupos de variable requiere técnicas estadísticas de diseño de experimentos muy complejas (Galeano y Peña, 2019).

4. Implantación de los métodos de análisis

La implantación efectiva de los nuevos métodos de análisis requiere la infraestructura y el *software* necesarios para llevar adelante estos cálculos. Desde el punto de vista del *software* destacaremos la existencia de sistemas potentes para el cálculo en paralelo y el desarrollo de lenguajes de programación en código abierto, que facilita a los inventores de nuevos métodos programarlos y ponerlos al alcance de todos los interesados.

Los programas de cálculo estadístico han evolucionado de los trabajos por lotes a los programas

interactivos que permiten acceso directo a los resultados intermedios y capacidad de programación dentro del paquete (como en SAS, S+, SCA, MATLAB, GAUSS, etc.). Un avance fundamental han sido los lenguajes en código abierto orientados a objetos, que permiten manejar indistintamente funciones, variables o gráficos. La aparición de los lenguajes R y Python, nacidos en los años noventa, ha creado un estándar donde cientos de investigadores de todo el mundo incorporan nuevas rutinas ampliando cada día las capacidades de análisis. Véase Ugarte, Militino y Arnholt (2005) para el uso de R en estadística y Pedregosa *et al*. (2011) para Python en *machine learning*. Esto ha dado a ambos lenguajes una ventaja imbatible frente a otros que no se enriquecen continuamente por los nuevos paquetes escritos por miles de investigadores en todo el mundo. Los dos lenguajes pueden integrarse con distintas bases de datos e incorporan cálculos en paralelo, necesarios con las grandes bases de datos actuales.

Los datos masivos requieren computación en paralelo y distribuida, y el almacenamiento en la nube. El cálculo en paralelo consiste en ejecutar muchas instrucciones simultáneamente. Esto exige una programación donde un cálculo secuencial se descompone en partes que pueden realizarse en paralelo, sea en *hardware* con procesadores con varios núcleos o con varios procesadores que se comunican entre sí. Esta forma de trabajos muestra su potencia cuando se conectan varios ordenadores a la web, en general de forma remota sin proximidad física, formando un clúster, es decir un grupo de ordenadores conectados entre sí mediante un sistema de red de alta velocidad. Además, debe existir un programa que controle la distribución de la carga de trabajo entre los equipos. Por lo general, este tipo de sistemas cuentan con un centro de almacenamiento de datos único.

Una infraestructura digital en código abierto, dentro de la licencia de la Fundación APACHE, es Hadoop, creado por Doug Cutting. Hadoop combina la computación en paralelo y distribuida permitiendo desarrollar tareas muy intensivas de computación dividiéndolas en pequeñas partes y distribuyéndolas en un conjunto tan grande como se quiera de máquinas. Al ser de licencia libre este *software* está siendo adoptado no solo por usuarios particulares sino también por grandes sistemas (Oracle, Dell, etc.), lo que está llevando a una aceleración tanto de su difusión como de sus prestaciones. A diferencia de las soluciones anteriores para datos estructurados, la tecnología Hadoop introduce técnicas de programación nuevas y más accesibles para trabajar en almacenamientos de datos masivos con datos tanto estructurados como no estructurados.

III. OPORTUNIDADES PARA LAS INSTITUCIONES FINANCIERAS

A continuación vamos a comentar algunas aplicaciones donde el uso del *Big Data* puede suponer ventajas para las instituciones financieras.

1. Mejorar las predicciones mezclando distintos tipos de información: *nowcasting*

Tradicionalmente, la obtención de predicciones ha sido llevada a cabo utilizando información proporcionada de manera periódica, que incluye la generada por institutos nacionales o internacionales de estadística. De esta manera, para obtener predicciones, por ejemplo del producto interior bruto (PIB), es muy habitual utilizar modelos de series temporales tales como modelos vectoriales autorregresivos (VAR) o modelos de regresión dinámica, donde el futuro de una variable se explica utilizando, además de su historia, otro conjunto de series temporales, con la misma frecuencia o periodicidad y el mismo período de observación. Recientemente, muchas instituciones financieras han comenzado a utilizar para hacer predicciones sobre sus clientes otro tipo de información más compleja y que mezcla una amplia variedad de bases de datos estructuradas y no estructuradas. Esta información puede incluir desde el uso que los clientes hacen de Internet, las búsquedas en Internet más frecuentes, su actividad en redes sociales como Twitter o Instagram, o información sobre transacciones entre diferentes clientes de la institución. Estos datos tienen la ventaja de que, en la mayoría de los casos, pueden obtenerse en tiempo real, permitiendo la actualización instantánea de las predicciones.

Para realizar predicciones en este nuevo entorno, ha aparecido una nueva área de investigación llamada *nowcasting* que permite combinar series temporales con diferentes periodicidades con el objetivo de mejorar las predicciones. Su nombre es una contracción de *now* (ahora) y *forecasting* (predicción). *Nowcasting* se desarrolló en meteorología para predecir el clima futuro a muy corto

plazo (tres o seis horas, por ejemplo). Pretende pronosticar, con alta fiabilidad, determinados fenómenos meteorológicos, como puede ser una tormenta en una localización geográfica concreta. En economía, Giannone, Reichlin y Small (2008) utilizaron dicha técnica con el objetivo de evaluar el impacto marginal que las publicaciones de datos intramensuales tienen en los pronósticos del trimestre actual del crecimiento del PIB real. De esta manera, se utiliza información actual para predecir a un breve horizonte temporal. Un aspecto importante es que *nowcasting* permite actualizar las predicciones cada vez que se obtienen nuevos datos de forma no sincronizada. Para una revisión extensa de este campo, referimos a Kapetanios y Papaillias (2018), quienes presentan metodologías para realizar *nowcasting*, que incluyen: (1) métodos basados en *machine learning*, como regresiones penalizadas, regresiones *sparse* y métodos *boosting*: (2) (optimización) entre otros, optimización heurística de criterios de información, como *simulated annealing* o algoritmos genéticos: y (3) métodos basados en reducción de la dimensión, como componentes principales y sus versiones *sparse*, y métodos textuales.

Desarrollar nuevos procedimientos que permitan mezclar información procedente de nuevas fuentes, como textos, imágenes o vídeos, con información más tradicional es una necesidad urgente. Un ejemplo en finanzas lo podemos encontrar en Chen *et al.* (2014), quienes, para pronosticar el mercado de valores, combinaron información estándar con información obtenida en webs financieras. De manera similar, Bartov, Faurel y Mohanram (2018) demostraron que no solo la actividad de Twitter predice el mercado de valores en su conjunto, si no que la opinión agregada de los *tweets* individuales predice con éxito las próximas ganancias trimestrales y los retornos de anuncios de una empresa. Otro ejemplo lo podemos encontrar en investigación de mercados y en *marketing* donde empieza a ser necesario combinar la información más tradicional sobre los clientes, como su género o su edad, obtenida mediante el uso de cámaras y micrófonos que captan las reacciones de varios clientes ante un dependiente o por un producto. Similarmente, las compras realizadas por un cliente, por su frecuencia, y/o las similitudes con las hechas por otros clientes están guiando los llamados sistemas de recomendación, tan habituales ya en compañías de comercio electrónico tales como Amazon, AliExpress o Ebay, para hacer ofertas atractivas de productos que los clientes tienen altas probabilidades de comprar. Por ejemplo, en Amazón, líder de estos sistemas de recomendación, un tercio de sus ventas se atribuye a su efecto (Mayer-Schonberger y Cukier, 2013).

