ManagementSolutions
Making things happen

Data science
y la transformación del sector financiero

Entidades Financieras

www.managementsolutions.com

Diseño y Maquetación
Dpto. Marketing y Comunicación
Management Solutions - España

Fotografías
Archivo fotográfico de Management Solutions
iStock

Índice

Introducción

Sin datos, no es usted más que otra persona con una opinión.
W. Edwards Deming[1]

El mundo se está transformando, y lo hace a gran velocidad. Estamos siendo testigos de una revolución tecnológica de magnitudes nunca antes observadas.

No se trata de un hecho coyuntural. El índice de cambio de paradigma (la velocidad de adopción de nuevas ideas) se está duplicando cada década: mientras que se tardó casi medio siglo en adoptar el teléfono, y aceptar la televisión y la radio llevó varias décadas, el ordenador, Internet y los teléfonos móviles se asumieron en menos de 10 años[2]. En 2014, el número de teléfonos móviles ya se equiparaba al número de personas en el mundo, 7.000 millones, un tercio de ellos smartphones; y el número de usuarios de Internet casi alcanzó los 3.000 millones[3].

Las tecnologías de la información duplican cada año su capacidad y su relación calidad/precio, como predice la Ley de Moore[4], que ha demostrado cumplirse hasta la fecha (Fig. 1). La consecuencia es un crecimiento exponencial en la disponibilidad de la tecnología y una reducción equivalente en su coste, indiferente a las crisis vividas en los últimos años, que previsiblemente continuará su evolución en las próximas décadas.

Pero esta revolución tecnológica está adquiriendo una nueva dimensión en los últimos años: al aumentar las prestaciones técnicas, también está aumentando la capacidad de generar, almacenar y procesar información, y lo hace a una tasa también exponencial, lo que ha dado en llamarse el fenómeno «big data». Algunas evidencias al respecto son:

- El volumen total de datos en el mundo se duplica cada 18 meses[5,6].
- Más del 90% de los datos que hoy existen han sido creados en los dos últimos años[6].
- La capacidad per cápita de almacenar información se ha duplicado cada 40 meses desde 1980 y su coste se ha reducido en más de un 90%[6].
- La capacidad de procesamiento se ha multiplicado por 300 desde el año 2000, permitiendo procesar millones de transacciones por minuto[6].

El impacto de esta transformación tecnológica está siendo especialmente relevante en el sector financiero, por cuanto viene a sumarse a otras cuatro grandes tendencias que están marcando su evolución:

1. Una coyuntura macroeconómica caracterizada por un crecimiento débil, bajas tasas de inflación y reducidos tipos de interés, que ha penalizado los márgenes de beneficio de la industria bancaria en las economías maduras durante un prolongado periodo de tiempo; y un comportamiento dispar en los países emergentes, con una tendencia a la ralentización del crecimiento y el repunte de la morosidad.

2. Un entorno normativo más exigente e intrusivo, donde la regulación adquiere un carácter global en términos de gobierno corporativo, solvencia, liquidez, limitación del bail-out, protección del consumidor, prevención del fraude y requisitos de información y reporte, entre otros.

3. Un cambio profundo en el comportamiento del cliente, que ahora tiene una mayor cultura financiera, espera y exige excelencia en el servicio, al tiempo que manifiesta una creciente confusión ante la complejidad y disparidad de la oferta, lo que le hace más dependiente de los líderes de opinión.

4. La entrada de nuevos competidores en el mercado financiero, algunos de ellos con nuevos modelos de negocio que impactan en el statu quo.

[1] William Edwards Deming (1900-1993). Estadístico estadounidense, profesor universitario, autor, consultor y difusor del concepto de calidad total, notable por su trabajo en el desarrollo y crecimiento de Japón tras la Segunda Guerra Mundial.

[2] Kurzweil [Director de Ingeniería de Google] (2014).

[3] International Telecommunication Union (2014).

[4] Observación de Gordon Moore, cofundador de Intel, en 1965: la tecnología evoluciona de modo que el número de transistores en un circuito integrado se duplica cada dos años aproximadamente. Moore (1965).

[5] Se estima que cada día de 2012 se produjeron 2,5 exabytes de datos, un volumen de información que equivale a 12 veces todos los libros impresos que hay en el mundo.

[6] Federal Big Data Commission (2014).

El efecto combinado de estos cuatro factores, junto con la transformación tecnológica, está conduciendo, entre otras cuestiones, a poner el foco en el uso eficiente de la información, dando así entrada en el sector financiero a una disciplina hasta ahora más centrada en el sector tecnológico: data science.

Data science, o la ciencia de los datos, es el estudio de la extracción generalizable de conocimiento a partir de los datos mediante el uso combinado de técnicas de aprendizaje automático, inteligencia artificial, matemáticas, estadística, bases de datos y optimización, junto con una comprensión profunda del contexto de negocio[7].

Todas estas cuestiones ya se empleaban en el ámbito financiero en diferente medida, pero esta disciplina tiene características que la hacen indispensable para afrontar la transformación del sector que ya está ocurriendo.

En concreto, todos los elementos del complejo contexto antes mencionado al que se enfrenta el sector financiero exigen datos abundantes y técnicas analíticas complejas para afrontarlos, lo que es exactamente el campo de especialidad de data science. Además, data science es una disciplina que se ve potenciada como consecuencia del fenómeno big data, y por tanto los data scientists son profesionales cualificados para tratar cantidades masivas de datos desestructurados (como por ejemplo los provenientes de redes sociales), cada vez más relevantes para las entidades.

Por otra parte, esta explosión en la generación, el acceso, el procesamiento y el almacenamiento de los datos, y en la toma de decisiones basadas en ellos, sumada a los otros factores coyunturales descritos, no ha pasado inadvertida a los reguladores. En efecto, hay una tendencia global sustanciada, entre otros, por el Comité de Supervisión Bancaria de Basilea (a través de la norma BCBS 239), hacia la exigencia de un marco robusto de gobierno de datos, que garantice su calidad, integridad, trazabilidad, consistencia y replicabilidad para la toma de decisiones, especialmente (pero no solo) en el ámbito de Riesgos.

Esta tendencia se complementa con la impulsada por la Reserva Federal y la OCC estadounidenses[8], que requiere a las entidades un marco robusto de gobierno de los modelos, para controlar y mitigar el riesgo que se deriva de su utilización, conocido como «riesgo de modelo»[9].

Las entidades financieras están avanzando de forma decisiva en el desarrollo de estos marcos de gobierno (datos y modelos), que conjuntamente conforman el gobierno de las capacidades de data science.

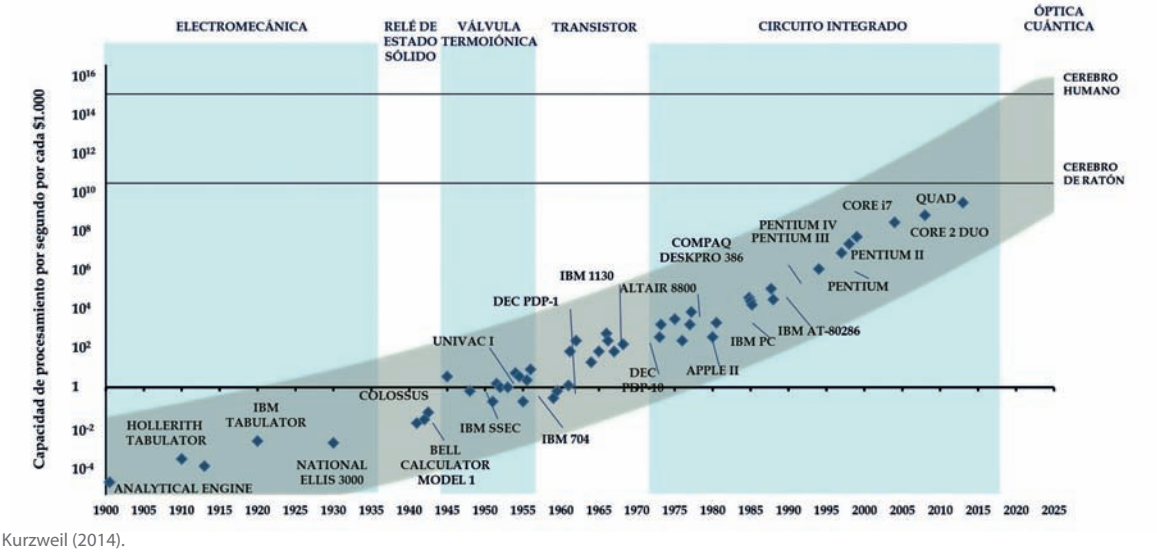
Ante este entorno cambiante, la transformación de las entidades financieras no es una posibilidad; es una necesidad para asegurar la supervivencia. Una transformación muy ligada a la inteligencia, que, en definitiva, es la capacidad de recibir, procesar y almacenar información para resolver problemas.

En este contexto, el presente estudio pretende describir de forma práctica el rol que desempeña la disciplina de data science, y más concretamente en el sector financiero. Para ello, el documento se estructura en tres secciones, que responden a tres objetivos:

- Describir la revolución tecnológica en la que está inmerso el sector financiero y sus consecuencias.
- Introducir la disciplina de data science, describir las características del data scientist y analizar las tendencias observadas a este respecto, así como su impacto en los marcos de gobierno de los datos y de los modelos en las entidades financieras.
- Exponer un caso de estudio para ilustrar la aplicación de data science en el sector financiero, consistente en el desarrollo de un modelo de scoring crediticio para particulares utilizando datos extraídos de redes sociales.

[7] Dhar [Center for Data Science, New York University] (2013).
[8] OCC/Fed (2011).
[9] Esta tendencia ha sido analizada en profundidad en Management Solutions (2014).

Figura 1. Ley de Moore: crecimiento exponencial de la capacidad de procesamiento por segundo y 1.000 dólares.

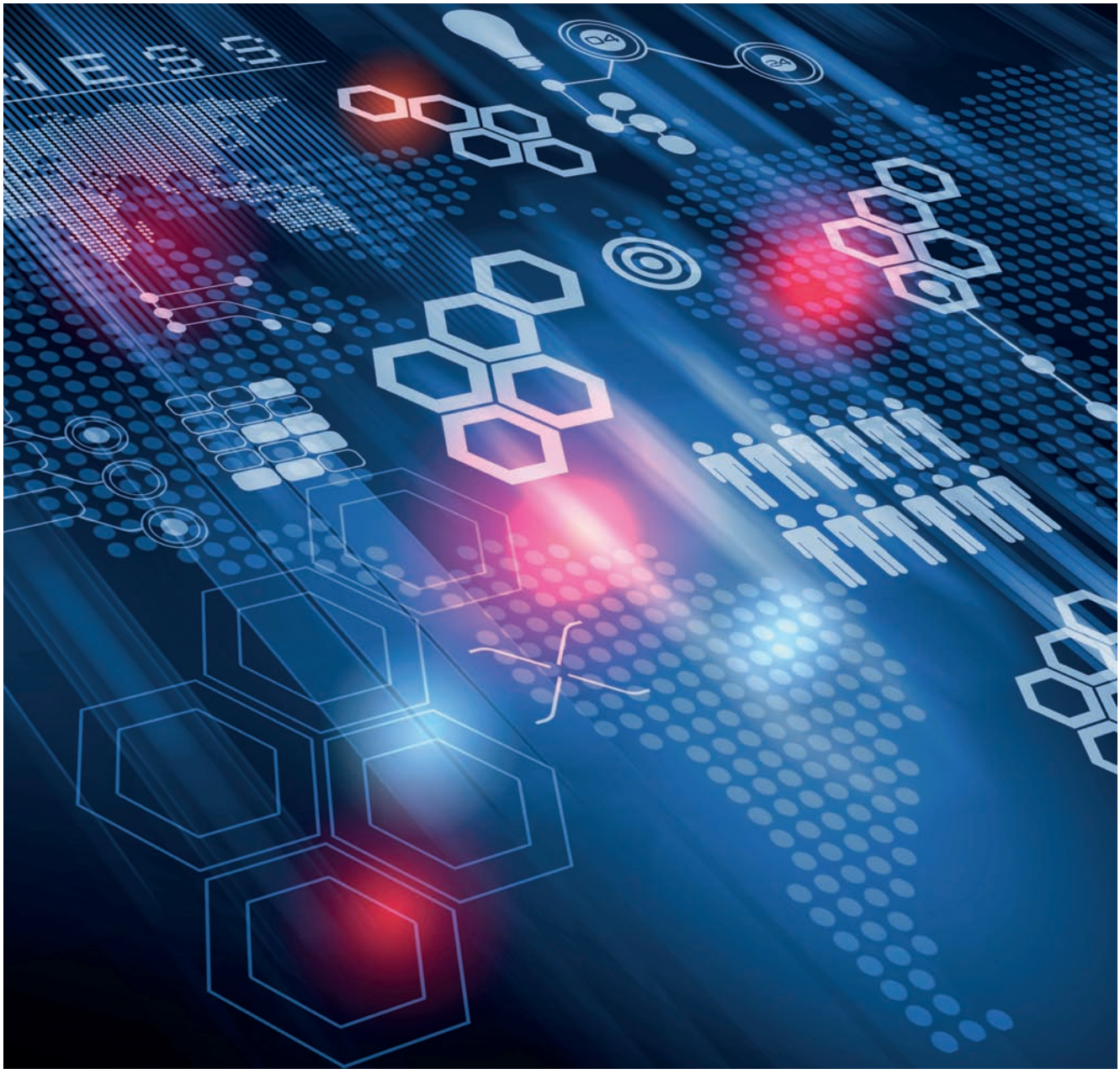


Fuente: Kurzweil (2014).

Resumen ejecutivo

Si no puede explicarlo de forma sencilla,
no lo comprende usted lo suficiente.

Albert Einstein[10]



Un sector financiero en transformación

1. Las entidades financieras se enfrentan a una revolución tecnológica sin precedentes, que está transformando el mundo en términos de lo que se puede hacer y del coste al que puede hacerse, y que, en consecuencia, impacta de forma sustancial en su actividad. Esta revolución se manifiesta tanto en la generación y el acceso a la información como en su almacenamiento, procesamiento y modelización.

   - La velocidad a la que se genera la información se está incrementando de forma vertiginosa: se estima que el volumen total de datos en el mundo se duplica cada 18 meses[11]. Desde hace años, estos datos son en su mayoría digitales, el 50% de ellos ya son accesibles mediante Internet, y el 80% son desestructurados (vídeos, imágenes, correos electrónicos, etc.)[11]. Además, gran parte de estos datos provienen de fuentes nuevas: redes sociales, logs de actividad, etc. Como consecuencia, el fenómeno big data se está caracterizando por una explosión de las «tres V»: volumen, variedad de fuentes y velocidad de generación de datos.

   - Se observa también una explosión del acceso a la información mediante dispositivos móviles: el mercado global de móviles se acerca al punto de saturación, y el de smartphones alcanzará una cuota del 51% en 2016[12]. Aunque aún hay recorrido, el crecimiento de los smartphones también se ralentiza, lo que apunta a una nueva tecnología de reemplazo en el futuro próximo, que posiblemente pase por la «Internet de las cosas»[13] y por dispositivos wearables[14], como gafas, relojes, etc., con la tecnología móvil integrada.

   - La capacidad de almacenamiento también crece de forma exponencial, y su coste unitario desciende al mismo ritmo: almacenar 1 GB de datos en 1980 costaba 10 millones de dólares, y hoy apenas llega a diez centavos de dólar[15]. Esto ha llevado a que la cantidad de información almacenada en el mundo sea masiva: en 2015 hay 8 ZB de datos, el triple que en 2012[16,17].

   - El mismo fenómeno se da en el procesamiento: la capacidad de ejecutar instrucciones por segundo por cada 1.000 dólares de procesador se ha multiplicado por casi 300 desde el año 2000[18]. Además, el desarrollo de la computación distribuida permite paralelizar las operaciones en numerosos núcleos y, avalada por gigantes tecnológicos y retailers como Google o Amazon, se perfila como el futuro del procesamiento. En el sector financiero, la banca digital y los requerimientos del mundo informacional hacen que las entidades necesiten mayores capacidades de procesamiento, y por ello ya estén adoptando máquinas de alto rendimiento y computación distribuida.

   - Por último, las capacidades de modelización están evolucionando con rapidez, impulsadas por las nuevas tecnologías y la disponibilidad de información, que abren un horizonte de posibilidades antes impensable[19]. El número de decisiones que se toman de forma automática empleando modelos en las entidades

[10] Albert Einstein (1879-1955). Físico alemán (nacionalizado después suizo y estadounidense), autor de la teoría de la relatividad, uno de los dos pilares de la Física moderna.
[11] Federal Big Data Commission (2014).
[12] International Telecommunication Union (2014).
[13] Interconexión digital de objetos cotidianos con Internet. Se estima que en 2020 habrá 26.000 millones de dispositivos conectados a la Internet de las cosas. (Gartner, 2013).
[14] Ropa o accesorios que incorporan tecnología electrónica avanzada.
[15] McCallum (2014).
[16] SiliconAngle (2014).
[17] El sector financiero también sigue esta tendencia ascendente: cada entidad maneja en media 1,9 PB en sus sistemas informacionales (DeZyre, 2014), lo que está impulsando el uso de plataformas de almacenamiento distribuido.
[18] Kurzweil [Director de Ingeniería de Google] (2014).
[19] Por ejemplo, la proliferación del «aprendizaje automático», que permite una modelización más ágil (5.000 modelos/año con 4 analistas vs. 75 modelos hoy).

financieras se multiplica cada año, lo que indudablemente produce beneficios (eficiencia, objetividad, automatización), pero también comporta riesgos.

2. Todo lo anterior está convirtiendo los datos en una nueva commodity: se generan, almacenan y procesan a un coste muy reducido, son fungibles (pierden vigencia) y, convenientemente transformados, tienen el potencial de aportar un enorme valor. Y todo ello requiere de profesionales y herramientas especializadas; en otras palabras: data science.

3. La industria financiera debería ser una de las más beneficiadas por la adopción de data science. No en vano se trata del sector que maneja la mayor cantidad y calidad de información de sus clientes (actuales y potenciales) para extraer conocimiento e incorporarlo en su propuesta de valor (entendimiento de sus necesidades, personalización de la oferta, adecuación del modelo de relación multicanal, etc.).