2. Predicciones personalizadas y geolocalización

El nuevo paradigma del *Big Data* está llevando a muchas compañías e instituciones a utilizar técnicas para personalizar sus predicciones. Por ejemplo, cada vez es más frecuente que un banco decida promocionar un producto a un determinado espectro de la población que tenga unas características específicas, como una hipoteca con unas condiciones ventajosas a todos sus clientes de entre 25 y 45 años, ya que en ese rango de edades es cuando la gran mayoría de personas decide comprarse una vivienda. Sin embargo, si solo se tiene en cuenta la edad es posible que la oferta publicitaria recibida por el cliente no sea de su interés, perdiendo la oportunidad de ofrecerle un producto más adecuado. Por ejemplo, una persona soltera que esté realizando estudios de posgrado, y que encaja perfectamente en ese rango de edades, posiblemente esté más interesada en un pequeño crédito que le permita comprar un equipo informático o realizar un viaje de vacaciones.

Conocer muchos datos de cada uno de los clientes puede ayudar a predecir mejor determinados comportamientos de consumo. En general, el éxito de las ofertas se incrementará al añadir aspectos que permitan realizar predicciones más personalizadas.

Un ejemplo de información relevante es la obtenida a través de la geolocalización que proporciona la ubicación exacta de un teléfono móvil, tablet o equipo informático. En base a ciertas coordenadas geográficas. El sistema de geolocalización más habitual es el denominado Sistema de Posicionamiento Global (*Global Positioning System, GPS,* por sus siglas en inglés). Este sistema es el que permite a determinadas aplicaciones que se pueden descargar desde cualquier teléfono móvil como Google Maps o Waze, conocer la ruta más rápida entre dos puntos geográficos concretos o planear rutas alternativas si durante el viaje, la ruta inicial deja de ser la más rápida debido a un atasco o un accidente de tráfico. El *GPS* no es el único sistema de geolocalización. Por ejemplo, se puede identificar la situación de un equipo informático a través de una

dirección de protocolo de Internet (*Internet Protocol , IP*), dirección de control de acceso a medios (*media access control, MAC*), o de otros sistemas de posicionamiento inalámbrico.

Conocer la geolocalización de un cliente es muy relevante en un gran número de aspectos. Un primer ejemplo es el de conocer cuándo un pago con una tarjeta de crédito es fraudulento o no. Por ejemplo, si un cliente realiza una compra desde una dirección *IP* de un país extranjero y el banco no tiene constancia de que el cliente esté en dicho país, por haber comprado un billete de avión con tarjeta de crédito o pagado una cuenta en el aeropuerto de la residencia habitual del cliente, lo más probable es que el banco determine que dicha compra es fraudulenta. La geolocalización del teléfono móvil también puede ser útil en este sentido ya que realizar un pago en un lugar diferente a donde está situado el teléfono móvil puede ser motivo para declarar la compra como fraudulenta. Un segundo ejemplo es el uso de la geolocalización de un cliente para conocer su entorno: donde vive, donde trabaja, donde va de vacaciones, etc. Esta información puede ser muy útil para ofertar productos de un determinado interés local u ofertas localizadas en lugares cercanos a los habituales lugares de movimientos del cliente. Un tercer ejemplo es que un banco puede dirigir a un cliente a la sucursal más próxima para la realización de trámites una vez se conecte a la banca *online* a través de su teléfono móvil. Un cuarto ejemplo es que un banco puede aplicar determinadas restricciones geográficas para la realización de operaciones o una compañía de venta *online* puede aplicar restricciones similares para el envío de determinados productos entre países para proteger los derechos de distribución.

3. Predicción del abandono de clientes (*churn or cutomer loyalty prediction*)

Debido al actual alto grado de competencia en todos los sectores, tan importante o más que intentar ganar clientes es retener los actuales. Su pérdida implica la ganancia de otros, y nuestra disminución de cuota de mercado. Además, un cliente enfadado difunde su descontento en redes sociales con mayor probabilidad que uno satisfecho, con efectos muy negativos sobre la imagen de la empresa. El término en inglés *churn rate* se refiere a la tasa de pérdida de clientes de una compañía o institución. Habitualmente, esta pérdida es pequeña, es decir, son clientes aislados que dejan de utilizar los servicios de la compañía o institución para, o bien irse a la competencia, o bien dejar de utilizar el servicio para siempre. Claramente, es de interés de las compañías entender qué factores hacen que los clientes tomen dicha decisión para predecir cuando un cliente dejará de serlo (*churn prediction*).

Existe un número importante de procedimientos para predecir el abandono de clientes mediante el uso de técnicas de *machine learning*, que utilizan tanto información propia de las compañías, como información externa. Por ejemplo, De Bock y Van den Poel (2011) comparan métodos de clasificación supervisada de clientes activos y clientes perdidos de varias compañías, incluyendo *random forest* y sus variantes como *AdaBoost*, con otras técnicas para la extracción de características, incluyendo análisis de componentes principales, análisis de componentes independientes y proyecciones aleatorias. Estos autores ilustran que el uso de combinaciones de clasificadores parece determinar muy bien qué características de los clientes son las que mejor determinan el abandono. Benoit y Van den Poel (2012) investigan el uso de información proporcionada por redes de clientes de servicios financieros para la retención de los mismos. Para ello, los autores investigan si además de los conjuntos convencionales de variables, sociodemográficas, historial de compras, etc., las variables basadas en la red de clientes mejoran el poder predictivo de modelos de retención de clientes. La principal conclusión obtenida es que el poder predictivo de un modelo de abandono se puede mejorar agregando variables basadas en redes sociales. Por ejemplo, la inclusión de determinadas características de una red, por ejemplo, el número de relaciones de un cliente, la importancia del cliente dentro de la red o la densidad de la propia red, aumenta la precisión predictiva. De hecho, las variables basadas en la red pueden tener mayor impacto en discriminar clientes que abandonan de los que no, que el resto de variables. Este y otros aspectos relacionados con redes de clientes serán tratados con mayor detalle en la siguiente subsección y en el ejemplo con datos reales de la sección cuarta.

Por último, Burez y Van den Poel (2008) demostraron que se puede predecir con mayor precisión la pérdida de clientes que se van a la competencia, o dejan de utilizar el tipo de servicios ofertados por la compañía, que su pérdida por morosidad.

4. Utilizar la red de clientes para orientar políticas

La red de clientes de una institución financiera, como puede ser un banco, es una fuente de información muy importante en varios aspectos. En primer lugar, es una pieza fundamental para entender la importancia de sus usuarios analizando su posición en dicha red. Tradicionalmente, se ha tendido a considerar que la importancia de un cliente se mide a partir de sus recursos económicos. Esta medida ignora su importancia estratégica dentro de la institución y su posición y relaciones en el conjunto de la red de clientes. La importancia de un cliente depende, claramente, de las consecuencias para el banco de que deje de serlo.