4. Sin embargo, en el sector financiero esta revolución tecnológica viene a sumarse a un entorno especialmente complejo, que combina elementos coyunturales con una fuerte presión regulatoria y con cambios en el comportamiento de los clientes.

5. En las economías desarrolladas, la coyuntura macroeconómica se caracteriza por un periodo prolongado de bajos tipos de interés, un crecimiento débil y bajas tasas de inflación, lo que penaliza los márgenes de beneficio de la banca. En las economías emergentes, con una elevada dependencia de la inversión pública y de políticas fiscales expansivas, se está produciendo una cierta desaceleración del crédito bancario, una ralentización del crecimiento y el repunte de la morosidad. Estos factores hacen necesario gestionar la cuenta de resultados con mayor intensidad.

6. En el ámbito normativo, se está experimentando un «tsunami regulatorio» caracterizado por la proliferación, la armonización, el endurecimiento y el carácter transnacional y más intrusivo de las normas en varios ámbitos: (1) capital y liquidez: buffers de capital, ratios de liquidez y apalancamiento, revisión de requerimientos por riesgo de crédito, mercado y operacional; (2) supervisión prudencial: refuerzo de SREP, ICAAP e ILAAP[20], stress tests supervisores; (3) limitación del respaldo público: planes de recuperación y resolución, TLAC y MREL, ring-fencing; (4) gobierno corporativo: mayores exigencias al Consejo y la Alta Dirección, nuevas figuras (CRO, CDO, CCO[21], etc.); (5) protección de los consumidores: función de Cumplimiento, control de calidad, gestión de quejas, riesgo de conducta; (6) lucha contra el fraude y los paraísos fiscales: FATCA, endurecimiento de sanciones por blanqueo; (7) ciberseguridad y seguridad de la información: FISMA, Convenio de Cibercrimen de Budapest, Directiva de Seguridad en las Redes, ISO 27032; y (8) información y reporte: RDA&RRF, COREP, FINREP, FR Y-14, FR Y-9C, etc.

7. Este costoso proceso de adecuación normativa supone no obstante un elemento diferencial para los clientes bancarios, al permitirles acceder a los procesos regulados más seguros y supervisados, aspecto que las entidades financieras acabarán poniendo en valor frente al resto de nuevos competidores.

8. Por su parte, el cliente bancario se ha vuelto más exigente, está permanentemente conectado (usa el móvil 110 veces al día[22]) y consulta las redes sociales antes de comprar. Además, ya no percibe a los bancos como los únicos proveedores de servicios financieros ni a las oficinas como el canal básico de relación, y se ha acostumbrado a la

[20] SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.
[21] CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer.
[22] KPCB (2014).

personalización en la oferta de servicios. Pero al mismo tiempo, da muestras de confusión ante la abundancia y la complejidad de la oferta, lo que favorece el poder de prescripción que ejercen los líderes de opinión. Este cambio está obligando a las entidades a un replanteamiento de su oferta de servicios y de sus canales, y a adoptar una visión más centrada en el cliente, con impactos relevantes en todos los ámbitos.

9. Además, afloran nuevos competidores en el sector financiero con nuevos modelos de negocio, que provienen de sectores no sujetos a la estricta regulación bancaria pero que tienen una imagen de marca que los clientes perciben de forma muy favorable.

Data science: una disciplina emergente

10. El contexto descrito está favoreciendo la adopción en el sector financiero de una disciplina emergente, proveniente en gran medida del sector tecnológico, y necesaria para abordar la transformación a la que se enfrentan las entidades: data science.

11. El rasgo esencial de data science[23] es su carácter de ciencia, se aproxima a la extracción de valor de los datos mediante un método científico, el «proceso data science»: la formulación de una pregunta o hipótesis; la obtención de información de diversas fuentes de datos masivos y posiblemente desestructurados para responderla; la exploración de los datos mediante estadística descriptiva; la modelización del fenómeno con los datos disponibles; y la visualización y comunicación de los resultados, que confirmarán o refutarán la hipótesis o pregunta formulada.

12. Data science supone, por tanto, la evolución de la modelización tradicional, en gran medida potenciada por el entorno big data, y emplea herramientas y técnicas novedosas[24] que permiten el self-service de datos, la fusión de datos de varias fuentes, la conectividad no relacional, la utilización de la nube y la visualización interactiva de datos[25].

13. Gracias a estas capacidades, la adopción de data science permite a las entidades formular y responder preguntas antes implanteables en todos los ámbitos (riesgos, marketing, finanzas, operaciones, etc.), sobre los clientes y su entorno, e incluso sobre la propia organización.

14. A modo de ejemplo, ya es posible enriquecer los modelos de calificación crediticia con información de redes sociales y del digital footprint, mejorar los modelos de estimación de la renta mediante datos públicamente disponibles en la red cruzados con geolocalización, prevenir la fuga de clientes analizando las grabaciones de call centers mediante procesamiento del lenguaje natural o detectar el fraude y el blanqueo mediante la detección de patrones de comportamiento en los logs de actividad, entre otras muchas posibilidades.

15. La evolución hacia estas capacidades, sin embargo, no está exenta de desafíos: el coste y la dificultad de manejar volúmenes masivos de datos, los aspectos de privacidad, ética y seguridad en el manejo de los datos, la captación y formación del talento en data science, el riesgo de hacer depender de modelos automáticos muchas decisiones relevantes, y el gobierno de los datos y de los modelos.

[23]Más allá de la definición de data science recogida en la introducción, la mayoría de estudios analizan las habilidades y conocimientos que necesita el data scientist: (1) formación en Matemáticas, Física, Estadística, etc., y aprendizaje automático, algoritmia, optimización, simulación o series temporales, entre otros; (2) habilidades tecnológicas, el dominio de lenguajes estadísticos, el manejo de bases de datos relacionales y no relacionales; y (3) un conocimiento profundo del negocio, lo que es clave para el éxito de los modelos.

[24]Nuevos dispositivos y canales de relación con el cliente, nuevas plataformas, nuevos medios de pago, soluciones BPM (business process management), redes sociales como canal de contratación, seguimiento de marca y atención de quejas, sistemas distribuidos y escalables horizontalmente, infraestructura como servicio, nuevas bases de datos (NoSQL e in-memory), captura y tratamiento de datos en tiempo real, nuevas ETL y motores de consulta para datos desestructurados, herramientas de data discovery y nuevos lenguajes de programación, y nuevas herramientas de visualización y de explotación de datos on-line.

[25]Para calibrar la importancia de este perfil, nótese que el presidente Barack Obama creó en febrero de 2015 el puesto de Chief Data Scientist y nombró personalmente a Dhanurjay 'DJ' Patil.

16. A este respecto, las entidades financieras se han ido adaptando al fenómeno descrito, transformando sus procesos de generación de datos y reporte, aunque en muchos casos de forma desestructurada y como consecuencia de peticiones incrementales de los supervisores y los reguladores, de necesidades de gestión no planificadas o de procesos de integración de entidades.

17. Los reguladores han señalado las carencias en la información como una de las causas de la crisis financiera, lo que ha llevado a la publicación de normativa específica con fuertes requerimientos en calidad, consistencia, integridad, trazabilidad y replicabilidad de los datos (especialmente en el caso de riesgos[26]). Todo ello lleva a la necesidad de revisar los marcos de gobierno de los datos de las entidades.

18. El esquema de gobierno de los datos debiera sustanciarse en un marco de gestión que describiera los principios, los intervinientes (con nuevas figuras como el CDO[27]), la estructura de comités, los procesos críticos relacionados con datos e información, las herramientas (diccionario de datos, arquitectura de datawarehouses, soluciones de explotación, etc.), y el control de la calidad de los datos.

19. El gobierno de los datos presenta varios retos, entre los que se cuentan la involucración de la Alta Dirección, la definición del perímetro de datos objeto del marco de gobierno, los aspectos de privacidad en el uso de los datos y ciberseguridad (que incluye la protección contra el «hacktivismo», los ciberdelitos financieros, el espionaje y el robo de información) o la adaptación a las arquitecturas de almacenamiento novedosas, como los data lakes[28], entre otros. Pero es un hecho la relevancia que dicho gobierno ha

[26] Risk Data Aggregation and Risk Reporting Framework; ver BCBS (2013).
[27] Chief Data Officer.
[28] Repositorios masivos de datos sin transformar (en su formato origen).

adquirido en la gestión de las entidades, convirtiéndose en condición necesaria para el correcto aprovisionamiento de los datos y la información disponible y, en suma, en un eje estratégico de las entidades.

20. En el caso de los modelos, la normativa también incide en la necesidad de disponer de un marco de gobierno de los mismos[29]. Los elementos de este marco fueron tratados en detalle en anteriores publicaciones de Management Solutions[30].

21. Las entidades más avanzadas en esta materia han desarrollado ya marcos de gestión del riesgo de modelo, que regulan el inventario y la clasificación de los modelos, su documentación y un esquema de seguimiento de los mismos.

22. En todo caso, el gobierno de los modelos también presenta retos, entre los que se incluyen la involucración de la Alta Dirección, la reflexión sobre el perímetro (qué es un modelo y qué modelos deben someterse a este gobierno), la segregación de funciones (ownership, control y compliance[31]), el effective challenge o la disposición de herramientas de inventario y workflow de modelos, entre otros. Pero es incuestionable que el gobierno de los modelos requiere una involucración al primer nivel, porque de él depende la correcta toma de decisiones en las entidades.

23. En síntesis, el gobierno de los datos y de los modelos constituye un elemento estratégico para las entidades financieras, impulsado por la normativa y como respuesta al fenómeno big data. Esta cuestión impacta en diferentes ámbitos de una entidad, desde la organización, pasando por las políticas y los procedimientos, hasta las herramientas y los sistemas de información, y se conforma como un eje clave de actuación en los próximos años.

Caso de estudio: redes sociales y credit scoring

24. Para ilustrar algunos de los conceptos descritos, y como caso de estudio, se presenta un modelo de scoring crediticio que utiliza datos de redes sociales, integrado con un modelo tradicional. Se analiza también en qué medida la incorporación de dichos datos mejora el poder predictivo del modelo tradicional.

25. Entre las conclusiones destacan: (1) los datos en las redes sociales tienen una calidad muy inferior a los internos; menos de la mitad de los clientes tienen datos, y de ellos un reducido número están completos; (2) existe un problema de desambiguación[32] para identificar de forma inequívoca a cada cliente; (3) pese a ello, el poder predictivo del modelo basado en redes sociales se equipara al del modelo tradicional, y al combinarlos se eleva el poder predictivo resultante[33]; (4) para ello, se han empleado variables de la formación reglada y no reglada del cliente, su experiencia profesional, su ubicación geográfica y otros datos relativos a aficiones e intereses.

26. El estudio demuestra el potencial de la aplicación de data science al ámbito de riesgos, y es extensible a otros tipos de modelos (valoración de garantías, fidelización, renta, abandono, propensión a la compra, etc.) y fuentes de información (logs internos, bases de datos públicas, información web, etc.).

27. Data science se perfila como una materia multidisciplinar emergente que abre un nuevo campo de posibilidades al sector financiero, al aplicar un enfoque científico al fenómeno big data, aprovechando la explosión de información y de capacidades tecnológicas para aumentar la inteligencia de las entidades[34].

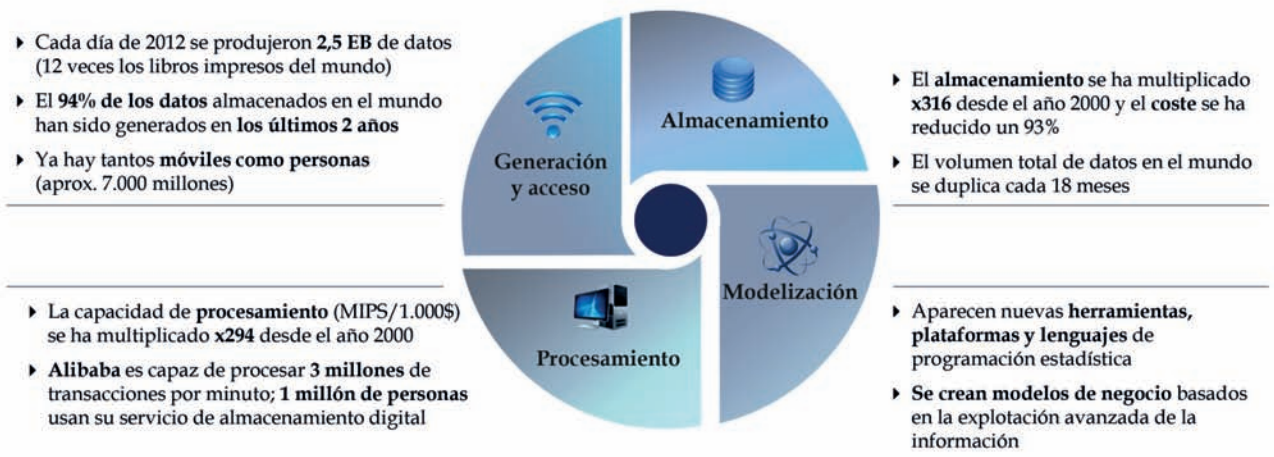
[29] OCC/Fed (2011-12).
[30] Ver Management Solutions (2014): Model Risk Management: aspectos cuantitativos y cualitativos de la gestión del riesgo de modelo.
[31] El model owner define los requerimientos del modelo y suele ser su usuario final. Control incluye la medición del riesgo de modelo, el establecimiento de límites y el seguimiento, así como la validación independiente. Compliance comprende los procesos que aseguren que los roles del model owner y de control se desempeñan de acuerdo a las políticas establecidas.
[32] Técnica que permite identificar de forma unívoca a un cliente entre varios que comparten el mismo nombre y otras características (ubicación, edad, etc.).
[33] Medido a través de la ROC (receiver operating characteristic), métrica de poder predictivo de un modelo de clasificación binaria.
[34] Capacidad de recibir, almacenar y procesar información para resolver problemas.

Un sector financiero en transformación

La tecnología y la infraestructura que se necesitan para conectar a las personas no tienen precedentes, y creemos que este es el mayor problema en el que debemos concentrarnos.

Mark Zuckerberg[35]

Fig. 2. La revolución tecnológica: una instantánea del momento actual.



La revolución tecnológica

El sector financiero hace frente a una revolución tecnológica de una magnitud y una extensión sin precedentes, que está transformando su actividad de forma sustancial.

El principal rasgo de esta revolución es su aceleración. Como predice la Ley de Moore, la potencia tecnológica, medida en varios ejes, está creciendo de forma exponencial, y las predicciones apuntan a que este fenómeno continuará. Esta revolución se manifiesta en cuatro dimensiones: generación y acceso a la información (incluyendo movilidad), almacenamiento, procesamiento y modelización.

Para entender este fenómeno y sus implicaciones, es necesario observarlo sobre cuatro dimensiones: generación y acceso a la información, almacenamiento, procesamiento y modelización (Fig. 2).

Fig. 3. Algunas magnitudes sobre generación e intercambio de información.

Fuente: Pethuru (2014)

En esta sección se tratarán los tres primeros; el detalle sobre modelización y data science se abordará en la próxima sección.

Generación y acceso a la información

La primera faceta de este fenómeno es el incremento de la velocidad a la que se generan los datos digitales. Las estimaciones de la velocidad y la aceleración de la generación de datos digitales varían según los analistas, pero todos coinciden en que se trata de una aceleración exponencial y en todos los ámbitos; por citar algunos ejemplos[36,37](Fig. 3):

- En 2012, se crearon 2,5 exabytes de datos cada día; esta tasa ha continuado aumentando.
- Más del 90% de todos los datos que hoy existen han sido creados en los dos últimos años.
- Cada minuto se suben 12 horas de vídeo a YouTube.
- Cada día se crean 12 terabytes de tweets en Twitter.
- Se producen 5.000 millones de transacciones financieras al día.
- En 2012 había 2.400 millones de usuarios de Internet en el mundo, casi la mitad de ellos en Asia…
- … que intercambiaron 144.000 millones de correos electrónicos al día, de los cuales el 69% eran spam.
- Y también en 2012, Facebook superó los 1.000 millones de usuarios, y se espera que en 2016 tenga más usuarios que la población de China.

[35]Mark Elliot Zuckerberg (n. 1984). Cofundador y director general de Facebook
[36]IBM (2014a).
[37]Pingdom (2015).

Fig. 4. ¿Cuánto es un zettabyte? Algunas magnitudes orientativas.

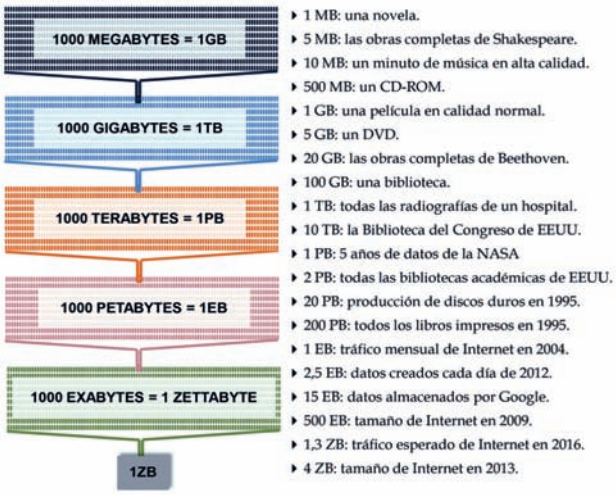


Fig. 5. Volumen, variedad y velocidad de generación de datos en el ámbito comercial.

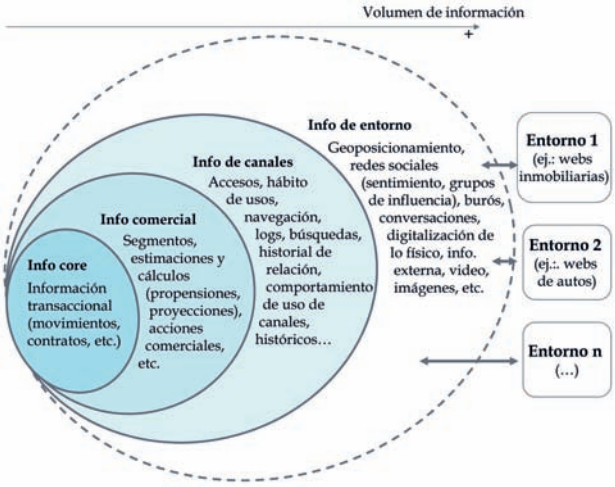
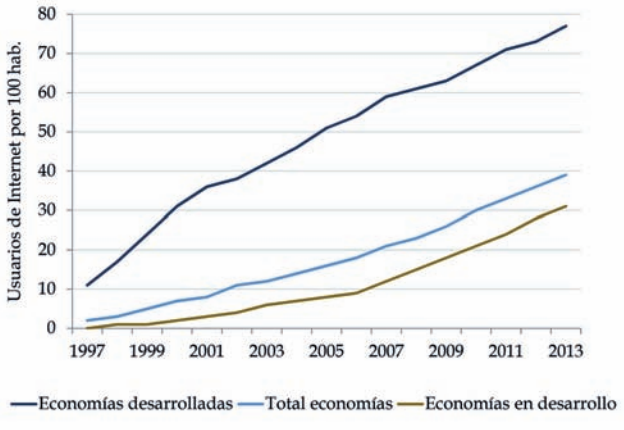


Fig. 6. Usuarios de Internet por cada 100 habitantes.



Fuente: International Telecommunications Union (2012).

En palabras de la Federal Big Data Commission, a la que el Gobierno de Estados Unidos ha encomendado la misión de comprender el fenómeno big data en las agencias gubernamentales[38]:

> *En los últimos años, los Gobiernos federales, estatales y locales hacen frente a un maremoto de cambio como resultado del incremento drástico en el volumen, la variedad y la velocidad de los datos en sus propios entornos y en todo el ecosistema gubernamental. […]*
>
> *Desde el año 2000, la cantidad de información que el Gobierno federal captura se ha incrementado exponencialmente. En 2009, el Gobierno de Estados Unidos produjo 848 petabytes de datos, y solo el sistema de Sanidad alcanzó los 150 exabytes. Cinco exabytes (10^{18} gigabytes) de datos contendrían todas las palabras pronunciadas por todos los seres humanos. A esta tasa, el fenómeno big data en la Sanidad estadounidense pronto alcanzará la escala de los zettabytes (10^{21} gigabytes), y poco después los yottabytes (10^{24} gigabytes).*

Estas cifras vertiginosas son difíciles de imaginar; veáse la Fig. 4 para tener una referencia orientativa de cada una de estas magnitudes.

Por otra parte, estos datos ya no se generan solo de forma estructurada; al contrario, el 80% de los datos que se crean cada día son desestructurados: vídeos, imágenes, correos electrónicos, etc., y provienen de una amplia variedad de fuentes nuevas: redes sociales, sensores, registros de navegación por Internet, logs de actividad, registros de llamadas, transacciones, etc.

En otras palabras, el fenómeno big data es una explosión en el volumen, la variedad y la velocidad de generación de datos, lo que ha sido llamado «las tres V del big data» (a lo que algunos autores añaden la cuarta V, de «veracidad»). Como ejemplo, en el ámbito comercial de las entidades financieras se está viviendo una expansión en estos tres ejes, que desde la información core transaccional ha ido creciendo, pasando por datos comerciales y provenientes de canales, hasta llegar la información proveniente del entorno del cliente, de gran riqueza, variedad y heterogeneidad (Fig. 5).

En cuanto al acceso a la información, si bien subsisten diferencias entre las economías desarrolladas y las emergentes, y algunos países apenas tienen acceso a Internet, la tendencia es clara, y se estima que en pocos años se alcanzará el pleno uso de Internet en casi todo el mundo (Fig. 6).

Asimismo, se observa una explosión en el acceso a la información mediante dispositivos móviles. El número de teléfonos móviles ya se equipara al número de habitantes, y en las economías desarrolladas lo supera en un 21%[39], mientras que las economías en desarrollo se acercan a la paridad, con nueve líneas móviles por cada diez habitantes.

[38] Federal Big Data Commission (2014).
[39] International Telecommunication Union (2014).

Las entidades financieras no han sido ajenas a esta proliferación de dispositivos, y en los países desarrollados se está observando una evolución constante de la banca digital (Fig. 7), que en gran medida viene impulsada por el propio cliente, contiene a la banca móvil y va más allá. Partiendo del e-banking y de la integración multicanal de la primera década del siglo XXI, y pasando por el fenómeno de la omnicanalidad que se ha desarrollado en los últimos años, de cara al futuro la tendencia viene marcada por la llamada «Internet de las cosas»: el predominio y la ubicuidad de los dispositivos inteligentes, con capacidad para generar información digital sobre su utilización.

En este sentido, en Estados Unidos casi el 60% de los clientes ya operan solo por canales digitales, pasan 400 veces más tiempo en estos canales que en la oficina y solo el 1% de las transacciones ocurren en la oficina. Pese a ello, el 90% de las ventas siguen ocurriendo en las oficinas (Fig. 8).

La evolución, por tanto, es heterogénea, y gran parte de las entidades se encuentran todavía en la fase de adaptación a la banca móvil; la mayoría han desarrollado aplicaciones para dar acceso móvil a los principales servicios. De acuerdo con el estudio llevado a cabo por la Reserva Federal de Estados Unidos[40], en este país más del 50% de los usuarios de un smartphone han utilizado la banca móvil en los últimos 12 meses, y los usos más comunes (Fig. 9) son consultar el saldo o transacciones recientes (93%) y realizar transferencias entre cuentas propias (57%).

La tendencia es clara: la banca digital está ganando cuota y, dado que los clientes de menos de 29 años utilizan los canales digitales casi cuatro veces más que los tradicionales, es previsible que esta expansión continúe en los próximos años. No obstante, la monetización de este fenómeno presenta evidentes recorridos de mejora.

[40]Fed (2014).

Fig. 7. Evolución de la banca digital.

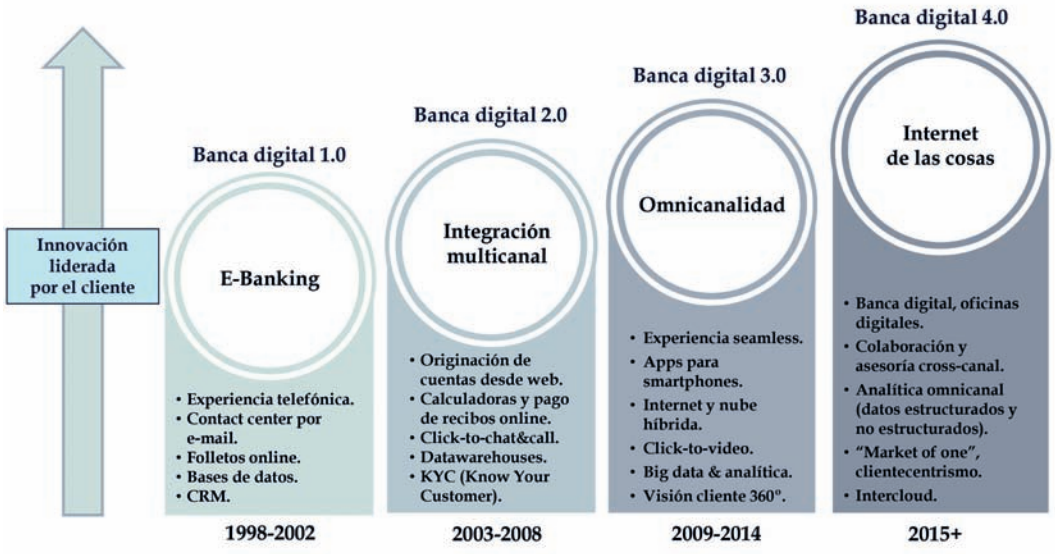
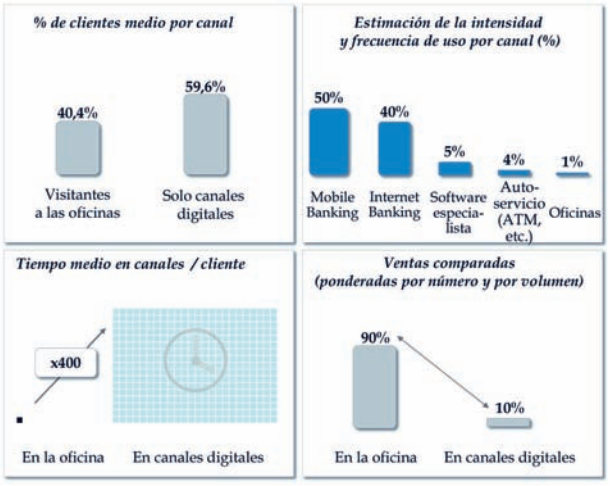


Fig. 8. Transformación de los canales en las entidades financieras.



Fuente: Digital Leadership GmbH (2014).

Fig. 9. Usos de banca móvil.

Fuente: Fed (2014).

Almacenamiento

De forma paralela a la generación y el acceso a la información, la capacidad de almacenamiento también está creciendo de forma exponencial, siguiendo la Ley de Moore, y su coste unitario continúa descendiendo al mismo ritmo (Fig. 10). Mientras que en 1980 almacenar un gigabyte requería dispositivos por valor de 10 millones de dólares, hoy apenas cuesta diez centavos de dólar en una diminuta fracción de un disco duro SSD.

Los dispositivos de almacenamiento han evolucionado de forma acelerada, desde la cinta magnética de 1920, pasando por los tubos de rayos catódicos de 1940, el primer disco duro de 1956 (Fig. 11), la cassette de 1963, la memoria DRAM de 1966, los disquetes (floppy disks) de la década de 1970, los CD de 1980, los zips y DVD de 1994 y 1995, las tarjetas flash de 1995, las tarjetas MMC de 1997, los pendrives de 1999, las tarjetas SD de 2000, el Blu-Ray de 2003, hasta la memoria sólida y el almacenamiento en la nube de la década de 2010. Obsérvese cómo la cantidad de nuevos formatos por década ha aumentado de forma exponencial, como también lo ha hecho la capacidad de los dispositivos.

Todo ello ha llevado a que la cantidad de información almacenada en el mundo haya crecido de forma masiva en los últimos años; se estima que en 2012 había un total de 2,75 zettabytes de datos almacenados digitalmente, y que esta cifra llega a los 8 zettabytes en 2015 y continúa su trayectoria ascendente[41].

En la industria financiera, la evolución de los sistemas de almacenamiento se ha desarrollado en paralelo a la necesidad de recopilar y gestionar grandes cantidades de información. A finales de la década de 1970, las entidades comenzaron la implantación de servidores host, entornos tecnológicos donde se recibe, procesa y almacena toda la información generada a través de la gestión de transacciones, que es su objetivo fundamental. Posteriormente, la incorporación de sistemas informacionales permitió desacoplar la consulta masiva de información de los procesos operacionales, abriendo la puerta a cantidades muy superiores de datos y a su historización.

Esta tendencia se materializó en la creación de datawarehouses, repositorios de información estructurada con cuatro características principales:

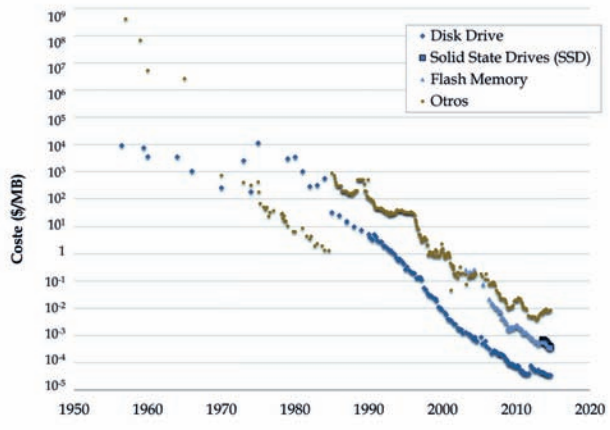
- Orientado: los datos almacenados se encuentran estructurados y agrupados por temáticas.
- Actualizado: permiten incorporar nueva información a lo largo del tiempo.
- Histórico: mantienen un registro de datos históricos, no es necesario eliminarlos.
- Integrado: aseguran la consistencia de los datos registrados por distintos sistemas.

A comienzos del siglo XXI, la industria financiera incorporó nueva tecnología para modernizar y optimizar el almacenamiento de información con grandes sistemas dedicados, que se caracterizan por emplear software y hardware diseñados específicamente para la explotación informacional. En paralelo, se comenzó la implantación de software de consulta más sofisticado, lo que permite mayor libertad al usuario (herramientas OLAP, query & reporting, etc.).

En la actualidad, las entidades financieras ya manejan en torno a 1,9 petabytes en sus sistemas informacionales[42], lo que supone un desafío para la arquitectura implantada. Como consecuencia, se está viviendo una nueva revolución en los sistemas de almacenamiento de información: las plataformas de almacenamiento distribuido. Esta tecnología permite el almacenamiento masivo de datos desestructurados y su gestión mediante una arquitectura nodular.

[41] SiliconAngle (2014).
[42] DeZyre (2014).

Fig. 10. Coste de almacenamiento, en dólares por megabyte (obsérvese que el eje Y es logarítmico).



Fuente: McCallum (2014).

Fig. 11. Disco duro IBM 350, de 1956, con una capacidad de 5 megabytes.

En efecto, algunas de las principales entidades financieras, y en Estados Unidos al menos tres de las cinco mayores de ellas[43], ya han adoptado plataformas de almacenamiento distribuido y comienzan a explotar su potencial de almacenamiento y procesamiento de datos, si bien de forma limitada todavía.

Procesamiento

Al igual que la generación y el almacenamiento de la información, la potencia de procesamiento también está experimentando el mismo crecimiento acelerado[44]. Teniendo en cuenta el coste, la capacidad de procesar instrucciones por segundo por cada 1.000 dólares de procesador se ha multiplicado por casi 300 desde el año 2000. Esto permite que algunos retailers sean capaces de procesar millones de transacciones comerciales por minuto, lo que está en la esencia de su modelo de negocio.

Por otra parte, la aparición de la computación distribuida ha permitido combinar las capacidades de numerosos procesadores (núcleos) para ejecutar operaciones de forma paralela. Se estima[45] que Google disponía en 2012 de unos 7,2 millones de núcleos en más de 1,8 millones de máquinas, y que con su potencia combinada era capaz de ejecutar unos 43.000 billones de operaciones por segundo, unas cuatro veces más que la máquina más potente del mundo (Fujitsu K). El mismo estudio calcula que Amazon podía alcanzar los 240 billones de operaciones por segundo.

Los principales players tecnológicos, conscientes de que su capacidad de procesamiento distribuido es en sí un servicio valioso, comercializan el acceso a computación distribuida en sus respectivas nubes; la sobreabundancia de esta capacidad les permite alquilarla a costes inferiores a un dólar por hora[46].

En el sector financiero, por su parte, la exigencia de procesamiento de la actividad tradicional bancaria (a través de las oficinas o los terminales punto de venta) es soportada por la tecnología implantada actualmente de forma razonablemente satisfactoria. Sin embargo, los nuevos canales, como la banca digital, requieren un aumento de la capacidad de procesamiento transaccional en paralelo.

En el mundo informacional, las sofisticadas técnicas de modelización hacen que las entidades requieran de una mayor capacidad de cálculo y tratamiento masivo de información. Para ello, la adopción de técnicas de computación paralela o distribuida, muy ligadas a la propia estructura de almacenamiento, permite a las entidades sacar el máximo rendimiento de la información disminuyendo el tiempo de procesamiento.

Una nueva commodity: los datos

Esta explosión en la capacidad de generar, almacenar y procesar información, y acceder a ella en cualquier momento y lugar mediante dispositivos móviles, está causando un fenómeno novedoso: los datos se han convertido en una nueva commodity. En efecto, los datos se generan, almacenan y procesan a un coste muy reducido, son fungibles por cuanto pierden vigencia con rapidez, y son la materia prima que, transformada, da lugar a servicios de todo tipo.

Sin embargo, esta nueva commodity tiene dos rasgos particulares: al igual que la energía, se ha vuelto indispensable para el funcionamiento de la mayoría de los negocios, incluidos los servicios financieros; y, como todas las commodities, requiere profesionales y herramientas especializadas para su procesamiento. Este es precisamente el campo de data science.

[43] Goldman (2014).
[44] Ver gráfico en la introducción.
[45] Pearn (2012).
[46] Para 8 núcleos con 30 GB de RAM en Amazon, Google y Windows, por ejemplo.

Fig. 12. Coyuntura macroeconómica y factores de riesgo.

Contexto del sector financiero

Aunque esta revolución tecnológica tiene un impacto relevante en todos los sectores, es esperable que el sector financiero sea uno de los más beneficiados por la adopción de data science como eje estratégico. Se trata de la industria que maneja la mayor cantidad y calidad de información de sus clientes y targets, y por tanto tiene un enorme potencial para extraer conocimiento e incorporarlo en su propuesta de valor, lo que es diferencial respecto a otros sectores.

Sin embargo, en el sector financiero esta revolución tecnológica se produce en un contexto singular, caracterizado por una difícil coyuntura macroeconómica, un entorno regulatorio exigente y un cambio en el patrón de comportamiento de los clientes, factores que no están impactando del mismo modo en otros sectores.

Coyuntura macroeconómica

En el aspecto macroeconómico (Fig. 12), se mantiene el carácter dual de la economía mundial (países desarrollados vs. emergentes) en términos de crecimiento, presiones inflacionistas y flujos de inversión, de manera que continúan ciertos patrones en las principales macromagnitudes (Fig. 13) que afectan a la evolución del negocio bancario tanto en sus fuentes de financiación como en sus inversiones y márgenes financieros.