En el análisis estadístico de redes encontramos varias maneras de medir la importancia de un vértice (cliente en este caso) en función de sus relaciones con otros vértices (otros clientes) para medir el efecto de que un vértice desaparezca de dicha red. Por ejemplo, una primera opción es considerar que un cliente puede ser importante si está relacionado con muchos otros clientes. Esto es relevante para un banco ya que este cliente puede tener una influencia importante sobre muchos otros y puede actuar como elemento transmisor de los productos del banco. Si el cliente dejara de serlo, se perdería dicho elemento transmisor. Una segunda opción es que lo sea si está cerca de muchos otros clientes. En este caso no sería necesario que el cliente tenga relación directa con muchos otros, sino que bastaría con que cualquier camino que debería recorrer el cliente para llegar a cualquier otro de la red fuese pequeño. De nuevo, este tipo de clientes deberían ser relevantes ya que si desapareciesen la distancia entre clientes aumentaría. Una tercera opción es que sea importante si es necesario cruzar por él frecuentemente para conectar unos clientes con otros. De nuevo nos encontramos con un cliente que puede actuar de catalizador entre muchos otros. Si el banco quiere ofertar un nuevo producto, este tipo de clientes son especialmente relevantes para extender la oferta entre muchos otros clientes.

En segundo lugar, la red de clientes es una pieza fundamental para captar la presencia de grupos de clientes cohesionados (con fuertes relaciones entre ellos). Conocer la presencia de estos grupos es muy relevante para orientar determinadas políticas. Por ejemplo, si un banco conoce un conjunto de clientes cohesionados, se podrían analizar las características específicas de dicho grupo (por ejemplo, si son todos padres de niños de un mismo colegio o de un club deportivo, o un grupo de personas que realizan actividades conjuntas), y crear productos específicos apropiados para ellos. Sería también de interés localizar los clientes más importantes dentro de la subred creada por ese grupo para conocer que cliente sería idóneo para la posterior propagación de dichos productos.

Determinar la presencia de grupos de clientes fuertemente enlazados se conoce con el nombre de detección de comunidades. El problema consiste en obtener una partición de la red en varias subredes de tal manera que los miembros de cada una de las subredes tuviesen una fuerte relación entre ellos y escasa con los miembros de otras subredes. Una medida de la calidad de una partición es la modularidad que se define como la suma para cada una de las subredes de las diferencias entre el número de relaciones dentro de la subred con respecto al número esperado de relaciones si no hubiese tal estructura de comunidades. Cuanto mayor sea el valor de la modularidad de una red, más significativa será la presencia de los grupos. Existen diferentes algoritmos que tratan de proporcionar particiones de la red con máxima modularidad. Posiblemente el más popular de ellos es el conocido como método Louvain (véase Blondel *et al*., 2008), debido a que fue propuesto por varios investigadores de la Universidad Católica de Lovaina. El método consiste en resolver el problema de maximizar la modularidad mediante un método conocido como *greedy optimization*. La ventaja fundamental de este procedimiento sobre otros propuestos, ver Clauset, Newman y Moore (2004) y Brandes *et al.* (2008) entre otros, es que el problema de optimización se puede resolver para un número muy grande de vértices y relaciones, tan grande como varios millones de cada una de ellas. Otro tipo de procedimientos algo menos utilizados es el uso de clustering jerárquico, muy habitual en el análisis de datos multivariantes. La idea de este tipo de procedimientos en redes, denominados algoritmos *walktrap* (véase Pons y Lapaty, 2005), es calcular distancias entre diferentes vértices (clientes) y aplicar clustering jerárquico sobre estas distancias para poder crear los grupos.

En tercer lugar, la red de clientes también puede ser utilizada para explicar y predecir determinados comportamientos de los clientes. Un ejemplo ilustrativo es la mora. Explicar las razones por las que los clientes entran en mora es *a priori*

un problema complicado. Una opción es utilizar la información personalizada del cliente, como su edad, su estado civil, su salario y el total familiar, su situación laboral, su lugar de residencia, la deuda contraída con el banco, los productos contratados por el mismo, etcétera. Sin embargo, la situación del cliente dentro de la red puede ser un aspecto fundamental para explicar las razones que hacen que un cliente entre en mora. Supongamos, por ejemplo, el trabajador de una empresa que entra en suspensión de pagos. Si el trabajador tiene una hipoteca con el banco, muy probablemente en los próximos meses entrará en mora ya que no recibirá su salario. Evidentemente otros factores pueden ser importantes, como los mencionados previamente, pero este hecho fundamental puede ser determinante para la entrada en mora del trabajador. El banco dispone de dicha información, puede utilizar mecanismos para evitarlo. En general, un aspecto importante para predecir la entrada en mora de un cliente es sus relaciones con otras personas en mora o que pueden estarlo a corto plazo. Para ello, es necesario que los analistas del banco realicen un análisis temporal de los datos. Por ejemplo, se pueden tomar diferentes momentos temporales de la información de los clientes, por ejemplo un mes, incluyendo su situación dentro de la red de clientes y cuáles de los clientes en dicha red están en mora. Con esta información, se pueden construir procedimientos de clasificación supervisada que permitan explicar qué variables son las más relevantes para explicar la mora de los clientes y con esta información construir procedimientos estadísticos que permitan predecir dicha mora a corto o medio plazo. Un ejemplo de este tipo de análisis se puede encontrar en la sección cuarta de este artículo donde podremos comprobar la efectividad de este tipo de análisis.

En cuarto lugar, una red formada por clientes y por no clientes de un banco puede ser útil para la captación de nuevos clientes. Para ello, un gestor de un banco puede contactar con uno de los clientes de su cartera y a través de quien podría contactar consecutivamente con otros clientes del banco que no estén en su cartera y que le permitan tener las mejores opciones para poder contactar y convencer a un potencial nuevo cliente. Para ello, el gestor debe tener algún tipo de herramienta que le permita construir el camino que le lleve al éxito con mayor probabilidad de entre todos los posibles caminos construidos a partir de las relaciones entre clientes. Para ello, una opción es determinar la probabilidad de tener éxito en cada una de las etapas del camino, es decir, de tener éxito en contactar a un cliente a partir de su relación con otro. Con todas estas probabilidades calculadas el camino óptimo de entre todos los posibles puede ser aquel que maximice el producto de las probabilidades asociadas a cada una de las relaciones. El cálculo de estas probabilidades puede depender de un número importante de aspectos como pueden ser la confianza o el vínculo entre los clientes, la satisfacción de los clientes con el banco, la influencia que puedan tener los clientes involucrados y la personalidad o el grado de conformidad de los mismos, entre otros aspectos, ver Quijano-Sánchez y Liberatore (2017). Evidentemente, estos conceptos son subjetivos y deben ser cuantificados de alguna manera. Para ello, es posible llevar a cabo una cierta estimación de cada uno de estos aspectos con la información que el banco posee de los clientes. Un ejemplo práctico de este tipo de análisis junto con una propuesta para el cálculo de las probabilidades de éxito basados en los aspectos descritos previamente se puede encontrar en el ejemplo de la sección cuarta.