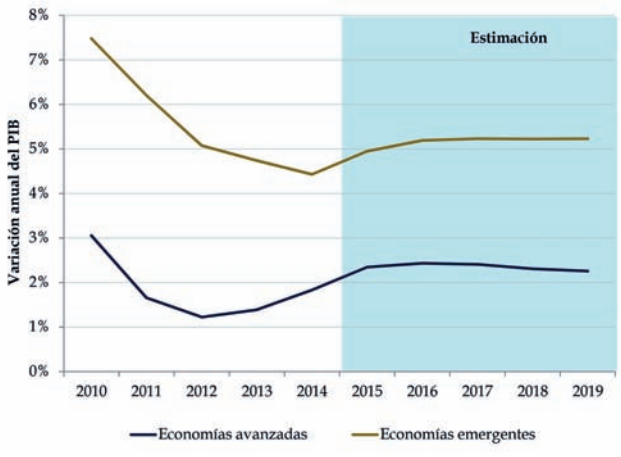
En el caso de las economías avanzadas, el escenario prolongado de bajos tipos de interés ha generado un aumento relativo del precio de las acciones, la compresión de los spreads y un descenso generalizado de la volatilidad hasta volver a niveles anteriores a la crisis. Sin embargo, esto no ha generado un repunte de la inversión, lo que contrasta con la evolución del ahorro, que se ha elevado y provoca una mayor debilidad de la demanda privada. De esta forma, en las economías avanzadas se espera que este estancamiento (que ha sido calificado de «estancamiento secular»[47]) se mantenga durante varios años, aunque de manera heterogénea entre países.

Además, cada vez existe mayor evidencia de que el crecimiento potencial de las economías avanzadas empezó a disminuir antes del estallido de la crisis financiera como consecuencia del envejecimiento de la población en la fuerza laboral y el débil crecimiento de la productividad de los factores[48].

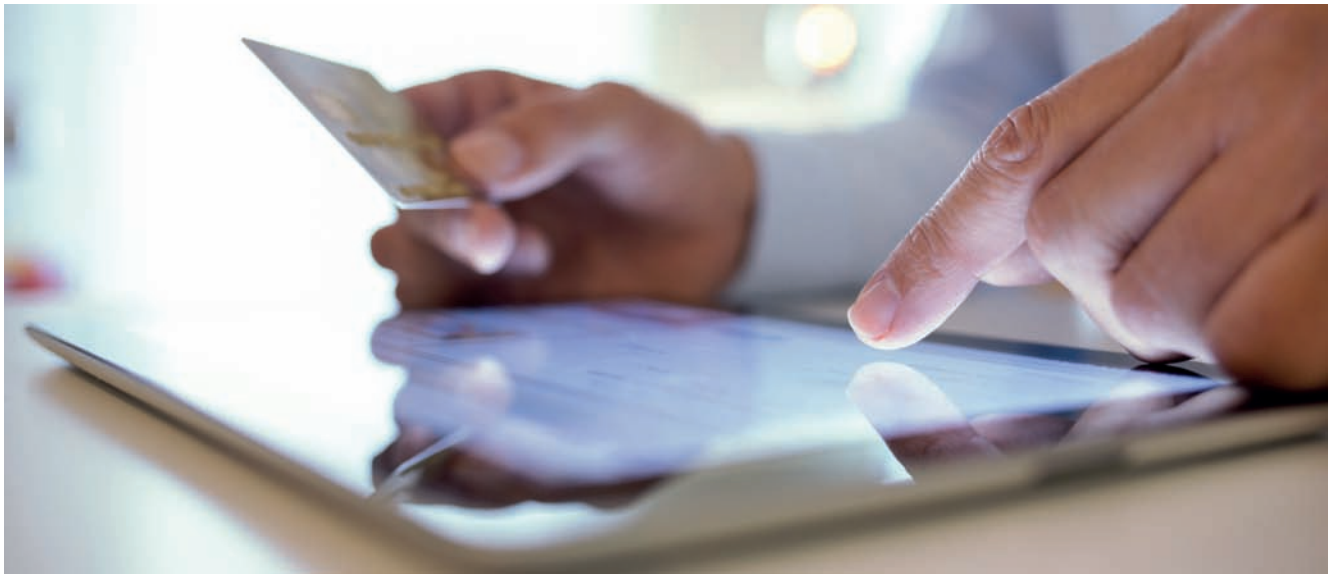
En algunas economías avanzadas, además, la inflación es baja o existen indicios de deflación, a lo que se suma que los tipos de referencia no tienen margen de reducción. Todo ello tiene un impacto negativo en la evolución del crédito y en los márgenes de las entidades financieras.

[47] IMF (2014).
[48] Para el caso de Estados Unidos, véanse por ejemplo Fernald (2014), Gordon (2014) y Hall (2014).

Fig. 13. Crecimiento del PIB en economías emergentes vs. avanzadas.



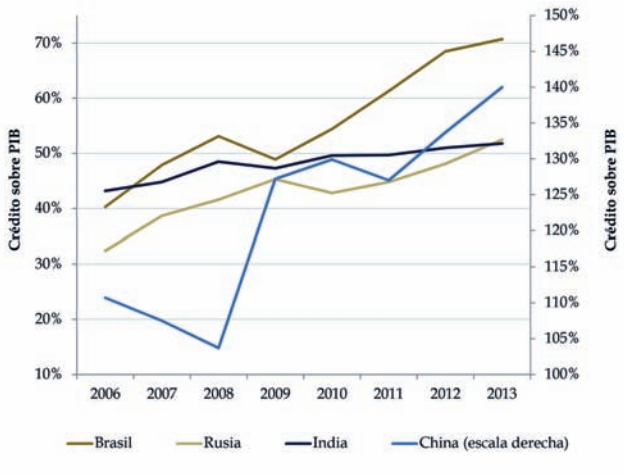
Fuente: IMF (2014).



En el caso de las economías emergentes, el crecimiento se ha ralentizado, aunque aún se mantienen niveles relativamente elevados respecto a las economías avanzadas. El consumo privado contribuye en gran medida a este crecimiento, aunque también existe una dependencia elevada de la inversión pública y de las políticas fiscales expansivas.

Aunque la expansión del crédito bancario se desacelera en algunos mercados emergentes (Brasil, India, Rusia), se mantienen tasas de crecimiento de dos dígitos (Fig. 14). Por otra parte, la tasa de morosidad está aumentando de forma generalizada en las economías emergentes debido a varias causas, entre las que se cuentan los problemas en determinados sectores (como el sector minero en Perú o el sector público en Chile y Brasil[49]) y la incorporación de nuevos clientes antes no bancarizados y con un peor perfil crediticio.

Fig. 14. Crédito sobre PIB en economías emergentes.



Fuente: IMF (2014).

Por último, coexisten dos elementos que generan incertidumbre en este entorno. Por un lado, la dependencia económica de China, cuya desaceleración provocaría una contracción masiva de las exportaciones en el resto del mundo, la reducción del precio de las materias primas y la caída de los índices de confianza de consumidores y empresas. A este respecto, se observan en China riesgos sobre el crecimiento a causa de la capacidad de producción excedentaria y una sobreabundancia de crédito, que son los principales impulsores de su crecimiento.

Por otro lado, la normalización de la política monetaria en Estados Unidos, Japón y la Unión Europea, que en los últimos años se ha expandido a países como Chile, México y Perú, supone un riesgo por el posible efecto deflacionista y por la atracción de flujos de inversión de los países emergentes.

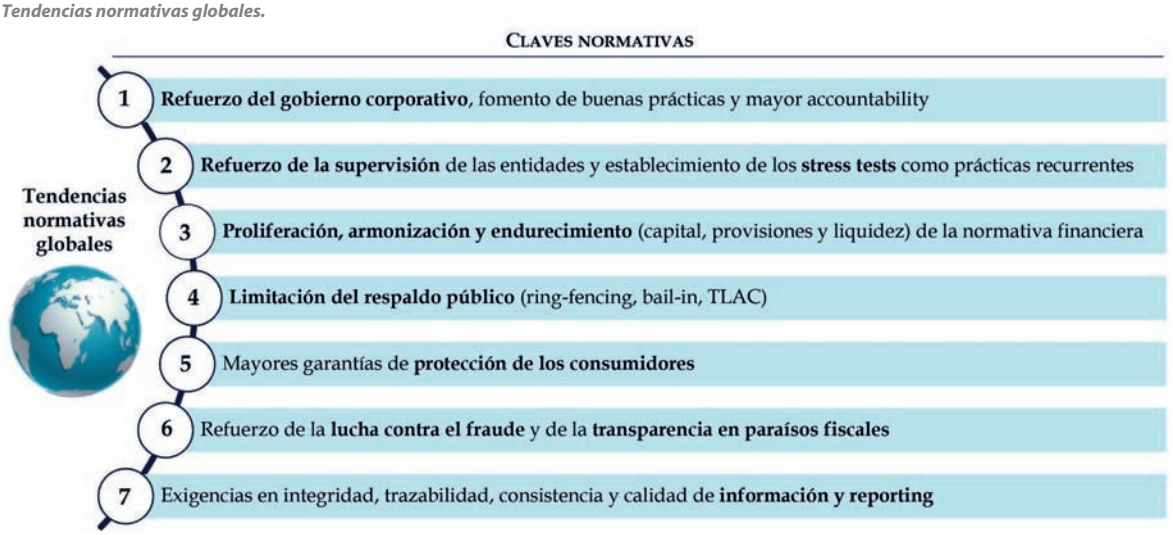
Esta coyuntura macroeconómica está generando un estrechamiento de los márgenes en el sector financiero, pero también introduce presión sobre el capital y la liquidez. La consecuencia es que las entidades han intensificado la gestión de la rentabilidad, del capital y de la estructura de balance, dotándolas de mayor inteligencia analítica y una visión de riesgos, con atención a las previsiones económicas y su potencial impacto.

Entorno regulatorio

El sector financiero está experimentando una notable proliferación, que cabe calificar de «tsunami», tanto de regulación supranacional como de normativa local en los ámbitos financieros que han tenido mayor influencia en la crisis iniciada en 2007: contable, de supervisión prudencial, de conducta y cumplimiento, de gobierno corporativo, de protección al consumidor y de riesgos en sentido amplio.

[49]BBVA Research (2014).

Fig. 15. Tendencias normativas globales.



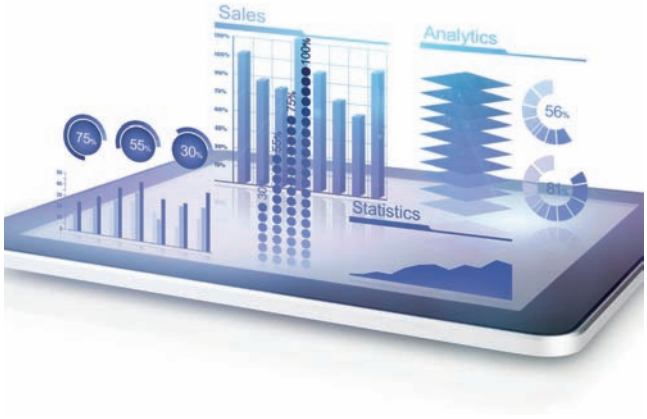
Asimismo, se observa una tendencia de armonización de la normativa entre los distintos países, a lo que está contribuyendo de forma decisiva la constitución de organismos reguladores y supervisores supranacionales, como la Autoridad Bancaria Europea (EBA), el Mecanismo Único de Supervisión (SSM) del Banco Central Europeo o el Financial Stability Board (FSB), entre otros.

Al mismo tiempo, el carácter de las normas está pasando a ser más intrusivo y prescriptivo, y deja menos espacio para adaptaciones e interpretaciones. Como ejemplo, en la Unión Europea se ha adoptado Basilea III a través de una combinación de un Reglamento (por tanto, de aplicación inmediata en todos los países de la Unión) y una Directiva (que debe ser traspuesta a las normativas locales); mientras que Basilea II fue adoptada solo en la forma de una Directiva.

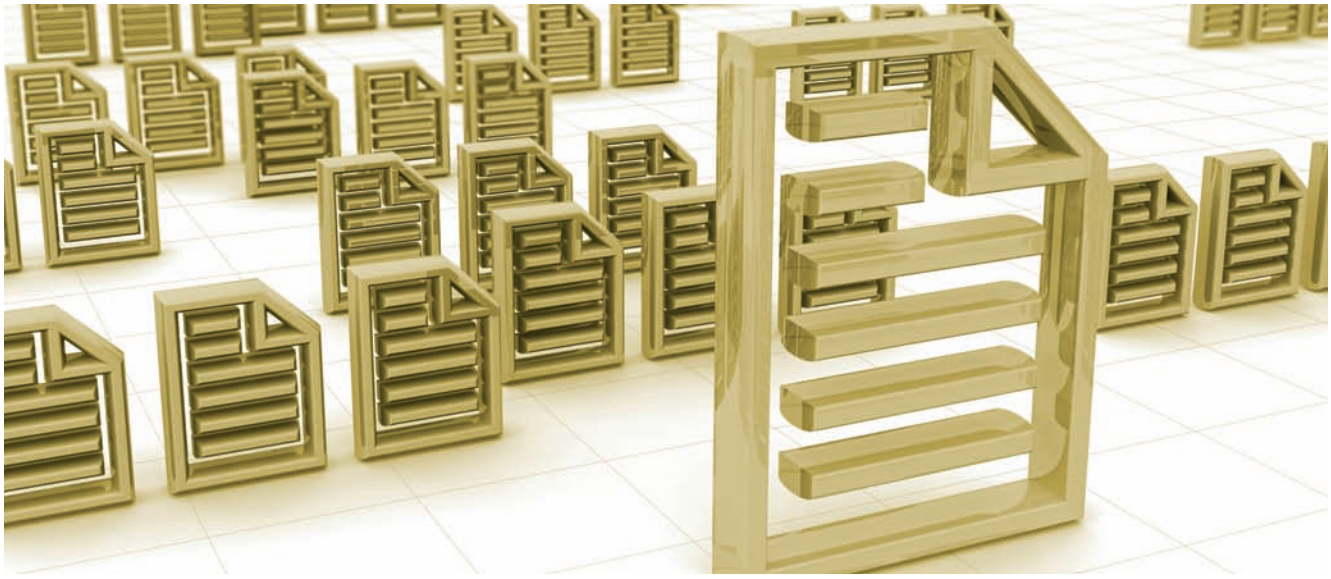
En concreto, esta proliferación y armonización de normativa financiera se está materializando en más normas (y de carácter más restrictivo) en varios ámbitos, entre los que cabe destacar (Fig. 15):

- Capital y liquidez: como consecuencia de Basilea III, aparecen mayores requerimientos de capital (tanto en cantidad como en calidad), un nuevo ratio de apalancamiento y dos ratios de liquidez (a corto y a largo plazo[50]). Asimismo, se revisan y simplifican los requerimientos de capital por riesgo de crédito, de mercado y operacional[51].

- Refuerzo de la supervisión prudencial: se establecen directrices comunes[52] para la supervisión de las entidades, que refuerzan los procesos SREP, ICAAP e ILAAP[53] (especialmente en Europa, con la entrada en vigor del Mecanismo Único de Supervisión en noviembre de 2014). Asimismo, se robustecen y establecen como prácticas recurrentes los stress tests supervisores[54].



[50] Management Solutions (2012).
[51] A este respecto, el Comité de Supervisión Bancaria de Basilea emitió durante 2014 varios documentos que revisan el método Estándar de riesgo de crédito, simplifican el cálculo de capital por riesgo operacional, fijan suelos de capital y revisan el cálculo de capital en la cartera de negociación, entre otros.
[52] EBA (2014) y ECB (2014).
[53] SREP: Supervisory Review and Evaluation Process; ICAAP: Internal Capital Adequacy Assessment Process; ILAAP: Internal Liquidity Adequacy Assessment Process.
[54] Management Solutions (2013).



- Limitación del respaldo público: especialmente en las economías avanzadas, se pretende que en ningún caso el Estado tenga que volver a rescatar a entidades con fondos públicos por ser sistémicas, lo que se conoce como «el fin del too-big-to-fail». Para ello, se obliga a las entidades a tener planes de recuperación y resolución[55], y a disponer de suficientes pasivos con capacidad de absorción de pérdidas (TLAC, MREL); y, en la Unión Europea, se crea una autoridad para gestionar la resolución de las entidades inviables (Single Resolution Board). Por otra parte, en Estados Unidos y en el Reino Unido, y posteriormente en el resto de la Unión Europea, se potencia la regulación sobre ring-fencing, que obliga a la separación jurídica entre las actividades mayoristas y la banca tradicional.

- Refuerzo del gobierno corporativo: se imponen mayores exigencias al Consejo de Administración y a la Alta Dirección sobre la aprobación y la supervisión del cumplimiento de la estrategia de negocio, el apetito al riesgo y el marco de gestión de riesgos, y se crean nuevas figuras clave (CRO, CDO, CCO[56], funciones corporativas de Risk MI, etc.).

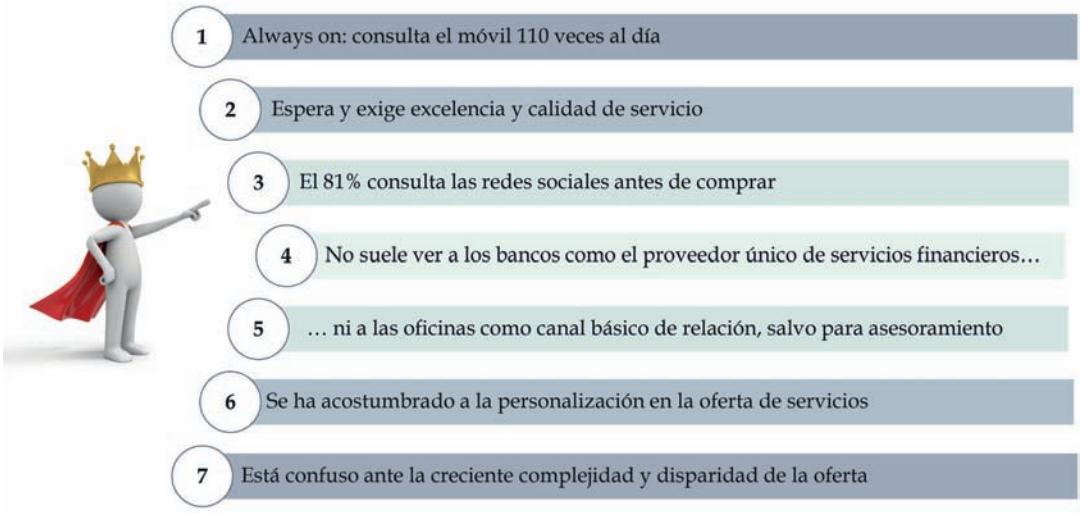
- Protección de los consumidores: como consecuencia de los escándalos en el sector financiero asociados a productos, canales de distribución, tecnología de pagos, abuso de mercado y blanqueo de capitales, aparece una regulación más intensiva y prescriptiva[57] que exige el refuerzo de la función de Cumplimiento (recursos, medios, capacidades y líneas de reporte), de control de calidad (mystery shopping) y de la política de gestión de quejas, así como el refuerzo de la medición, gestión, control, supervisión y reporte del riesgo de conducta frente a mercados y clientes. Esta tendencia es abanderada por el Reino Unido, con la creación de una autoridad específica, la Financial Conduct Authority, que solo entre 2013 y 2014 impuso casi 2.000 millones de libras en multas por temas de conducta[58].

- Lucha contra el fraude y los paraísos fiscales: debido al incremento del fraude por el uso intensivo de los canales electrónicos y por los constantes cambios en las organizaciones, se regula la necesidad de contar con un control intensivo del fraude interno y externo. Aparecen políticas agresivas por parte de algunos países para evitar fraude fiscal por parte de sus ciudadanos (por ejemplo, FATCA[59]). Se incrementan las exigencias en la lucha contra el blanqueo de capitales (sanciones elevadas[60]), lo que demanda adaptaciones importantes en los procesos y sistemas de las entidades.

- Ciberseguridad: se desarrolla normativa específica para combatir el incremento de los ataques a la seguridad de las entidades («hacktivismo», ciberdelitos financieros, espionaje, robo de información, etc.): en Estados Unidos, el Federal Information Security Management Act (FISMA), entre otras; en Europa, el Convenio contra el Cibercrimen de Budapest o la Directiva de Seguridad en las Redes (SRI); y en el ámbito global, la norma ISO 27032, que proporciona directrices específicas sobre ciberseguridad.

[55] En la Unión Europea, mediante la Directiva de Recuperación y Resolución de Entidades Bancarias (BRRD), resumida en European Commission (2014).
[56] CRO: Chief Risk Officer; CDO: Chief Data Officer; CCO: Chief Compliance Officer; MI: Management Information.
[57] Mortgage Market Review y Retail Distribution Review en el Reino Unido, Directiva sobre Abuso de Mercado y sobre Blanqueo de Capitales e Informe sobre Tendencias en los Consumidores en la Unión Europea, entre otros.
[58] FCA (2015).
[59] Foreign Account Tax Compliance, ley federal que requiere a las entidades financieras de todo el mundo que reporten a la agencia fiscal de EE.UU. las cuentas en el extranjero de las personas estadounidenses.
[60] Como la multa de 1.900 millones de dólares a HSBC por fallos en sus controles contra el blanqueo de capitales; ver Bloomberg (2013).

Fig. 16. El cliente de hoy.



- Información y reporte: los procesos de generación de información y reporte de riesgos han ido perdiendo efectividad por diversas razones, lo que se ha señalado como una de las causas de la crisis, ante lo cual los reguladores han emitido normativa[61] que obliga a una revisión integral de los datos y el reporte de riesgos para garantizar su calidad, integridad, trazabilidad y consistencia, conocida como «RDA&RRF[62]». El objetivo es reforzar las capacidades de agregación de datos de riesgos y las prácticas de presentación de informes, para permitir así mejorar la gestión y la toma de decisiones. Por otra parte, se unifican los criterios de reporte de capital, liquidez e información financiera (COREP, FINREP). Todo ello obliga a las entidades a una revisión profunda de los sistemas y procesos de generación de información y reportes.

Hacer frente a este «tsunami regulatorio» está suponiendo un enorme coste para las entidades y las está obligando a ambiciosos procesos de transformación. No obstante, esta transformación es un elemento claramente diferencial de las entidades, ya que permite ofrecer a los clientes la seguridad de disponer de los procesos más seguros, regulados y supervisados de todas las industrias digitales. Se trata de un aspecto clave que las entidades acabarán poniendo en valor frente al resto de nuevos competidores.

Comportamiento del cliente

En el plazo de unos pocos años, el sector financiero ha visto transformarse el comportamiento de sus clientes: están más informados, más conectados, tienen más cultura financiera y muestran una demanda más ajustada (Fig. 16). Exigen un servicio que les ofrezca comodidad, velocidad, personalización y trato justo, además de acceso desde dispositivos móviles.

El cliente se caracteriza por sus elevadas expectativas; compara la calidad de servicio proporcionada por su entidad financiera con la de proveedores de otros sectores (tecnológico, retail, etc.) y espera un nivel similar de prestaciones y de respuesta en tiempo real.

También es competente y activo en el uso de redes sociales, que utiliza tanto para cotejar información (el 81% consulta las redes sociales antes de comprar) como para difundir su disconformidad ante una experiencia deficiente.

Los estudios[63] demuestran que la experiencia del cliente está positivamente correlacionada con la retención. Pese a ello, revelan que, aunque se va observando una mejora paulatina de la calidad de la experiencia del cliente con las entidades financieras, todavía no es suficiente: más del 50% de los clientes manifiestan su intención de cambiar de banco antes de seis meses.

Más aún, el cliente ya no contempla a los bancos como los únicos proveedores de servicios financieros ni a las oficinas como el canal básico de relación (salvo para asesoramiento). Todo ello está obligando a las entidades financieras a un replanteamiento integral de su oferta de servicios y de sus canales, y a adoptar, en suma, una visión «clientecéntrica» o «360°», que tiene impacto en todos los ámbitos, desde los procesos a los sistemas, pasando por la organización, el control de riesgos o la planificación de negocio.

No obstante, se percibe también en los clientes una creciente confusión ante la complejidad y la diversidad de la oferta. Esto está llevando a las entidades a adoptar una visión comercial orientada a simplificar la oferta, adecuándola a las necesidades del cliente (revisando así su catálogo de productos y servicios).

De forma simultánea y muy relacionada con el cambio del perfil del cliente, se está observando un efecto disruptor y novedoso:

[61] BCBS 239 (2013).
[62] Risk Data Aggregation and Risk Reporting Framework.
[63] EFMA (2013).

la entrada de nuevos competidores en el sector financiero, algunos de ellos provenientes de otros sectores (Fig. 17), que satisfacen necesidades no cubiertas del todo por la banca tradicional.

Esta nueva competencia puede clasificarse en tres familias:

- Competencia conocida que ofrece nuevos servicios (como bancos 100% móviles).
- Nuevos players financieros que antes no existían en el mercado y que cubren nichos desatendidos.
- Nuevos modelos de negocio, que provienen de otros sectores; en especial, la tecnología, la venta retail y las telecomunicaciones.

En el caso de la competencia proveniente de otros sectores, esta amenaza es particularmente lesiva por varias razones: por una parte, los nuevos competidores no están sujetos a la estricta regulación bancaria; por otra, tienen modelos de negocio de una gran eficiencia, con costes muy reducidos; en tercer lugar, disponen de «ecosistemas» que agrupan muchas necesidades del cliente: dispositivos físicos, entorno de trabajo, música, películas, libros, revistas, etc., donde resulta natural integrar los servicios financieros; y por último, tienen imágenes de marca que son percibidas de forma muy positiva por los clientes.

La banca todavía percibe a estos competidores como una amenaza moderada (Fig. 18), y es cierto que por el momento se están concentrando en determinados nichos, como los medios de pago (por ejemplo, PayPal, Google Wallet o Apple Pay), y que sus barreras de entrada regulatorias son elevadas para acceder a los servicios core de depósito y crédito. Sin embargo, por su tamaño e influencia, estos competidores tienen el potencial de salvar las barreras y alterar el mercado de forma significativa en un futuro cercano.

Fig. 17. Nueva competencia en el sector financiero.

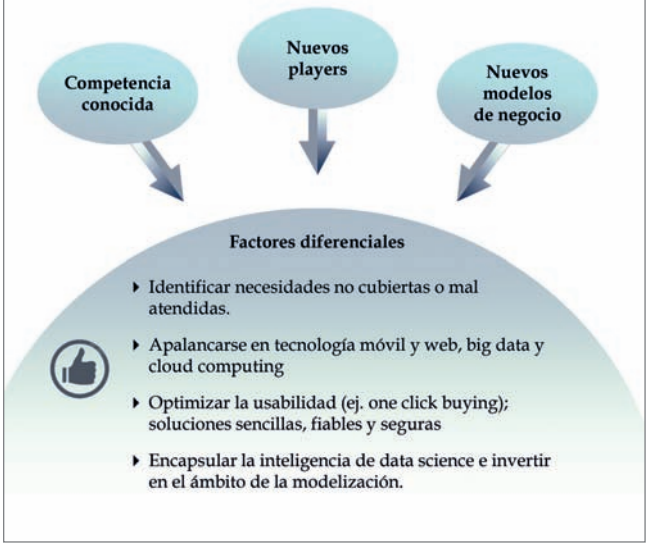
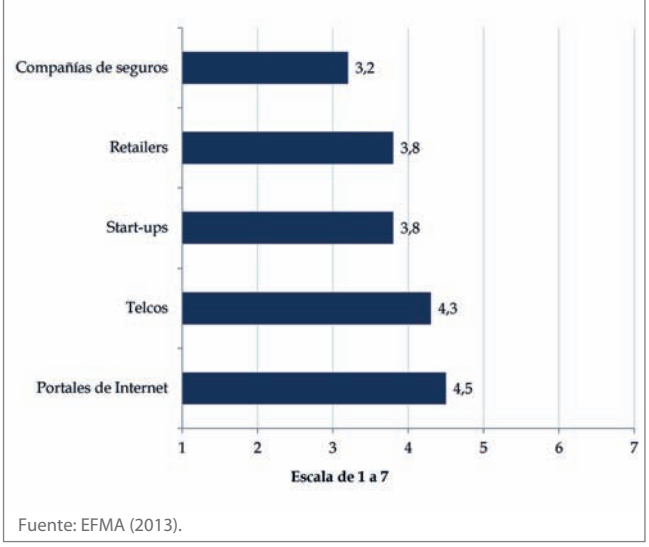


Fig. 18. Percepción de amenazas por parte de las entidades financieras.



Fuente: EFMA (2013).

Data science: una disciplina emergente

Toda compañía tiene big data en su futuro
y toda compañía estará en el negocio de los datos tarde o temprano
Thomas H. Davenport[64]

¿Qué es data science?

La comoditización de los datos y el gobierno de los datos y de los modelos que se impone como consecuencia conllevan, como sucede con cualquier materia prima, la aparición de nuevas herramientas y técnicas para procesarlos. El conjunto de estas herramientas y técnicas conforma una disciplina que, si bien no es nueva, tiene un carácter emergente y está recibiendo una creciente atención en todos los sectores, incluido el financiero: data science.

Definición

La naturaleza novedosa del interés por esta disciplina, junto con su carácter innovador y ligado a las tecnologías de big data, hace que no exista una definición formal y comúnmente aceptada de data science. El Center for Data Science de la New York University se aproxima así al término[65]:

> *Data science, o la ciencia de los datos, es el estudio de la extracción generalizable de conocimiento a partir de los datos mediante el uso combinado de técnicas de aprendizaje automático, inteligencia artificial, matemáticas, estadística, bases de datos y optimización, junto con una comprensión profunda del contexto de negocio.*

Sin embargo, la mayor parte de acercamientos al concepto de data science pasan más bien por describir las habilidades y conocimientos que necesita el profesional para ser considerado un data scientist:

> *Un profesional con entrenamiento y curiosidad para tomar decisiones en el mundo del big data. […] Realiza descubrimientos estando completamente sumergido en datos y es capaz de estructurar grandes cantidades de datos sin formato, así como identificar fuentes de información ricas aunque incompletas y combinarlas para obtener conjuntos de datos mucho más completos y de gran valor.*[66]

> *Un profesional con un conocimiento profundo de datos que puede trabajar de manera efectiva con datos de una manera escalable.*[67]

> *Una evolución del analista de negocio: tiene una base sólida en computación y sus aplicaciones, matemáticas, modelización, estadística y análisis de datos. Sin embargo, el data scientist destaca por la agudeza de su sentido de negocio del sector y su capacidad de comunicación.*[68]

Como se puede apreciar, el data scientist es un profesional con un perfil multidisciplinar, y en concreto combina al menos tres características:

- Una formación de base en alguna ciencia o disciplina cuantitativa, que incluya conocimientos de aprendizaje automático, algoritmia, optimización, simulación, series temporales y modelos de asociación y clasificación, entre otros.

- Unas habilidades tecnológicas avanzadas, que incluyen el dominio de lenguajes de programación estadística, pero también conocimientos técnicos para la extracción eficiente, el uso y el almacenamiento de información, el manejo de bases de datos relacionales y no relacionales y la capacidad para extraer datos de Internet y procesar grandes cantidades de información.

[64] Thomas Hayes Davenport (n. 1954). Académico estadounidense especialista en gestión del conocimiento e innovación de procesos de negocio. Fue nombrado uno de los tres mejores analistas de negocio del mundo en 2005 por la revista Optimize.
[65] Dhar [Center for Data Science, New York University] (2013).
[66] Harvard Business Review (2012).
[67] Berkeley (2015).
[68] IBM (2014b).

- Y, lo que posiblemente marca una mayor diferencia con otros perfiles similares, un conocimiento profundo del negocio en el que desarrollan su labor como data scientists.

El tercer aspecto, la experiencia en el negocio, es particularmente relevante porque acerca de forma definitiva las capacidades analíticas a, en el caso de la banca, el conocimiento financiero; esto es clave para que los modelos se integren de forma plena en la gestión, lo que es una condición indispensable para su éxito y buen uso (Fig. 19).

Para calibrar la importancia de este perfil, nótese que en Estados Unidos el presidente Barack Obama creó en febrero de 2015 el puesto de Chief Data Scientist, y nombró personalmente a Dhanurjay 'DJ' Patil[69] para este puesto, con la misión de impulsar nuevas aplicaciones de big data en todas las áreas del Gobierno[70].

El proceso data science

Por otra parte, como señalan algunos autores[71], la característica más relevante de data science es precisamente su cualidad de ciencia: ante la cantidad masiva de datos a la que se enfrenta una entidad, la aproximación de un data scientist es formular una teoría (una pregunta o una hipótesis) proveniente de la realidad del negocio, y aplicar sus conocimientos y habilidades sobre los datos para verificarla o descartarla. Esto es lo que se conoce como el «proceso data science», que se compone de cinco etapas (Fig. 20):

1. **Formulación:** se plantea una pregunta relevante para el negocio, que deberá ser respondida empleando los datos y las técnicas disponibles. Uno de los cambios esenciales que aporta la disciplina de data science es precisamente la formulación de cuestiones o hipótesis que antes resultaba imposible verificar y que, con la abundancia actual de datos, herramientas y técnicas, abre nuevas posibilidades. Por ejemplo, «a juzgar por sus comentarios en las últimas llamadas al call center, ¿cuál es la probabilidad de que cada uno de mis clientes cambie de banco en los próximos seis meses, y qué debería hacer para evitarlo?».

2. **Obtención de datos:** se localizan todas las fuentes disponibles, incluyendo fuentes estructuradas (datawarehouses, datamarts, etc.) y no estructuradas (logs de actividad, redes sociales, etc.). La cantidad masiva de datos y, en su caso, su naturaleza desestructurada son el núcleo del desafío computacional de todo el proceso. En esta fase también se tratan los aspectos legales, como la protección de datos, la confidencialidad o las cláusulas de restricción de uso.

3. **Exploración de datos:** se aplican técnicas de estadística descriptiva para realizar un primer análisis exploratorio. En este punto, data science aporta técnicas nuevas de exploración que facilitan la labor y, dado el potencial de paralelización en estas tareas, se beneficia de las plataformas de computación distribuida.

4. **Modelización:** la construcción y la validación tradicional de modelos se ven enriquecidas por algoritmos de alto rendimiento desarrollados ad hoc para grandes volúmenes de información, así como por tipos de modelos alternativos a los clásicos, que aportan mejoras sobre ellos en términos de estabilidad, robustez y aprovechamiento de la riqueza de la información, como los random forests y las support vector machines, entre otros (Fig. 21). Para ello, los vendors tradicionales de herramientas analíticas están completando sus suites de productos, y surgen nuevos lenguajes de programación estadística, muchos de ellos de código abierto.

[69]Reputado data scientist que ha trabajado en LinkedIn, eBay, PayPal y Skype, entre otros, y a quien se atribuye la creación del término «data scientist».
[70]Wired (2015).
[71]O'Neil y Schutt (2013).

Fig. 19. Características del data scientist.