Por último, dadas las múltiples relaciones en una red de clientes es necesario tener en cuenta todas las características de los clientes. No puede ser igual una relación consistente en un pago puntual por la venta de un coche de segunda mano, que otra que implica transferencias periódicas por una cantidad importante. Las relaciones entre clientes deben ponderarse de acuerdo a los objetivos del análisis de la red. Por ejemplo, hay que considerar: 1) el tipo de clientes que definen la relación, es decir, cuando los clientes son personas, entidades, representantes, etc.; 2) la dirección de la relación, es decir, si una empresa paga a un empleado, o un arrendado paga un alquiler al arrendatario, etc.; 3) el número de relaciones entre los clientes, es decir, si existe un único movimiento o existe un número repetido de movimientos; y/o (4) la cantidad económica transferida. Un ejemplo de estas ponderaciones se presentará en la sección cuarta.

5. Prevenir y detectar el fraude

El aumento de las operaciones de pago con medios digitales, tarjetas, teléfonos móviles, etc. ha generado nuevas formas en que los delincuentes pueden cometer fraudes. Cómo prevenir y detec-

tar fraudes financieros ha sido un amplio campo de investigación en los últimos veinte años. Véase Bolton y Hand (2002), Kou *et al*. (2004), Phua *et al*. (2010) y Abdallah, Maarof y Zainal (2016), para una revisión de los tipos de fraude y los métodos para prevenirlos y detectarlos.

Los dos tipos de fraude financiero más importantes en la actualidad son el fraude a través de compras con tarjeta de crédito y mediante la utilización fraudulenta de la cuenta bancaria. La utilización de las tarjetas de crédito ha crecido mucho en los últimos tiempos y también el fraude, aunque el uso creciente de técnicas de prevención y detección parecen haber detenido ese avance: El Banco Central Europeo en su último informe de 2018 (https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport201809.en.html), estima que este fraude ha sido en 2016 de 1,8 billones de euros en el *Single Euro Payments Area*. El fraude estuvo aumentado fuertemente desde 2012 a 2015, pero se ha estabilizado en los últimos años, representando un crecimiento del 14,8 por 100 comparado con el de 2011. Respecto a la composición del fraude, el más frecuente (73 por 100) corresponde a pagos no presenciales realizados por Internet o teléfono, 19 por 100 a pagos con la tarjeta en la venta y 8 por 100 a transacciones al sacar dinero con la tarjeta. Además, los pagos por medios no tradicionales, como el teléfono móvil o Internet, suponen ya el 60 por 100 de fraude cuando en 2008 era del 46 por 100. El mercado ha empezado a desarrollar herramientas para la prevención del fraude y su detección, al mismo tiempo que los gobiernos han revisado la normativa sobre las transacciones electrónicas (en la Unión Europea el *Payment Services Directive* (PSD2) de 2015 y el *Regulatory Technical Standards* en 2017).

Los primeros métodos utilizados a finales del siglo pasado para la detección del fraude fueron métodos de discriminación o de aprendizaje supervisado, en la terminología de *machine learning*. Estos métodos requieren tener una base de operaciones que se sabe han sido correctas y otra base de datos de operaciones fraudulentas. Ambos grupos se caracterizan por un conjunto de variables y se calcula la media de estas variables en ambos grupos. Cuando se observa una operación se calcula su distancia a ambos grupos y se clasifica en el más próximo. La comparación puede hacerse respecto a las medias, como en la discriminación lineal o logística, o respecto a todos los miembros de cada grupo, como los métodos no paramétricos de vecinos más próximos. En este siglo se introdujeron métodos no lineales con variables cualitativas como los árboles de decisión (*CART*), los bosques aleatorios (*random forest*), las redes neuronales (*NN*) y las máquinas de vector soporte (*vector support machines*). Los bancos no proporcionan información precisa de los métodos utilizados pero uno muy popular es el desarrollado por FICO (Fair&Isaac Cooporation), que asegura que su *Falcon Fraud Manager*, basado en *NN*, se utiliza en más de la mitad de las transacciones mundiales (véase Maruatona, Vamplew y Dazeley, 2012). En España, Dorronsoro *et al.* (1997) crearon un sistema de detección, también basado en *NN*, ampliamente utilizado.

Los métodos supervisados o de discriminación se han aplicado con éxito a las operaciones con tarjetas de crédito, pero son menos útiles para los fraudes en cuenta corriente donde no siempre disponemos de los dos conjuntos de datos buenos y fraudulentos. Además, el tipo de operaciones fraudulentas puede ser muy amplio, y cambiante en el tiempo. Por ejemplo, si el posible fraude es realizar una transferencia a una cuenta controlada por el delincuente, no dispondremos en general de una base de transferencias fraudulentas hechas en condiciones similares para clasificar una nueva solicitud. Este tipo de fraude se detecta mejor con métodos de aprendizaje no supervisado, que incluye los métodos de identificación de datos anómalos y de detección de grupos, y que se utilizan cuando tenemos que basarnos únicamente en una secuencia disponible de transacciones, presumiblemente no fraudulentas. Los métodos que se utilizan se basan en comparar la transacción con las históricas y determinar la probabilidad de que la transacción sea significativamente diferente de las anteriores.

La disponibilidad de mucha más información respecto al uso de la tarjeta y los hábitos y características de su propietario permiten detectar mejor las operaciones fraudulentas pasando de modelos generales de detección del fraude a modelos específicos para cada tipo de cliente, en función de su situación. Además, el estudio de los datos pasados permite prever el tipo y características del defraudador que suele estar asociado a un tipo de cliente. También, se hace una graduación más fina de la probabilidad de que la operación sea fraudulenta, de manera que cuando esta probabilidad sea baja se acepte, cuando sea dudosa se consulte con el cliente y cuando sea alta se rechace y se notifique la situación de riesgo al cliente. El *Big Data* permite

especializar mucho mejor la población de referencia para cada cliente, incorporando no solo su propia actividad, sino la de personas con características similares a las suyas, generando una información mucho más precisa para la detección instantánea del fraude.

IV. UN EJEMPLO DE ANÁLISIS DE REDES DE CLIENTES

Esta sección presenta algunos resultados derivados del estudio de diferentes redes de clientes del Banco Santander (BS en adelante) desarrollado por varios miembros del Instituto Mixto UC3M-BS en *Big Data* de la Universidad Carlos III de Madrid (IBiDat en adelante) con el objetivo de ilustrar en un caso concreto algunas de las oportunidades del uso de información del *Big Data* para instituciones financieras descritas en la tercera sección.

1. Objetivos del proyecto

El objetivo del proyecto consistía en determinar si la información contenida en diferentes redes de clientes del BS y en la estructura de sus relaciones es de utilidad para mejorar las políticas comerciales del banco. Los resultados del proyecto demostraron que la respuesta es claramente afirmativa: en primer lugar, disponer de la red de clientes permite clasificarles mejor por su importancia; segundo, utilizar la red mejora la captación de nuevos clientes; en tercero mejora la predicción de la morosidad.

Durante el proyecto se ha elaborado un método para determinar el camino óptimo que debe seguir un gestor para contactar con un potencial nuevo cliente que tenga relaciones con clientes del BS y que será brevemente detallado posteriormente. Además, se ha comprobado que la información contenida en las redes de clientes es muy útil para explicar la evolución financiera de clientes en mora.

2. Datos disponibles

Para poder construir y analizar las diferentes redes de clientes del BS, fue necesario trabajar con grandes bases de datos proporcionados por el BS de sus clientes en España y de las relaciones de dichos clientes con otros (clientes y no clientes). Esta información ha estado disponible en tres momentos temporales separados en seis meses. Más concretamente, el BS nos permitió trabajar con cinco bases de datos de ámbito nacional correspondientes a: 1) relaciones entre clientes y no clientes; 2) perfiles de clientes y no clientes incluidos o no en la base de datos anterior; 3) importes de mora de clientes españoles; 4) datos de gestores-clientes; 5) datos de no clientes objetivos y de sus perfiles. Estas bases de datos fueron depuradas y estructuradas para construir diferentes redes de clientes necesarias para los objetivos propuestos. Es importante resaltar que para poder utilizar toda la información relevante fue necesario preservar el anonimato de los clientes, prestar una atención extrema a la confidencialidad de la información, y tener el máximo respeto a la Ley de Protección de Datos.

Vamos a resumir brevemente algunas de las variables que aparecen en estas bases de datos. En primer lugar, nos vamos a centrar en la base de datos de perfiles de clientes y no clientes. Esta base de datos se estructura en un conjunto de filas y columnas, donde cada fila corresponde a un cliente y cada columna son valores de variables asociadas a cada cliente. Algunas de estas variables son:

- El tipo de persona, es decir, si es física, jurídica, compañía, etc.
- El código de cliente.
- La presencia de mora o no en cada uno de los tres períodos analizados.
- El importe de la mora en cada uno de los tres períodos analizados.
- La edad del cliente.
- La ocupación del cliente.
- El segmento al que pertenece el cliente.
- El grado de vinculación del cliente con el BS en cada uno de los tres períodos analizados.
- Los recursos que tiene el cliente en el BS.
- El dinero del cliente en cuentas corrientes del BS.
- La deuda del cliente en hipotecas, préstamos al consumo, etc.
- Muchas variables dicotómicas, del tipo: si tienen nómina domiciliada en el BS, si disponen de tarjetas del BS, si tiene recibos domiciliados, si utiliza banca por Internet/móvil, etc.

La base de datos de relaciones entre clientes y/o no clientes del BS se estructura en un conjunto de filas y columnas, donde cada fila proporciona información sobre la relación entre un cliente y un cliente/no cliente del BS y cada columna proporciona variables como:

- Los códigos identificativos de los sujetos que forman la relación.
- El sentido en el que se debe leer un registro.
- El tipo de relación entre los sujetos.
- La intensidad de la relación.
- Las fechas de primera y última relación en el período considerado.

Las bases de datos sufrieron un fuerte preprocesamiento antes de poder construir las diferentes redes utilizadas en el proyecto. Por ejemplo, se realizó una tarea de limpieza para que en la red solo apareciesen los clientes con relaciones y/o que estén en carteras de gestores. Adicionalmente se substituyeron datos faltantes correspondientes a exclientes que aparecen en las relaciones con datos admisibles. También se eliminaron operaciones no deseadas entre clientes, clientes irrelevantes, relaciones repetidas entre clientes y relaciones entre clientes no relevantes para los objetivos del estudio.

El preprocesamiento ha reducido el tamaño de los datos finalmente utilizados, aunque sigue siendo enorme. Por ejemplo, la red de clientes construida para la captación de nuevos clientes contiene 6.329.506 relaciones entre 4.783.145 clientes relevantes, una reducción importante de la base de datos inicial que contenía 81 millones de relaciones entre 33 millones de clientes y no clientes.

3. Análisis descriptivo de la red de clientes

Como hemos comentado en el apartado cuatro de la tercera sesión, una red es un objeto llamado grafo formado por vértices y aristas. En nuestro caso particular, los vértices del grafo (red) serán los clientes (personas, empresas, organizaciones y demás clientes) del BS, y las aristas representan relaciones o flujos entre dichos clientes. Para realizar un análisis descriptivo de la red se utilizan varias características, incluyendo medidas que midan la centralidad (importancia) de los clientes con el objetivo de cuantificar las relaciones de poder, protagonismo, confianza, etc., y la detección de comunidades específicas que puedan tener características interesantes a señalar y estudiar.

El primer paso del análisis es crear el grafo que represente la red de interés. Para ello, y tal como se ha mencionado previamente, se utilizaron las bases de datos relativas a las operaciones entre clientes y la descripción de clientes descritas en el punto 2 anterior. En este grafo, cada vértice representará un cliente del BS y cada arista representará al menos una relación entre dos clientes. Además, cada arista está valorada por un peso que toma valores en el intervalo [0,1], para representar la cercanía entre los clientes que une, de nuevo en función de la red creada. Por tanto, un peso próximo a 1 representa una gran cercanía entre los dos clientes. Un ejemplo de definición de peso se describe en el apartado 4, donde se describe la determinación de caminos óptimos.

Una vez construida la red, el siguiente paso es la obtención de sus componentes conexas. Una componente conexa de una red es una subred, es decir, una parte de la red, en la que todos sus vértices están conectados a través de trayectorias entre las relaciones que forman la subred, y a la que no se pueden añadir más vértices que cumplan dicha propiedad. Si la mayor componente conexa del grafo incluye una fracción muy significativa del mismo, se la denomina componente gigante. El resto de componentes se denominan componentes aisladas. En el caso particular de la red construida para la captación de nuevos clientes, existe una componente gigante de 4.783.145 clientes y 6.329.506 relaciones, frente a los 604.020 clientes y 302.010 relaciones que pertenecen a relaciones aisladas entre ellos, es decir, relaciones entre dos clientes totalmente aislados del resto de la red. Por tanto, el resto del análisis se puede concentrar en dicha componente gigante mientras que los clientes aislados pueden ser identificados por el BS para diseñar estrategias para fidelizar a estos clientes periféricos.

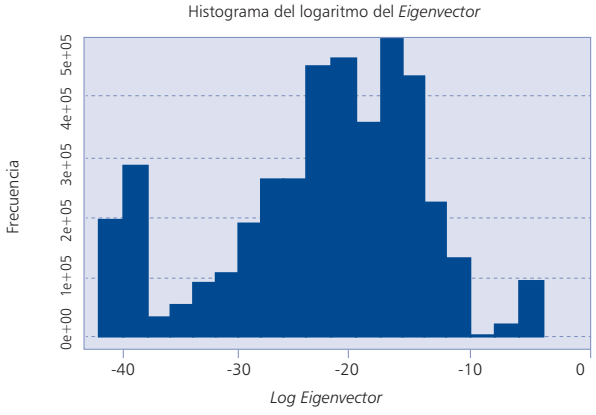
El siguiente paso es obtener medidas de centralidad de los clientes ya que nos sirven para conocer que clientes son los más importantes dentro de la componente gigante de la red. Como se mencionó en la sección tercera, hay diferentes opciones para definir centralidad, siendo la más popular el grado definido como el número de relaciones coincidentes con dicho cliente ponderadas por el peso de cada una de las relaciones. Es razonable pensar que los vértices con más enlaces son vértices centrales de la

red. Sin embargo, hay que tener en cuenta que el grado solo mide la importancia con respecto a los clientes más cercanos. Es decir, se asume que las conexiones de los clientes no importan, solamente importa las relaciones directas con los vecinos. Una segunda opción es la centralidad por vector propio o *eigenvector* que proporciona un mayor valor a aquellos clientes que están conectados a muchos clientes que a su vez están bien conectados en este sentido, y por tanto son buenos candidatos a difundir información útil para el BS. De esta manera se tiene en cuenta tanto la cantidad como la calidad de las mismas relaciones.