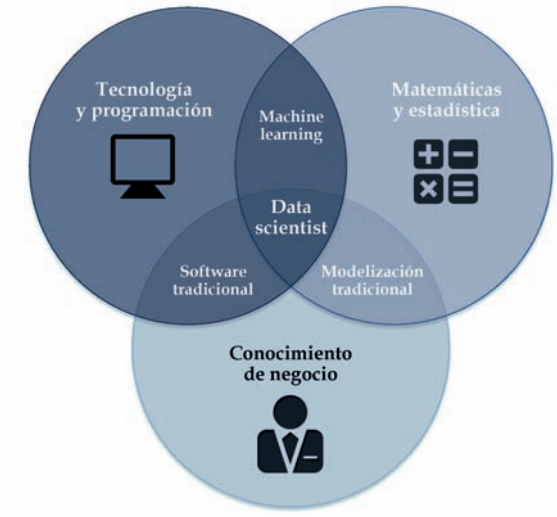
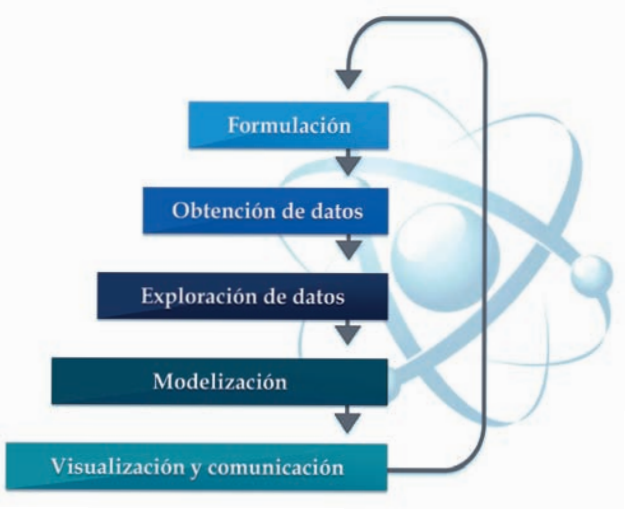
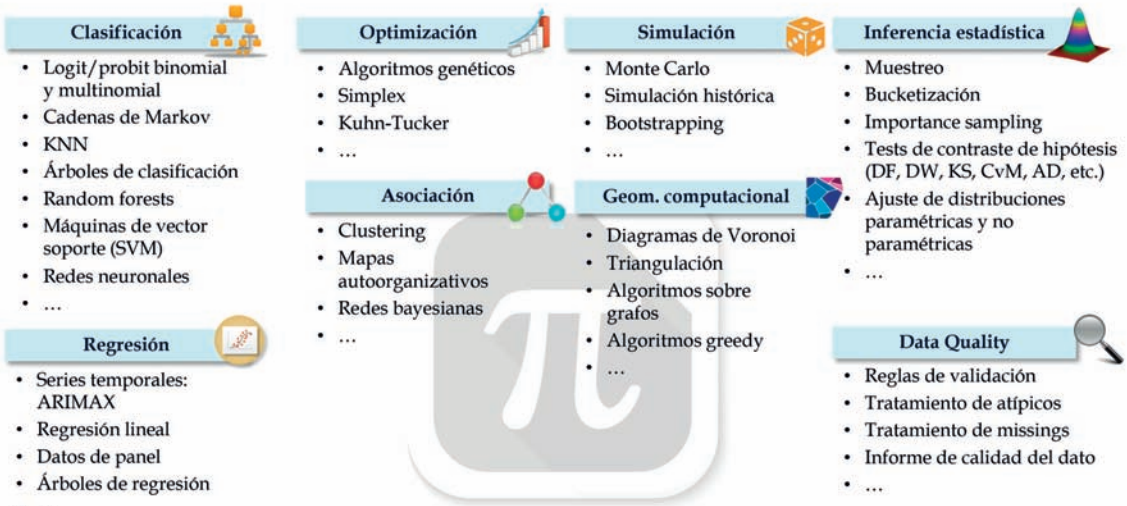


Fig. 20. El proceso data science.



Fuente: adaptado de Harvard (2014).

Fig. 21. Técnicas de data science.



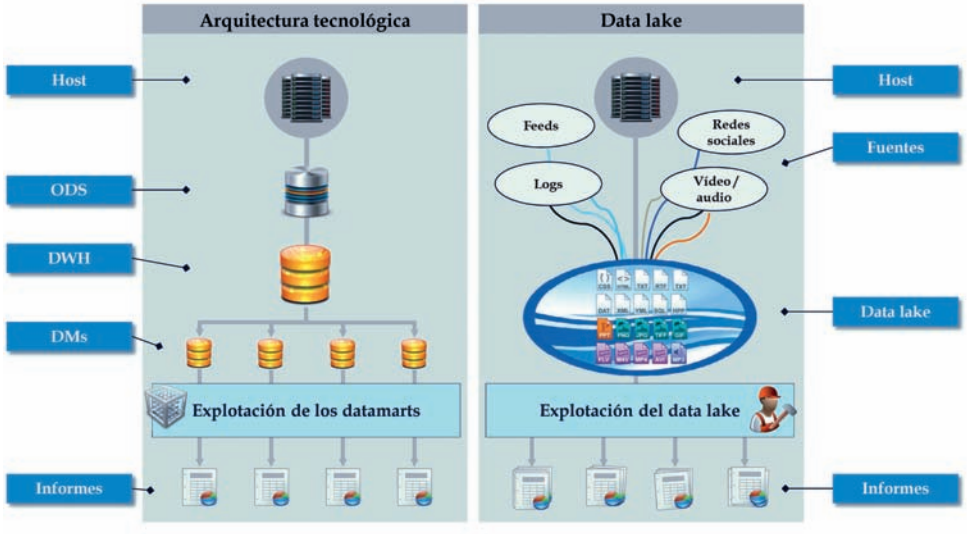
Data lake: una nueva arquitectura informacional

Ante el volumen creciente de información y la naturaleza heterogénea de las fuentes, es necesario contar con nuevas técnicas y tecnologías capaces de almacenar de forma optimizada la información. En este escenario surge el concepto de «data lake» o «data pool», un inmenso repositorio de datos en su formato de origen con las siguientes características:

- ***Integridad.*** Se trata de un único repositorio donde se almacena toda la información, asegurando la trazabilidad de la información.
- ***Flexibilidad.*** El data lake almacena cualquier información de interés, independientemente del formato, eliminando cualquier estándar concreto sobre captura y almacenamiento de datos.
- ***Independencia.*** Elimina en gran medida las dependencias con los departamentos de TI en tanto que el volcado de información es flexible y el usuario puede realizar las extracciones necesarias directamente desde el data lake.

El uso y la explotación de data lakes y las técnicas de data science asociadas permite a los data scientists trabajar con la información sin tratar, en su formato original, permite escalabilidad horizontal y elimina los límites en manejo de grandes volúmenes de información sin estructurar. Esta arquitectura no sustituye necesariamente a la tradicional; al contrario, en general es complementaria a los datawarehouses y datamarts.

Data lake vs. arquitectura tradicional.



5. **Visualización y comunicación:** dos de los aspectos que más atención han recibido en data science, la visualización de los resultados y su comunicación de forma inteligible a terceras partes, son dos cualidades que se esperan en un data scientist, y también se ven potenciadas por nuevas herramientas que integran el código con la documentación de forma intuitiva y natural.

Aunque podría parecer un enfoque de sentido común, esta aproximación a los datos mediante un método científico conlleva un cambio de metodología de trabajo: con frecuencia los analistas abordan el problema de manera inversa (lanzar modelos al azar sobre una cantidad masiva de datos en busca de relaciones ocultas), lo que puede suponer un consumo elevado de recursos sin un objetivo claramente establecido ni una hipótesis que contrastar.

En suma, data science supone la evolución de la modelización tradicional en el entorno big data, y abre nuevas posibilidades antes implanteables, que incluso llegan a transformar los modelos de negocio establecidos. La adopción de data science como un eje estratégico de desarrollo es una prioridad para el sector tecnológico y, como se verá, también comienza a serlo para el sector financiero.

Herramientas de data science

La comoditización de los datos también está favoreciendo la evolución y la aparición de nuevas herramientas tecnológicas de data science que facilitan su tratamiento, análisis y visualización. Todos los vendors tradicionales están impulsando ecosistemas analíticos, y continuamente aparecen start-ups con propuestas novedosas, así como herramientas y lenguajes open source (Fig. 22), lo que convierte a este mercado en un foco de competencia y desarrollo acelerado.

Estas herramientas permiten superar las limitaciones de los sistemas tradicionales, que resultaban insuficientes ante la heterogeneidad de los datos (no podían analizar información estructurada y no estructurada conjuntamente), su desfragmentación (la información estaba distribuida en silos diferentes bajo modelizaciones imprecisas), la dependencia de IT (los usuarios de negocio debían delegar en áreas de Sistemas la tarea de recopilar y organizar la información en datawarehouses, lo que conllevaba un excesivo tiempo de preparación de datos) y, en general, la falta de adaptación a las fuentes de datos actuales (los sistemas tradicionales no se integraban con redes sociales, call centers, sensores, posicionamiento geográfico, etc., ni eran adecuados para tratar el volumen de información generada por ellos).

Así, entre las principales aportaciones que estas herramientas incorporan, cabe destacar[72]:

- Self-service: en el esquema tradicional, solo unos pocos profesionales de la entidad, muy especializados, tenían acceso a los datos y a las herramientas analíticas. En el esquema de data science, aparecen herramientas de una mayor sencillez, que permiten a más profesionales explorar, analizar y visualizar datos. Este fenómeno se da en todos los sectores y contribuye a potenciar las capacidades analíticas de los profesionales.

- Fusión de datos: la fusión de datos se refiere a la combinación de información proveniente de distintas fuentes y en formatos diferentes. En el esquema tradicional, esto se hace mediante procesos de ETL y el despliegue de modelos de datos que pueden llegar a ser muy complejos. En el esquema de data science más avanzado, los datos se vuelcan en un data lake común, bien documentado con un diccionario de datos, y las herramientas son capaces de tomar los archivos y fusionarlos en un tiempo reducido.

[72] Adaptado de Gigaom Research (2014).

Fig. 22. ¿Quién es quién en data science? Algunas de las principales herramientas y lenguajes open source.

hadoop	**Hadoop:** infraestructura de programación de código abierto que permite almacenar, procesar y analizar grandes volúmenes de datos, diseminándolos en grandes clusters de servidores y que se procesan en paralelo.
HIVE	**Hive:** sistema de datawarehouse sobre Hadoop, desarrollado por Facebook. Se utiliza para lanzar queries y manejar grandes cantidades de datos en un almacenamiento distribuido. Emplea un lenguaje similar a SQL denominado HiveQL (HQL), lo que evita a los usuarios la necesidad de usar Java o APIs de Hadoop.
	Pig: aplicación open source creada por Yahoo y construida sobre Hadoop, centrada en el procesamiento de grandes volúmenes de datos (estructurados y semi-estructurados) en modo batch.
	Impala: plataforma open source que proporciona consultas SQL (procesamiento y análisis de datos) en tiempo real. Utiliza los componentes de Hadoop.
Lucene	**Lucene:** API open source que sirve como motor de búsqueda. Se utiliza para el indexado y búsqueda de datos sobre datasets acotados. Está implementado en las web apps de Twitter, LinkedIn, Apple, AOL, Eclipse, etc.
mahout	**Mahout:** proyecto open source para implementar algoritmos escalables y distribuidos de machine learning. Es un framework de Java y utiliza Hadoop. Lo utilizan Adobe, AOL, Intel, LinkedIn, Twitter y Yahoo, entre otros.
R	**R:** lenguaje y entorno de programación estadístico muy versátil por su modularidad (se pueden instalar paquetes de funcionalidades avanzadas ya creadas por otros programadores) y su naturaleza open source.
python	**Python:** lenguaje de programación de propósito general que hace énfasis en la legibilidad y la intuición del código. En data science, está especialmente indicado para la captura de datos de fuentes online.

Nuevas capacidades tecnológicas

El fenómeno big data aporta nuevas capacidades tecnológicas, que se estructuran en cuatro capas:

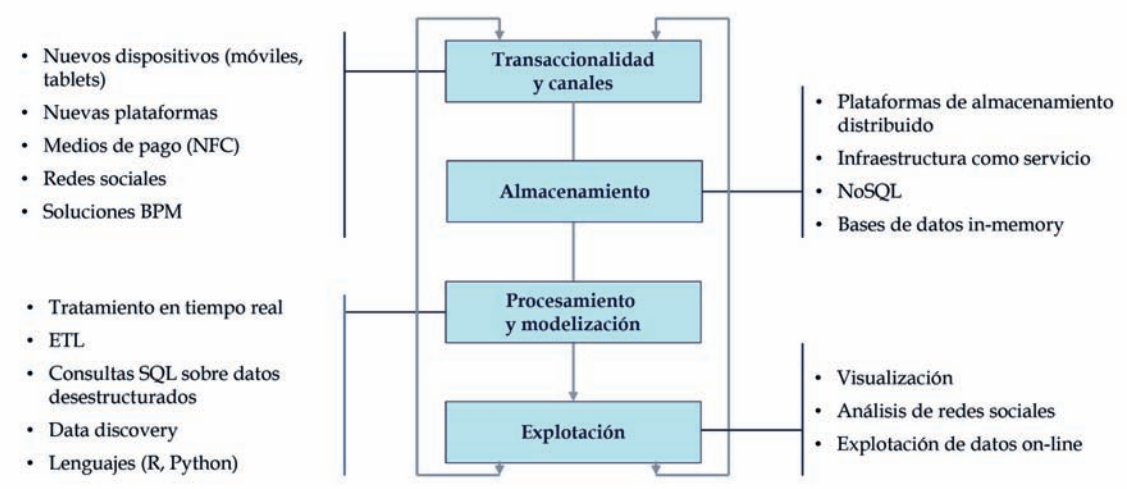
- En la capa **transaccional**, aparecen nuevos dispositivos y canales de relación con los clientes (móvil, tablet, etc.), nuevas plataformas para el desarrollo de aplicaciones y la prestación de servicios (CRM, apps para móviles con objeto de cubrir nuevas necesidades financieras y operativas del cliente, etc.), nuevas tecnologías de medios de pago como el pago por móvil (como NFC) y soluciones BPM (Business Process Management) para la integración de plataformas y automatización de procesos, como la contratación on-line o la gestión documental. Asimismo, se potencian las redes sociales como nuevo canal de relación con el cliente, que se perfilan como potencial canal de contratación y a través de las cuales se realizan análisis de sentimiento de marca y la atención a quejas y reclamaciones.

- En la capa de **almacenamiento**, aparecen nuevos sistemas de almacenamiento diseñados para ejecutarse en hardware de bajo coste, ofreciendo alta disponibilidad, tolerancia a fallos (con datos replicados en varios nodos), con escalabilidad horizontal y que permiten el tratamiento masivo de datos. Surge la infraestructura como servicio en las modalidades de cloud público, cloud privado o cloud híbrido. Aparecen nuevas bases de datos (NoSQL) orientadas al tratamiento batch de grandes volúmenes de información y nuevas estructuras de datos: bases de datos columnares y documentales, y también nuevas bases de datos in-memory para el tratamiento de información en memoria con una alta velocidad de respuesta a consultas.

- En la capa de **procesamiento y modelización,** surgen herramientas para la captura y tratamiento de información en tiempo real, así como nuevas ETL para transformación de datos desestructurados, como Pig, y nuevos motores de consulta de datos desestructurados en lenguaje SQL. También aparecen herramientas para la implementación de mecanismos que garanticen el gobierno del dato: catalogación, transformación, trazabilidad, calidad, consistencia y control de acceso, y herramientas de data discovery para la extracción de conocimiento de forma libre desde fuentes diversas, estructuradas y no estructuradas. Y, por último, aparecen nuevas técnicas, algoritmos matemáticos y lenguajes para el reconocimiento de patrones en los datos, el análisis predictivo, la implementación de los modelos y el aprendizaje automático (machine learning).

- Por último, en la capa de **explotación** aparecen nuevas herramientas de análisis multidimensional y reporting con capacidad de acceso a grandes volúmenes de información en memoria, soluciones específicas para el análisis de la información proveniente de redes sociales y para la explotación de flujos de datos on-line para la toma de decisiones y desencadenamiento de eventos en tiempo real, como la detección del fraude, la detección y el lanzamiento de eventos comerciales o los scorings de riesgos, entre otros muchos usos.

Nuevas capacidades por capa de arquitectura tecnológica.



- Conectividad no relacional: a diferencia de las herramientas tradicionales, que preveían solo la conexión con bases de datos, las herramientas de data science permiten conectar con otras fuentes de información: NoSQL, plataformas de computación distribuida, e información de redes sociales, en la nube o en sistemas de software as a service, de importancia creciente para las entidades.

- La nube: una de las novedades más relevantes es la utilización de la nube en el ámbito analítico, que aporta la funcionalidad de almacenamiento de datos en su versión más básica, lo que permite desligar la labor del data scientist de una ubicación o un servidor concretos y facilita el trabajo desde distintas geografías. En algunos casos integra además los servicios de ETL, visualización de datos y despliegue en dispositivos móviles, creando un ecosistema analítico completo, que simplifica la labor de análisis.

- Visualización de datos: uno de los rasgos diferenciales de la disciplina de data science, ligado al conocimiento de negocio, es la visualización de los datos. Algunas herramientas van más mucho más allá de la generación de gráficos con anotaciones, y ya son capaces de producir de forma automática cuadros de mando y presentaciones interactivas que permiten al usuario profundizar de forma dinámica en los análisis.

Data science en el sector financiero

El sector financiero está experimentando la misma explosión en materia de generación y necesidad de almacenamiento de datos que otros sectores, y tiene el potencial de extraer un conocimiento profundo tanto sobre sus clientes como sobre su entorno (competencia, actividad económica sectorial, geolocalización, etc.) al que antes resultaba imposible acceder.

Para ello, las entidades están desarrollando una serie de capacidades tecnológicas y metodológicas, que están abriendo un horizonte de nuevas posibilidades en el sector, aunque al tiempo plantean una serie de desafíos.

Nuevas oportunidades en el sector financiero

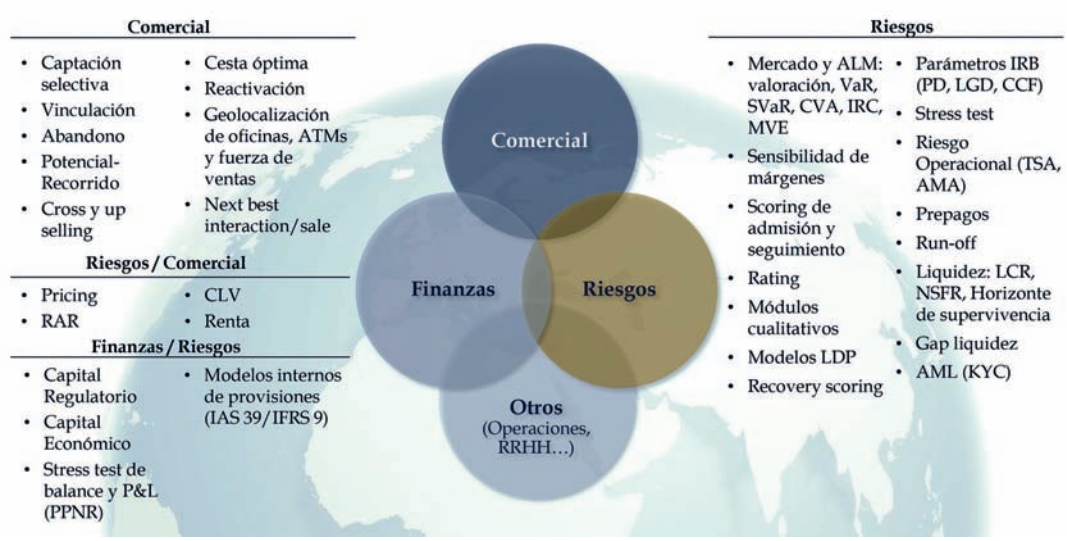
Gracias a estas nuevas capacidades, los modelos tradicionales se ven ampliados y enriquecidos en todos los ámbitos de las entidades financieras: riesgos, marketing, finanzas, operaciones, etc. (Fig. 23), contribuyendo a aprovechar toda la información disponible para mejorar la toma de decisiones en estos ámbitos y, en algunos casos, a automatizar algunos de los procesos.