El histograma mostrado en el gráfico 2 muestra el logaritmo del valor de la centralidad por vector propio de la red de clientes del BS que sugiere tres grupos de clientes distintos de más a menos importantes: 1) un grupo con alto valor de la log-centralidad, entre -10 y 0; 2) un grupo con valor medio de log-centralidad, entre -10 y -38 aproximadamente; y 3) un grupo con valores bajos de log-centralidad, por debajo de -38. Evidentemente, el grupo formado por los clientes más importantes es el menos numeroso, mientras el grupo que engloba a una mayor cantidad de clientes es el de importancia intermedia. De esta manera, el BS puede identificar qué clientes son los más relevantes y estudiar posibles acciones que permitan su retención.

El siguiente paso del estudio descriptivo de la red es la detección de comunidades. Las redes complejas, como la estudiada, tienden a mostrar una alta concentración de enlaces en ciertas regiones del grafo (comunidades o clústers), y una baja concentración de enlaces fuera de esas regiones. Esta propiedad suele darse como consecuencia de la heterogeneidad global y local de la distribución de los enlaces en un grafo. Por tanto, las comunidades o clústers se definen como grupos de vértices densamente conectados que presentan conexiones dispersas entre sí. Como se definió en la tercera sección, la modularidad es una medida con valor en el intervalo [-1,1] que mide la presencia de los grupos. Cuanto mayor sea el valor de la modularidad de una red, más significativa será la presencia de los grupos. En este caso, el valor de la modularidad es 0.8452562, por lo que la red está claramente formada por comunidades. Utilizando el algoritmo de partición de comunidades de Louvin mencionado, en la sección tercera, se han detectado 119.389 comunidades. El histograma mostrado en el gráfico 3 muestra la distribución de los tamaños de dichas comunidades. La gran mayoría de las comunidades tiene un tamaño muy pequeño. Un estudio de cada comunidad permite investigar sus características e identificar los factores que relacionan los clientes que las componen. Esta información se podría aprovechar, por ejemplo, para el diseño de campañas y productos financieros del BS para cada comunidad.

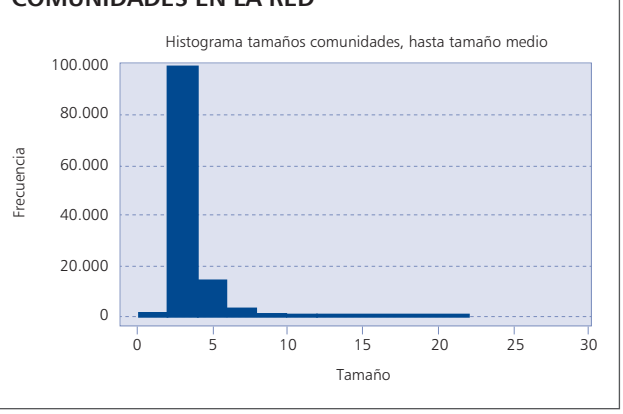
GRÁFICO 2
HISTOGRAMA DEL LOGARITMO DE LAS CENTRALIDADES DEL VECTOR PROPIO O *EIGENVECTOR*



4. Captación de clientes

A continuación mostramos una metodología desarrollada para apoyar a los gestores de clientes del BS en su tarea de captar nuevos clientes y recursos. Esto incluye la atracción al BS de personas que no son clientes, pero también, la intensificación de la colaboración entre los clientes y el BS. A partir de

GRÁFICO 3
DISTRIBUCIÓN DEL TAMAÑO DE LAS COMUNIDADES EN LA RED



una red de clientes y no clientes, los perfiles de los clientes y las carteras de los gestores, el método indica una colección de posibles caminos óptimos para llegar al cliente objetivo a través de sus contactos con clientes del BS. Dichos caminos óptimos se originan en el gestor y utilizan su cartera de clientes.

El objetivo es determinar la secuencia de personas clientes del BS que un gestor tiene en su cartera y que debería contactar para poder llegar a un potencial cliente que se define como objetivo. Supongamos, por ejemplo, que un gestor quiere conectar con un cliente objetivo a través de tres clientes en su cartera que, a su vez, se relacionan con otros clientes que pueden conducirle al cliente objetivo. Cada relación entre clientes lleva asociada un peso que represente la probabilidad de éxito del gestor. Por lo tanto, suponiendo independencia entre las diferentes relaciones, la probabilidad de éxito de cada uno de los caminos posibles se puede obtener como el producto de las probabilidades de éxito de cada relación utilizada.

Dicho lo anterior, el problema a resolver es: dados unos clientes iniciales (los que forman la cartera de clientes del gestor) y un objetivo, y contando con los miembros de la red de clientes del BS que están conectado con los clientes iniciales (la cartera de clientes del gestor), se desea determinar el camino hacia el objetivo con la mayor probabilidad de éxito. Para encontrar la solución deseada son necesarias tres etapas: 1) construir la red propia del cliente objetivo, que es aquella que contiene todas las conexiones de los clientes de la cartera del gestor con el cliente objetivo; 2) estimar la probabilidad de éxito de cada posible camino; y 3) seleccionar entre los caminos posibles el de máxima probabilidad de éxito. Dado que la red construida con todos los clientes que permiten alcanzar el cliente objetivo puede ser muy extensa e incluir caminos muy largos, parece adecuado fijar un número máximo de clientes por los que se debe pasar para llegar al objetivo.

La asignación de las probabilidades de éxito de cada una de las relaciones es la clave del método. Supongamos que queremos determinar la probabilidad de éxito entre un cliente *i* y un cliente *j*, p_{ij}, que puede representar, por ejemplo, que el cliente *i* influya positivamente al cliente *j* acerca de un producto del BS, o que el no cliente *j* se convierta en cliente por la interacción con el cliente *i*. Para determinar esta probabilidad p_{ij} se han considerado cuatro factores que parecen determinar la influencia de una persona sobre otra (ver Masthoff y Gatt, 2006). Estas son:

- La confianza entre las dos personas (también referido como la fuerza del vínculo). Definiremos la variable tie_{ij}, (*tie strength*, o confianza), entre *i* y *j*.

- La satisfacción del posible influenciador (*i* en nuestro caso) con el objeto a recomendar (el banco en sí o un producto de este). Lo caracterizaremos con la variable sat_i, que mide la satisfacción aparente del cliente *i* con el BS.

- La influencia que tiene *i* sobre *j* (o sobre la red de usuarios en general, su autoridad). Definiremos la variable aut_i, influencia de *i*, o capacidad de persuasión.

- La facilidad con la que el (potencial) cliente (*j*) puede ser persuadido, es decir, su tipo de personalidad o grado de conformidad. Se caracteriza mediante la variable per_j, *personality* de *j*, o grado de conformidad.