Así, a modo de ejemplo se pueden citar algunas aplicaciones emergentes de data science en el sector, basadas en datos de redes sociales, geolocalización, multimedia o logs, entre otros, que antes no recibían atención:

- Credit scoring con digital footprint: la calificación crediticia de los particulares, normalmente basada en unas pocas variables (entre 5 y 20, dependiendo de la cartera y la información disponible), se ve enriquecida y ampliada con la información presente en las redes sociales y en Internet en general, lo que se conoce como el «digital footprint» o «huella digital». Se construyen modelos basados en esta información, que mejoran sustancialmente el poder predictivo y por tanto el control de la morosidad, especialmente en la población de no clientes, tradicionalmente peor calificada por las limitaciones en la disponibilidad de datos.

- Prevención del abandono mediante procesamiento del lenguaje natural (NLP): las grabaciones de los call centers, que se solían emplear casi exclusivamente para el control interno de calidad, se revelan como una fuente valiosa de prevención de la fuga de clientes. Mediante modelos de reconocimiento del habla, se transcriben de forma automática todas las conversaciones con los clientes, y sobre los textos resultantes se aplican técnicas de text

Fig. 23. Principales modelos en las entidades financieras.



mining y lingüística computacional para identificar la probabilidad de que un cliente concreto decida cambiar de entidad en las próximas semanas. Para ello, se comienza por un análisis lexicográfico (la detección de ciertas palabras que se asocian con la intención de cambio), pero de forma experimental también se avanza hacia un nivel semántico, donde el modelo comprende patrones más complejos de significado en el discurso del cliente.

- Modelos de renta y propensión basados en redes sociales cruzadas con geolocalización: se cruza la información disponible de un cliente en las redes sociales con datos censales, inmobiliarios, de Google Maps y de otras fuentes, y con ello se realizan estimaciones mejoradas de su nivel de renta, su capacidad de ahorro, sus necesidades de productos financieros, el valor del inmueble donde reside (utilizado también para valorar el subyacente en una titulización), etc. Todo ello complementa la información disponible y contribuye a mejorar las acciones comerciales sobre los clientes.

- Personalización de promociones para reducir los costes de adquisición: se recopila y cruza toda la información disponible sobre cada cliente en sus transacciones por todos los canales y los datos de redes sociales, obteniendo así una visión 360° del cliente. Con ello, se reduce al máximo el nicho objetivo de cada promoción, y en consecuencia se aumenta el ratio de captación de clientes y se reducen los costes de adquisición.

- Campañas de bonificación mediante análisis de transaccionalidad de tarjetas: partiendo de los movimientos de las tarjetas, se pueden conocer las costumbres de los titulares, los momentos en que realizan compras, sus viajes, tiendas habituales, etc., y proponer campañas de bonificación en el momento adecuado, con mayor probabilidad de éxito.

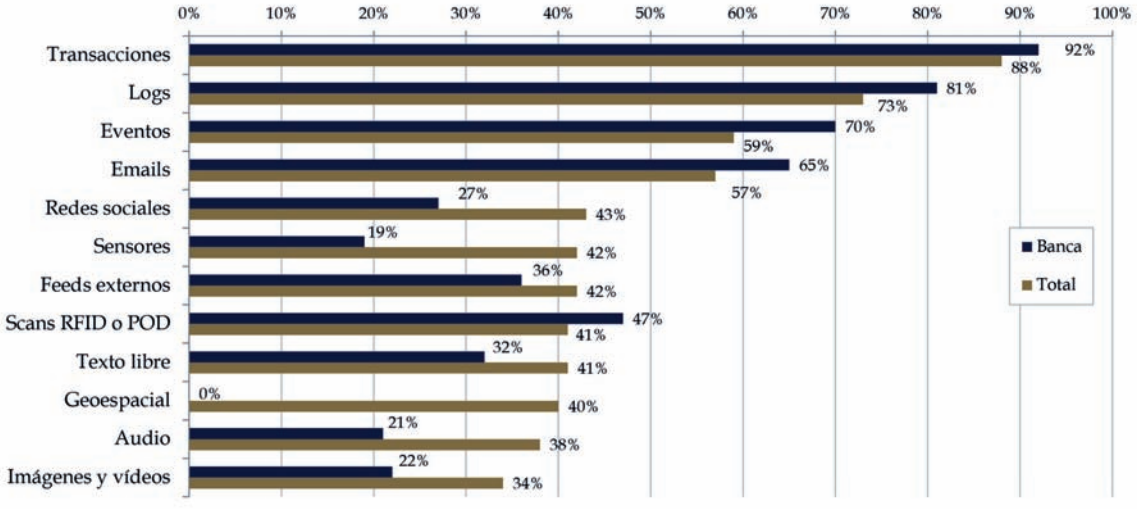
- Detección del fraude y del blanqueo de capitales, y mejora del control de calidad, mediante logs: los registros de actividad o logs son grandes ficheros poco estructurados donde constan todas las acciones que realiza un cliente o un empleado en una plataforma digital (ordenador, dispositivo móvil, cajero automático, etc.). La detección de patrones de comportamiento en los logs es compleja, porque requiere un tratamiento de información especialmente masivo, pero puede servir para identificar intentos de fraude (tanto interno como externo) y de blanqueo de capitales. Asimismo, es la base de una modalidad nueva de control de calidad, potencialmente muy extensa, que puede abarcar desde los tiempos de respuesta en una oficina hasta la dificultad para utilizar una nueva aplicación, pasando por la preferencia de los clientes por uno u otro canal para cada tipo de transacción, entre otros muchos.

Estos son solo unos pocos ejemplos; las oportunidades son tantas como preguntas se puedan formular, considerando la proliferación de fuentes de datos en las entidades financieras y las crecientes capacidades en data science (talento y herramientas) de las que ya se están dotando.

En este sentido, la automatización de procesos y la mejora de los modelos empleados en el sector financiero están estrechamente ligadas a la capacidad de las entidades de capturar información relevante de sus clientes, procesos, productos, etc. y posteriormente, mediante herramientas de data science, proceder a su almacenamiento, procesamiento y explotación.

En la actualidad, las fuentes de información disponibles para ser explotadas por las entidades son prácticamente ilimitadas; esto pone de relieve que cualquier información, con independencia de su procedencia (interna o externa) y su carácter (estructurado o desestructurado), es potencialmente relevante para la toma de decisiones. Más aún, se estima que la mayoría de los datos masivos en los bancos provienen de las transacciones, los logs, los eventos y los emails, a distancia de las demás fuentes (Fig. 24).

Fig. 24. Fuentes de big data en banca y otros sectores.



Fuente: IBM & University of Oxford (2015).

Desafíos ante la adopción de data science

Como queda patente, las posibilidades que abre la disciplina de data science en el sector financiero son numerosas y tienen el potencial de mejorar sustancialmente las métricas de desempeño en todos los ámbitos: la experiencia del cliente, la eficiencia, el control del riesgo, la eficacia de las acciones comerciales, etc.

Sin embargo, la evolución hacia estas capacidades no es sencilla ni inmediata; las entidades se enfrentan a una serie de desafíos (Fig. 25), que han sido relacionados con los desafíos de big data en encuestas realizadas al sector (Fig. 26), entre los que cabe citar:

- Volumen y compartición de datos: las mismas cantidades masivas de datos que posibilitan la existencia de data science suponen un reto en términos de dimensionamiento de las bases de datos, arquitectura de almacenamiento, coste de la capacidad de procesamiento, tiempo de computación y necesidad de algoritmos optimizados, que se soluciona en parte con las nuevas herramientas y plataformas, pero que requiere una cuidadosa planificación tecnológica. Asimismo, la puesta en común de los datos entre las áreas de las entidades supone un reto tecnológico y organizativo complejo de resolver.

- Restricciones presupuestarias: ligado con lo anterior, la evolución hacia data science lleva aparejada una necesidad de inversión en infraestructura y en talento, que en el entorno de márgenes presionados y abundante regulación es necesario articular con las restricciones presupuestarias.

- Privacidad, ética y seguridad: la utilización de información abundante sobre los clientes levanta cuestiones sobre la privacidad de los datos y la ética de su uso. El servicio es tanto más personalizado cuanta más información se emplee, y, dado que la regulación todavía es incipiente sobre muchos de los matices de la privacidad, y es complicado que avance a la velocidad a la que se mueve el fenómeno big data, para encontrar el equilibrio entre privacidad y experiencia de cliente es necesario hacer intervenir a las áreas de asesoría jurídica y escuchar al propio cliente. En este punto también se engloban los aspectos de seguridad de la información y la garantía de que los datos recopilados de fuentes diversas están a salvo de hacking, robos de identidad y venta indebida a terceras partes.

- Captación y formación del talento: los data scientists constituyen todavía un perfil relativamente escaso, y la demanda de estos profesionales supera con creces la oferta. Por ello, uno de los principales desafíos es la captación de data scientists en el mercado y la formación de los metodólogos tradicionales para adquirir nuevas capacidades. Algunos estudios[73] apuntan a una escasez en

[73] McKinsey (2011).

Fig. 25. Desafíos en la evolución hacia data science en las entidades financieras.

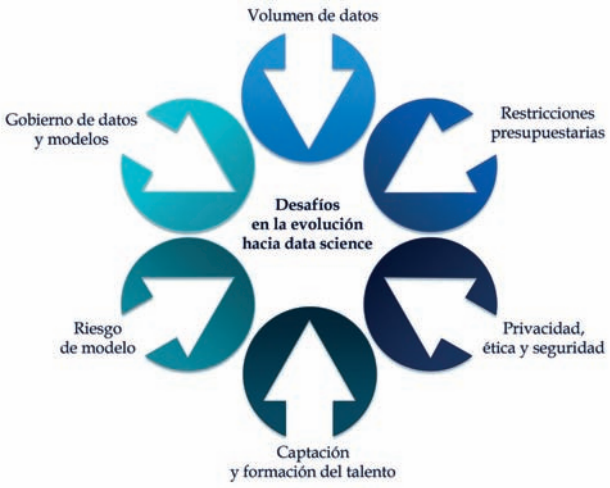
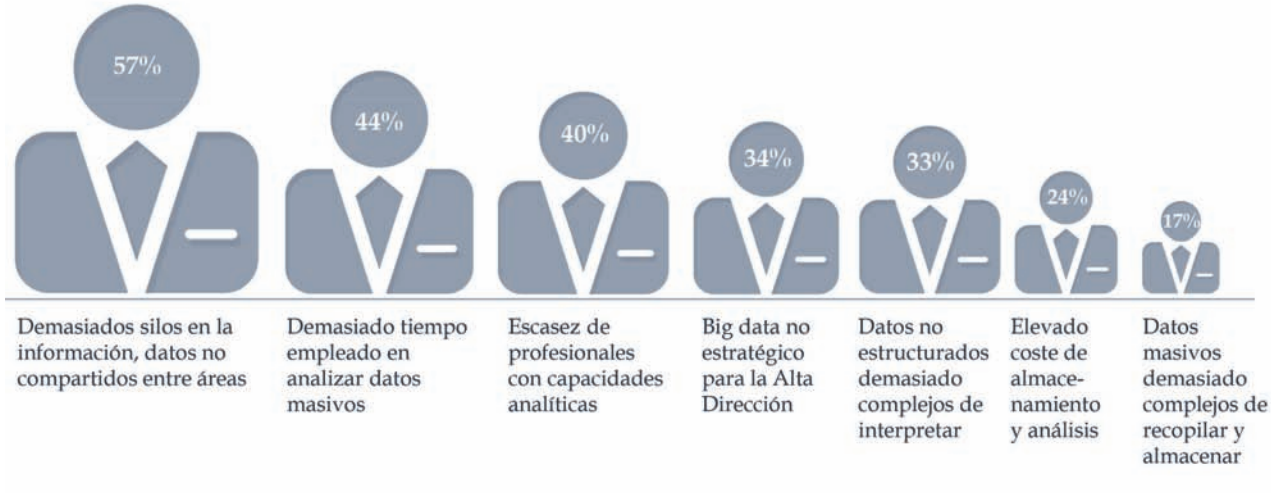


Fig. 26. Resultados de la encuesta sobre «mayores obstáculos en los bancos para el éxito en big data» (porcentajes sobre el total de respuestas).



Fuente: adaptado de Inetco (2015).

2018 de más de 140.000 data scientists solo en Estados Unidos.

- Riesgo de modelo: la utilización de modelos conlleva riesgos, que emanan de los datos que emplean, pero también de la propia estimación del modelo y de su potencial mal uso. La gestión y el control del riesgo de modelo son un foco de atención creciente ante la aparición de data science[74].

- Gobierno de los datos y de los modelos: por último, la organización y el gobierno de las estructuras internas necesarias para gestionar adecuadamente los datos y los modelos de una entidad financiera son un elemento clave para el éxito de la evolución hacia data science en la organización, como se detallará a continuación.

Impactos en el gobierno de datos y modelos

Aunque la aceleración en la generación y acceso a los datos, y las posibilidades que esto ofrece, constituyen una cierta novedad en el sector financiero, la realidad es que ni las entidades financieras ni los reguladores y supervisores están siendo ajenos al fenómeno descrito.

Por el contrario, las entidades han venido realizando transformaciones sobre sus sistemas y procesos de generación de información y de reporte, especialmente en los ámbitos de riesgos, financiero y comercial. Sin embargo, en muchos casos estas transformaciones se han realizado de forma poco estructurada y con una perspectiva limitada, como consecuencia de peticiones incrementales de los reguladores y supervisores, de necesidades de gestión no planificadas con visión global y, en muchos casos, de las migraciones de datos ocasionadas por fusiones y adquisiciones. Esto ha causado que los procesos de generación de información y reporte hayan ido perdiendo efectividad, y en ocasiones la consistencia de los datos no esté garantizada.

Más aún, los reguladores han señalado las carencias en los datos como una de las causas de la crisis financiera iniciada en 2007:

> *Una de las principales lecciones de la crisis financiera mundial iniciada en 2007 fue que la inadecuación de las tecnologías de la información (TI) y las arquitecturas de datos de los bancos impidió realizar una gestión integral de los riesgos financieros. […] En algunos bancos, la incapacidad para gestionar adecuadamente los riesgos respondía a carencias en la agregación de datos sobre riesgos y en las prácticas de presentación de los correspondientes informes. Esto tuvo consecuencias graves para los propios bancos y para la estabilidad del sistema financiero en su conjunto.*[75]

Por otra parte, la utilización de modelos para la toma de decisiones es un fenómeno que también está proliferando con gran rapidez, lo que aporta beneficios indudables, como la mejora de la objetividad, la automatización y la eficiencia. Sin embargo, su uso también conlleva un «riesgo de modelo», entendido como los potenciales perjuicios (económicos, reputacionales, etc.) provocados por decisiones basadas en modelos erróneos o utilizados de forma inapropiada[76].

Tanto la regulación como los aspectos de gestión derivados de la abundancia de información y su utilización en modelos para la toma de decisiones llaman a la necesidad de establecer un nuevo framework para gobernar de forma apropiada los datos y los modelos en cada entidad financiera. En esta sección se analizarán las prácticas en el sector en lo relativo a estos marcos de gobierno.

Gobierno de los datos

El establecimiento de mecanismos de gobierno de los datos es una necesidad estratégica compleja en las entidades, que se hace especialmente acuciante al hilo del fenómeno big data y es una condición imprescindible para obtener el máximo aprovechamiento de la información.

[74] Management Solutions (2014).
[75] BCBS (2013).
[76] Management Solutions (2014).

Este aspecto no ha pasado inadvertido a los supervisores, que, en algunos ámbitos, han emitido normativa específica relativa al gobierno de los datos e información, destacando especialmente, en lo relativo a riesgos, los *Principios para una eficaz agregación de datos sobre riesgos y presentación de informes de riesgos* (BCBS, 2013), conocidos como «RDA&RRF»[77], con requerimientos en materia de calidad, consistencia, integridad, trazabilidad y replicabilidad de los datos. Dicha normativa es vinculante para las entidades globalmente sistémicas y sus filiales, y en un futuro para las localmente sistémicas.

Asimismo, la importancia de estos aspectos se manifiesta en que ciertas entidades ya han articulado una involucración máxima del Consejo de Administración y de la Alta Dirección en los aspectos relativos a información y reporting, sustanciada en algunos casos en la creación de comités delegados del Consejo para la gestión de datos. Por otro lado, las entidades están abordando de forma generalizada la creación de figuras organizativas como el Chief Data Officer (CDO) o los responsables de Risk Management Information (RMI), y están abordando iniciativas estratégicas para robustecer la infraestructura soporte a los procesos de generación de información.

Los beneficios de un gobierno robusto de los datos son claros: consigue un reporte homogéneo y consistente, utilizando conceptos uniformes en toda la organización y alineando a la matriz y las filiales en el caso de grupos internacionales; garantiza la consistencia entre el reporting regulatorio y de gestión; consigue una mayor eficiencia (mayor automatización, menores redundancias), mejora el time-to-market y hace más flexible la generación de reporting; y, en el caso de riesgos, facilita el conocimiento preciso de los riesgos por la Alta Dirección y contribuye, en definitiva, a mejorar la gestión y el control de riesgos de la entidad.

Elementos de un marco de información y gobierno del dato

Lo anterior deriva, por tanto, en que las entidades más avanzadas en la materia disponen de un marco de gobierno de la información y del dato que detalla los principios básicos, los intervinientes y sus roles, la estructura de gobierno, y los elementos (procesos y herramientas) de soporte en relación con la gestión de los datos y la generación de información.