Para calcular el valor de la variable tie_{ij} utilizamos el artículo de Granovetter (1973) quien define la *fuerza del vínculo* como una combinación lineal de cuatro factores: cantidad de tiempo, intensidad emocional, intimidad y servicios recíprocos que caracterizan el vínculo. Siguiendo estas definiciones, podemos definir el *tie strength* como la combinación de: 1) $time_{ij}$, el tiempo que *i* y *j* han estado interactuando; 2) $intensity_{ij}$, la intensidad de las transacciones entre *i* y *j*; 3) $intimacy_{ij}$, el estado de estar en una relación privada o íntima; y 4) $reciprocal_{ij}$, situación en la que *i* y *j* se prestan servicio mutuo. En resumen y sin entrar en los detalles concretos para preservar información confidencial: 1) la variable $time_{ij}$, se calcula como el tiempo (años, meses) durante el cual *i* y *j* han estado interactuando mediante transacciones a través del BS; 2) la variable $intimacy_{ij}$, se calcula asignando un peso a la intensidad de las relaciones entre dos vértices *i* y *j*; 3) la variable $intensity_{ij}$ se calcula como el peso máximo entre las operaciones entre *i* y *j* dividido por el rango de los pesos; y 4) La variable $reciprocal_{ij}$ toma el valor 1, si hay relaciones recíprocas, es decir existen relaciones de *i* a *j* y de *j* a *i*, y el valor $reciprocal_{ij} = 0{,}5$, si las relaciones son unidireccionales. A continuación, se construye una combinación lineal de estas cuatro variables para ciertos pesos, lo que nos permite determinar el valor de la variable tie_{ij}.

Para calcular el valor de la variable sat_i utilizamos el compromiso del cliente con el BS, para lo que damos diferentes pesos según el nivel de compromiso con el banco.

Para calcular el valor de la variable aut_i utilizamos el valor de diferentes medidas de centralidad del cliente. Por ejemplo, podemos definir:

$$inf_i = (grado_i - minGrado_G) / (maxGrado_G - minGrado_G), \quad [1]$$

donde G es la red total tomada como no dirigida y sin repetición de aristas, $minGrado_G$ y $maxGrado_G$ son el grado mínimo y máximo en la red, respectivamente, mientras $grado_i$ es el grado del nodo *i*.

Para calcular el valor de la variable per_j hay que tener en cuenta que no existe información de la misma para los no clientes. Sin embargo, para un cliente podemos interpretar este atributo como el éxito que ha tenido hasta ahora el BS en vender productos a este cliente, por lo que podemos interpretar per_j como la satisfacción y definirla como vimos anteriormente.

Una vez definidas todas las variables, se supone por sencillez que la probabilidad p_{ij} *es una combinación lineal de estas cuatro variables tal que:*

$$p_{ij} = peso\_tie * tie_{ij} + peso\_sat * sat_i + peso\_per * per_j + peso\_aut * aut_i \quad [2]$$

para unos ciertos pesos, *peso_tie, peso_sat, peso_per* y *peso_aut*. Como se mencionó previamente, una vez calculados todos los pesos necesarios, la probabilidad de éxito de un camino *P* se calcula como el producto de las probabilidades p_{ij} de todas las aristas *(i, j)* pertenecientes a *P*. Por tanto, el camino óptimo será aquel que maximiza el valor de dicho producto. Para ello, se requiere la resolución de un problema de optimización tal como se describe en Quijano-Sánchez y Liberatore (2017).

5. Mejora de la predicción de mora con variables de red

El tercer y último punto que vamos a considerar es el de tratar de explicar la mora de los clientes del BS con la ayuda de ciertas variables construidas con determinadas características de la red. Para ello, se han construido modelos para explicar la mora en dos momentos temporales separados por seis meses utilizando un conjunto amplio de variables que incluyen el uso de los servicios del BS, los recursos del cliente, sus relaciones en la red de clientes y los cambios en estas variables durante el período considerado.

Presentamos a continuación la metodología utilizada y un breve resumen de las variables explicativas seleccionadas. Hemos utilizado regresión logística para explicar la variable mora por dos razones principales: 1) este modelo permite determinar la importancia de cada una de las variables utilizadas para explicar la variable respuesta lo que permite explicar mejor la morosidad de los clientes del BS y medir sus efectos en términos de la probabilidad de mora; y 2) este modelo ha demostrado su utilidad en problemas similares.

La variable mora describe si un cliente es moroso (valor igual a 1) o no moroso (valor igual a 0), y está disponible para tres momentos temporales que vamos a denominar, M1, M2 y M3, respectivamente, separados seis meses cada uno de ellos. Nos centramos en explicar: 1) la variable mora en M2 con información disponible en el período M1-M2; (2) la variable mora en M3 con la información disponible en el período M2-M3; y (3) la variable mora en M3 con la información disponible únicamente en M2. Los experimentos 1 y 2 requieren modelos explicativos y de predicción a corto plazo, ya que calculan la probabilidad de mora de un cliente hoy utilizando la situación financiera actual y pasada del cliente. El experimento 3 requiere un modelo de predicción con horizonte de seis meses.

Los modelos se han construido para distintas clases de clientes que resultan de segmentarlos por el tipo de persona («Autónomos», «Empresas» y «Físicas») y cuatro categorías diferentes de su vinculación con el BS, muy fuerte, fuerte, media y baja. Por tanto, se obtienen doce clases de clientes (desde «Autónomos con vinculación muy fuerte», AMF, hasta «Físicas con vinculación baja», FB). La frecuencia de aparición de la variable mora es distinta en los diferentes grupos en los tres momentos temporales, siendo más alta entre los Autónomos con baja vinculación y la más baja en Físicas con vinculación muy fuerte. En general, la morosidad disminuye con el grado de vinculación.

Las variables explicativas disponibles pueden clasificarse en tres grupos o bloques: 1) variables categóricas que miden el uso de productos y servicios por el cliente tanto en el momento de predicción

de la mora como el cambio experimentado entre el momento que se explica y el momento observado anterior; 2) variables cuantitativas que miden recursos disponibles: cantidades en cuentas corrientes, fondos, etc., así como el cambio de recursos experimentado respecto al momento observado anterior; y 3) variables de red que miden la relación de ese cliente con otros clientes morosos, así como las correspondientes variables de cambio. En este tercer bloque de variables, se consideran 12 variables: 1) proporción de vecinos morosos a 1 Paso (una variable para Autónomos, otra para Empresas y otra para Físicas); 2) cambio en la proporción de Autónomos morosos a 1 Paso (tres variables como antes); 3) como en 1, pero vecinos a dos pasos; y (4) como en 2, pero vecinos a dos pasos.

Una vez definidas las variables explicativas, se introducen en el modelo de regresión logística y se lleva a cabo un procedimiento estadístico automático e iterativo de eliminación «hacia atrás» (*backward*) de variables. Este procedimiento automático se ha combinado con algunas reglas diseñadas *ad hoc* para estos modelos, como por ejemplo la eliminación de variables con valores diferentes de 0 en menos del 1 por 100 de las observaciones de un modelo.

Los resultados obtenidos en las dos primeras regresiones logísticas, es decir a los modelos explicativos, se pueden resumir en tres puntos fundamentales:

1. La variable «Mora en el período anterior» entra en todos los modelos de tal manera que indica que los clientes en mora hace seis meses tienen una probabilidad muy alta de continuar en mora seis meses después.

2. Las relaciones en la red con morosos hacen aumentar mucho la probabilidad de mora. Existe una relación positiva entre la variable mora en el período a explicar y las variables «Proporción de Autónomos Morosos a 2 Pasos» y «Proporción de Empresas Morosas a 2 Pasos». Un resultado similar se observa para las variables de cambio asociadas a estas variables («Cambio en la Proporción de Autónomos Morosos a 2 Pasos» y «Cambio en la Proporción de Empresas Morosas a 2 Pasos»).

3. Hay un pequeño grupo de variables que en muchos de los modelos, si bien su importancia relativa con respecto a las variables mencionadas en los dos puntos anteriores es muy inferior.

Estos resultados sugieren que a corto plazo es posible prever la mora con muy pequeño error. De hecho, con estas variables explicativas un modelo logístico es capaz de prever la mora a corto plazo con un error del orden de uno por mil para los falsos positivos (se prevé que el cliente entra en mora cuando no lo hace) y del uno por cien para los falsos negativos (se prevé que el cliente no entra en mora cuando sí lo hace). Además, los resultados obtenidos demuestran la importancia que tienen las variables de red para la explicación de la mora. Si eliminamos dichas variables, los modelos explicarían la mora mucho peor.

Sin embargo, los resultados obtenidos en la tercera regresión logística, es decir para el modelo predictivo a seis meses, son notablemente inferiores. Los errores de predicción obtenidos para la mora en el momento M3 son mucho más altos como consecuencia de que prevemos con información sobre la situación de los clientes hace seis meses, y no uno o dos meses antes. Es decir, que las variables importantes en los casos explicativos no son capaces de prever con precisión la aparición de mora en un horizonte de seis meses. Esto sugiere que un modelo con información dinámica de la evolución de las variables detectadas como clave en los pocos meses previos a la entrada en mora será capaz de prever la entrada de un cliente en mora en uno o dos meses con un pequeño error.

V. CONCLUSIONES

Los bancos, cajas de ahorros y otras instituciones financieras pueden mejorar su situación diseñando mecanismos para recoger y analizar la enorme cantidad de información que genera la actividad económica de sus clientes y utilizarla para adaptar más sus políticas comerciales a las necesidades de sus clientes. Por ejemplo, el análisis detallado de distintas series de ingresos y gastos en la cuenta de cada cliente puede anticipar cambios que indican un aumento de su probabilidad de abandono. Por otro lado, esta información permitiría anticiparse a las necesidades de los clientes y proponerles estrategias para mejorar su comportamiento financiero. Además, esta información puede enriquecerse con otros datos disponibles sobre los clientes en las redes sociales y otros sistemas públicos de captación de datos, mejorando la segmentación de los clientes y permitiendo mejores predicciones sobre su actividad. Un uso inteligente de toda esta información

puede aumentar la lealtad de los clientes con su institución, mejorar la imagen pública del banco y situarlo en una posición más sólida ante los retos futuros derivados de la entrada en el sector financiero de empresas líderes en la recogida y manejo de información.

La información detallada sobre sus clientes puede además reducir sus costes, al disminuir el fraude y la morosidad, y conducir a mejores predicciones que disminuirán los costes asociados al riesgo.

Por otro lado, los grandes bancos disponen al agregar la evolución de sus clientes, preservando la confidencialidad de los datos individuales, de una información muy valiosa para prever el ciclo económico y anticipar problemas en sectores concretos. La experiencia en áreas indica que esta información, convenientemente utilizada, puede generar nuevo valor económico.

BIBLIOGRAFÍA

Abdallah, A., Maarof, M. y Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, *68,* pp. 90-113.

Bartov, E., Faurel, L. y Mohanram, P. S. (2018). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review*, *93,* pp. 25-57.

Benoit, D. F. y Van Den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert System with Applications*, *39,* pp.11435-11442.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blondel, V. D., Guillaume, J-L., Lambiotte, R. y Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *(10),* P10008.

Bolton, R. y Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17,* pp. 235-249.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. y Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge and Data Engineering*, *20,* pp. 172-188.

Burez, J. y Van den Poel, D. (2008). Separating financial from commercial customer churn. *Expert System with Applications*, *35,* pp. 497-514.

Chen, H., De P., Hu Y. J., y Hwang, B-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, *27,* pp. 1367-1403.

Clauset, A., Newman, M. E. J. y Moore, C. (2004). *Finding community structure in very large networks*. http://www.arxiv.org/abs/cond-mat/0408187

De Bock, K. y Van den Poel, D. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Application*, *38,* pp. 12293-12301.

Dorronsoro, J. R., Ginel, F., Sánchez, C. y Cruz, C. S. (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks*, *8,* pp. 827-834.

Einav, L. y Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14,* pp. 1-24.

European Central Bank (2018). *Fith Report on card fraud*. https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport201809.en.html

Fang, B. y Zhang, P. (2016). Big data in finance. En *Big Data Concepts, Theories, and Applications*. Springer, pp. 391-412.

Galeano, P. y Peña, D. (2019). Statistics, Big Data and Data Science (with discussion). *TEST*, *28,* pp. 289-368.

Giannone, D., Reichlin, L. y Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55,* pp. 665-676.

Granovetter, M. S. (1973). The strength of weak ties. *The American Journal of Sociology*, *78,* pp. 1360-1380.

Hastie, T., Tibshirani, R. y Friedman, J. H. (2009). *The Elements of Statistical Learning Data* (2nd edition). Springer.

Kahneman, D. (2012). *Pensar rápido, pensar despacio*. Debate.

Kapetanios, G. y Papaillias, F. (2018). Big Data & Macroeconomic Nowcasting: Methodological review. *ESCoE Discussion paper*, 2018-12.

Kou, Y., Lu, C-T., Sirwongwattana, S. y Huang, Y-P. (2004). Survey of fraud detection techniques. En *IEEE International Conference on Networking, Sensing and Control, 2004*. IEEE, pp. 749-754.

Maruatona, O., Vamplew, P. y Dazeley, R. (2012). Prudent fraud detection in Internet banking. *2012 Third Cybercrime and Trustworthy Computing Workshop*. IEEE, pp. 60-65.

Masthoff, J. y Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. *User Modelling and User-Adapted Interaction*, *16,* 281-319.

Mayer-Schönberger, V. y Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Paradis, L. y Han, Q. (2007). A survey of fault management in wireless sensor networks. *Journal of Network and Systems Management*, *15,* pp. 171-190.

Pedregosa, F. *et al*. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12,* pp. 2825-2830.

Peña, D. (2014). Big Data y Estadística: ¿Tendencia o cambio? *Boletín de Estadística e Investigación Operativa*, *30,* pp. 313-324.

— (2015) Big Data, Ciencia y Estadística. *Revista de Ciencia y Humanidades*, *14,* pp. 97-106. Fundación Ramón Areces.

Phua, C., Lee, V., Smith, K. y Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Pons, P. y Latapy, M. (2005). *Computing communities in large networks using random walks*. http://arxiv.org/abs/physics/0512106

Quijano-Sánchez, L. y Liberatore, F. (2017). The BIG CHASE: A decision support system for client acquisition applied to financial networks. *Decision Support Systems*, *98,* pp. 49-58.

Seth, T. y Chaudhary, V. (2015). Big Data in Finance. En Li, K.-C., Jiaang, H., Yang, L. T., Cuzzocrea, A. (eds.), *Big Data: Algorithms, Analytics, and Applications*, 1 ed., Chapter 17, CRC Big Data Series, p. 29. Chapman&Hall.

Ugarte, M. D., Militino, A. F. y Arnholt, A. T. (2015). *Probability and Statistics with R.* Chapman and Hall/CRC.