En lo relativo a los principios, el marco debe identificar las directrices básicas que rigen el gobierno de la información y los datos, incluyendo el alcance del marco y su ámbito de aplicación, la propiedad de los datos, la consistencia entre los diferentes ámbitos o los mecanismos desplegados para garantizar la calidad de la información.

En cuanto a la organización y gobierno, el marco identifica los intervinientes en el proceso y sus roles, incluyendo, en particular, los responsables de generación de la información, de garantizar su calidad y de los repositorios de información. Destacan figuras clave como el Chief Data Officer (CDO), responsable de asegurar la calidad y trazabilidad end-to-end del dato en los informes a la Alta Dirección y la consistencia de la información (para lo que se apoya en herramientas como el diccionario de datos), los responsables del dato (por ámbito) o, en el ámbito de Riesgos, el responsable de Risk Management Information (RMI), entre otros.

Asimismo, el marco contempla la definición de los órganos de gobierno responsables de la información y el dato, entre cuyas atribuciones se incluye el promover la elaboración e implantación efectiva del modelo del gobierno del dato, la revisión y aprobación de modificaciones relevantes en el proceso de generación de información, la aprobación de los

[77] Risk Data Aggregation and Risk Reporting Framework.

objetivos de calidad del dato y, en general, la definición de la estrategia de gestión de datos. En estos comités deben estar representados todos los usuarios de los datos, lo que suele incluir a las divisiones de Negocio, Riesgos y Finanzas, así como los responsables de Sistemas, el CDO y los responsables de la generación de información.

En los comités se pueden distinguir varios niveles, incluyendo comités técnicos responsables de dar soporte a los comités de mayor rango y resolver posibles conflictos sobre datos que afecten a varios ámbitos o geografías. Adicionalmente, en entidades internacionales es preciso garantizar la extensión del gobierno de los datos de manera consistente a todas las geografías, para lo que se precisa de la constitución de los comités pertinentes en los distintos países y el establecimiento de los adecuados mecanismos de reporting y escalado a los comités de ámbito corporativo.

El gobierno de los datos requiere de una serie de elementos para su correcta articulación, que faciliten el cumplimiento de los principios de calidad, trazabilidad, consistencia y granularidad requeridos por los reguladores[78], entre los que destacan:

- Diccionario de datos: se trata de un inventario unificado de métricas, dimensiones y componentes asociados a los informes, con definiciones funcionales claras y unificadas para toda la entidad.

- Metadatos: es la información específica sobre cada dato, contenida en el diccionario de datos, que permite su catalogación y condiciona su utilización. La tendencia es enriquecer los metadatos actuales (de negocio y técnicos), completándolos con metadatos adicionales sobre el origen del dato, su transformación y su calidad, que condicionan la decisión sobre el uso que se puede dar a cada dato.

- Datawarehouses y data lakes: son las bases de datos y otras fuentes de información que reúnan la calidad suficiente para construir las métricas candidatas a ser incluidas en los informes.

- Herramientas de explotación: son las herramientas analíticas de tratamiento y visualización de la información, entre las que desempeñan un rol esencial aquellas con capacidades de tratamiento de big data.

Finalmente, es crítico asegurar la calidad de los datos utilizados. Para ello las mejores prácticas consideran:

- El establecimiento de un modelo de control de los datos que incluya la monitorización de los procesos de generación de los informes para la Alta Dirección y la definición de los niveles de tolerancia de calidad que se deben aplicar a los datos que serán reportados.

- La identificación, definición e implantación de KPIs que permitan medir el grado de calidad de la información a múltiples niveles en el ciclo de vida del dato, así como la definición de herramientas (cuadros de mando) de agregación y seguimiento de los niveles de calidad de los datos en los informes a la Alta Dirección.

- La ejecución de planes de calidad de los datos a distintos niveles (sistemas operacionales, repositorios de información e informes), que se materializan en iniciativas de depuración de datos históricos (que se suele abordar mediantes planes de choque), y de mejora de la nueva producción de información (a través de modificaciones en los procesos).

[78]Por ejemplo, el Comité de Supervisión Bancaria de Basilea, la Fed y la OCC.

Retos en el gobierno de los datos

El desarrollo de un gobierno de los datos sólido conlleva, sin embargo, una serie de retos que las entidades deben afrontar, entre los que se cuentan:

- Garantizar la involucración de la Alta Dirección en el gobierno de la información, los datos y su calidad.
- Definir el perímetro de datos sujeto al modelo de gobierno (en especial considerando el crecimiento exponencial en cuanto a su diversidad y volumen), y asegurar que el modelo sea operativo y garantice los niveles adecuados de calidad, trazabilidad y consistencia de los datos sin implicar un menoscabo de las capacidades de la organización para optimizar su uso.
- Resolver los aspectos relativos a la privacidad y seguridad de la información y la garantía de que los datos están a salvo de usos fraudulentos.
- Reforzar la ciberseguridad, que incluye la protección contra el «hacktivismo» (los ataques contra las entidades por motivos ideológicos a través de virus, malware, etc.), el uso fraudulento de los datos, los ciberdelitos financieros, el espionaje y el robo de información. (Cabe mencionar que en 2014 se incorporó el riesgo de ciberataques al top 5 de riesgos globales del Foro Económico Mundial).
- Identificar e implantar herramientas que faciliten el gobierno de los datos y adaptar los mecanismos de gobierno del dato al caso de arquitecturas novedosas, como los data lakes.
- Implantar un diccionario único de conceptos, que permita una homogeneidad de entendimiento a lo largo de la entidad y, en su caso, las filiales.
- Involucrar a las diferentes filiales, en el caso de un grupo financiero, en el gobierno conjunto del dato.

Gobierno de los modelos

De acuerdo con la Fed y la OCC, el término «modelo» se refiere a «un método cuantitativo, sistema o estrategia que aplica teorías, técnicas e hipótesis estadísticas, económicas, financieras o matemáticas para procesar datos y obtener estimaciones cuantitativas»[79].

Hasta la fecha, existe poca normativa que regule de forma concreta el riesgo de modelo, y tiende a ser inespecífica tanto en su delimitación como en el tratamiento esperado. La excepción es la Supervisory Guidance on Model Risk Management publicada en 2011-12 por la OCC y la Fed estadounidenses.

En ella se define por primera vez el riesgo de modelo como «el conjunto de posibles consecuencias adversas derivadas de decisiones basadas en resultados e informes incorrectos de modelos, o de su uso inapropiado», y se establece, a través de unas directrices, la necesidad de que las entidades dispongan de un marco para identificarlo y gestionarlo, aprobado por sus consejos de administración.

Estas directrices cubren todas las fases del ciclo de vida de un modelo: desarrollo e implantación, uso, validación, gobierno, políticas, control y documentación por parte de todos los intervinientes. Entre los principales aspectos requeridos se encuentra la necesidad de tratar el riesgo de modelo con el mismo rigor que cualquier otro riesgo, con la particularidad de que no puede ser eliminado, solo mitigado a través de un cuestionamiento efectivo («effective challenge»).

[79] OCC/Fed (2011-12).

Elementos de un marco objetivo de MRM

Las entidades más avanzadas en esta materia disponen de un marco de gestión del riesgo de modelo (MRM) que se sustancia en un documento aprobado por el Consejo de Administración y que detalla aspectos relativos a organización y gobierno, gestión de modelos, etc.

En lo relativo a organización y gobierno, el marco de MRM se caracteriza por su transversalidad (involucra a varias áreas, como las líneas de Negocio, Riesgos, Auditoría Interna, Tecnología, Finanzas, etc.), la definición explícita de los tres roles que el regulador demanda (ownership, control y compliance[80]) y su asignación a funciones concretas de la organización y, sobre todo, el establecimiento de una función de Gestión de Riesgo de Modelo, cuya responsabilidad sea crear y mantener el marco de MRM.

En lo referente a la gestión de modelos, el marco de MRM incluye aspectos tales como: (a) el inventario de modelos, que cense todos los modelos de la entidad en todos sus ámbitos (riesgos, comercial, finanzas, etc.), soportado normalmente en una herramienta tecnológica apropiada que guarde traza de todos los cambios y versiones; (b) un sistema de clasificación o tiering de los modelos según el riesgo que comporten para la entidad, del que depende el nivel de exhaustividad en el seguimiento, la validación y la documentación de los modelos; (c) una documentación completa y detallada de cada modelo, que permita la réplica por parte de un tercero y el traspaso a un nuevo modelizador sin pérdida de conocimiento; y (d) un esquema de seguimiento de los modelos que permita detectar de forma temprana desviaciones del desempeño del modelo respecto a lo previsto, así como usos inadecuados, para tomar acciones en consecuencia.

La validación de los modelos es un elemento central para la gestión del riesgo de modelo, y debe tomar como principio fundamental el cuestionamiento (challenge) crítico, efectivo e independiente de todas las decisiones tomadas en el desarrollo, el seguimiento y el uso del modelo. La periodicidad y la intensidad de la validación de cada modelo deben estar proporcionadas a su riesgo, medido a través de su tier, y el proceso y el resultado de la validación deben documentarse exhaustivamente a su vez.

Retos en el gobierno de modelos

Todo lo anterior lleva a la necesidad de definir un gobierno de modelos robusto y estable, lo que plantea una serie de retos a las entidades financieras, entre los que cabe mencionar:

- La reflexión sobre qué es un modelo y qué modelos deben someterse a estos procedimientos (posiblemente dependiendo del tipo de modelo y su clasificación o tiering) y cómo compatibilizar esta necesidad de gobierno de los modelos con un mayor uso de ellos para múltiples fines.

- Resolver las dificultades planteadas por la existencia de mayores volúmenes y tipos de datos (no todos sometidos a los mismos controles de calidad) utilizados en el proceso de modelización.

- Lograr la involucración de la Alta Dirección en el gobierno de los modelos, y en concreto definir y aprobar el marco de riesgo de modelo al más alto nivel.

- Definir el esquema organizativo de la función (o funciones) de data science en términos de centralización o descentralización tanto geográfica como entre las áreas de la entidad, y delimitar las responsabilidades entre las áreas corporativas y locales, en el caso de grupos internacionales.

- Construir o reforzar los mecanismos de gobierno alrededor de cada uno de los procesos asociados a la función analítica.

En definitiva, gobernar los datos y su transformación en conocimiento, que implica a su vez gobernar los modelos que articulan esta transformación, ha pasado a ser un eje estratégico de actuación para cualquier organización, y en particular para las entidades financieras. En consecuencia, la tendencia indudable en los próximos años será el impulso decisivo de sus respectivos marcos de gobierno.

[80] El model owner define los requerimientos del modelo y suele ser su usuario final. Control incluye la medición del riesgo de modelo, el establecimiento de límites y el seguimiento, así como la validación independiente.
Compliance comprende los procesos que aseguren que los roles del model owner y de control se desempeñan de acuerdo a las políticas establecidas.

Caso de estudio: redes sociales y credit scoring

Es un error capital teorizar antes de tener datos.
Sin darse cuenta, uno empieza a deformar los hechos para que se ajusten a las teorías,
en lugar de ajustar las teorías a los hechos.

Sir Arthur Conan Doyle[81]

Objetivo

Con el propósito de ilustrar de forma directa la aplicación de la disciplina de data science en el sector financiero, se ha considerado de interés realizar un ejercicio cuantitativo que utilice algunas de las herramientas descritas para un uso específico en una entidad financiera.

En concreto, el objetivo del estudio es desarrollar un modelo de scoring crediticio para particulares utilizando datos extraídos de redes sociales, integrarlo con un modelo tradicional de préstamos personales, y comprobar en qué medida mejora el poder predictivo.

Datos del estudio

El estudio se ha llevado a cabo empleando los siguientes datos y modelos:

- Una muestra real de construcción de un modelo de scoring de préstamos de particulares, compuesta por aproximadamente 75.000 registros, con una tasa de incumplimiento en el entorno del 12%.
- Variables adicionales sobre los clientes de la muestra, que permiten su búsqueda en las redes sociales.
- Un modelo de scoring construido sobre la muestra anterior, que utiliza 12 variables y tiene un poder predictivo medio (ROC[82] del entorno del 73%).

Principales conclusiones

Las principales conclusiones que se desprenden del estudio son las siguientes:

- La cantidad y la calidad de la información disponible en las redes sociales son notablemente inferiores a las de los datos internos del banco: solo un 24% de los clientes tienen datos, y de estos, solo el 19% tienen información completa o casi completa.
- Además, la extracción de datos de redes sociales se caracteriza por un problema de desambiguación: las personas físicas no se identifican de manera inequívoca con un documento de identidad en la red, por lo que existe una probabilidad de error en la identificación de cada cliente con su perfil en las redes sociales. Para este estudio se han descartado los clientes en los que esta probabilidad se ha estimado superior al 25%.
- Las variables extraídas de las redes sociales, además, son en su mayoría cualitativas y pueden tomar una gran cantidad de valores, lo que dificulta su tratamiento, pero a cambio permite construir variables de una gran riqueza.
- El poder predictivo del modelo de scoring basado en redes sociales emplea 9 variables numéricas y categóricas (algunas discretizadas) que cubren varios aspectos del perfil profesional del cliente (en especial su historial laboral, pero también el sector, los estudios y los idiomas) y alcanza un poder predictivo equiparable al del modelo original, con una ROC del 72%.
- La combinación de ambos modelos, sin embargo, eleva sustancialmente el poder predictivo, hasta alcanzar un 79%.

[81] Sir Arthur Ignatius Conan Doyle (1859-1930). Escritor y médico escocés, célebre por su personaje Sherlock Holmes.
[82] Receiver operating characteristic, medida del poder predictivo de un modelo de respuesta binaria.

En síntesis, este estudio revela que la información de las redes sociales aporta una información sustancialmente distinta, que complementa y enriquece significativamente al modelo de scoring tradicional. No obstante, subsisten ciertas dificultades inherentes a su uso, que en el futuro previsiblemente serán resueltos en gran medida mediante la captura ordenada de los datos por parte de las entidades como parte de su proceso de admisión y seguimiento del crédito.

Descripción del estudio

El estudio se ha abordado en cuatro fases: extracción de datos de las redes sociales, limpieza y tratamiento, construcción del «módulo social» e integración con el «módulo tradicional».

Para la extracción de datos de las redes sociales, se ha utilizado una combinación de una herramienta específicamente diseñada en Python, que se conecta mediante APIs nativas de las propias redes sociales, con un módulo en VBA que extrae y ordena la información en un formato accesible.

La información de las redes sociales es muy irregular en su completitud y calidad, suele componerse de variables cualitativas y no se adapta a listados de valores preestablecidos. Además, no se extrae en un esquema relacional clásico, sino en registros que requieren un proceso de análisis o parsing para convertirlos en información utilizable.

Para este estudio se encontraron datos de aproximadamente 18.000 de los 75.000 clientes, y de ellos la información estaba razonablemente completa en cerca de 4.000. No obstante, se aplicó un tratamiento exhaustivo de los valores ausentes o missings, de modo que la cantidad de registros finalmente utilizable fue mayor.

Para esos clientes, se extrajeron 30 variables originales, con distintos niveles de falta de información. Las variables abarcaban varios ámbitos del perfil profesional y personal del cliente: su formación reglada y no reglada, su experiencia profesional, su ubicación geográfica y otros datos relativos a aficiones, intereses, etc.

El tratamiento de los datos comprendió, además de las transformaciones necesarias para manejar la información, la construcción de variables compuestas a partir de las originales. Así, a partir de las 30 variables extraídas se crearon más de 120 campos, en su mayoría categorizaciones basadas en análisis univariantes y bivariantes, y análisis temporales del historial profesional de cada cliente.

La fase de construcción del modelo es muy similar a la de cualquier modelo de scoring; se combinó el análisis experto con un proceso stepwise de selección de variables, se eliminaron las variables redundantes desde un punto de vista de negocio, y se aplicó un criterio de aceptación del 95% de confianza (p-valor inferior a 0,05).

El algoritmo utilizado fue una combinación de un árbol de decisión (reforzado con un algoritmo de poda para reducir la entropía, y por tanto mejorar la robustez y la estabilidad) y una regresión logística binomial:

$$P(Y=1)=\frac{1}{1+\exp(-\beta_0-\beta_1 x_1-\cdots-\beta_n x_n)}$$

Como resultado, se obtuvo un modelo de 9 variables, que se resume en la Fig. 27. La curva ROC, que mide su poder predictivo, se muestra en la Fig. 28.

Como se puede observar, se trata de un modelo con un poder predictivo medio, pero equiparable al del modelo tradicional, pese a emplear tan solo 9 variables; esto es atribuible en cierta medida a la presencia de variables cualitativas. Los demás estadísticos muestran que el modelo es robusto y tiene buenas propiedades estadísticas.

La última fase es la integración del módulo social con el módulo tradicional del scoring. Para ello, se ha construido una nueva regresión logística que toma como variables independientes las puntuaciones o scores de cada uno de los módulos:

$$P(Y=1)=\frac{1}{1+\exp(-\delta_0-\delta_1 score\_trad-\delta_2 score\_social)}$$

Ambos scores son significativos al 95% de confianza y, como se puede apreciar en la Fig. 29, el modelo final tiene un área bajo la curva ROC de 79%, lo que mejora sustancialmente el poder predictivo de cada modelo por separado.

El presente estudio se ha centrado en el caso de un modelo de scoring crediticio y la adición de datos provenientes de redes sociales, pero el ejercicio es extensible a otros tipos de modelos (valoración de garantías, fidelización, renta, abandono, propensión a la compra, etc.) y fuentes de información (logs internos, bases de datos públicas, información web, etc.).

En conclusión, como se ha demostrado, la incorporación de variables provenientes de otras fuentes tiene el potencial de incrementar significativamente la capacidad discriminante de los modelos tradicionales.

Fig. 27. Variables del módulo social del scoring crediticio.

Variable	Descripción	Peso relativo
Duración cargo actual	Duración en meses del cargo actual	18%
Antigüedad laboral	Antigüedad laboral en meses	15%
Mínima duración en cargo	Mínima duración en un cargo en su trayectoria profesional	15%
Máxima duración en cargo	Máxima duración en un cargo en su trayectoria profesional	13%
Sector de actividad	Categorización INE del sector profesional	12%
Número de trabajos	Número de trabajos actuales e históricos	9%
Tiempo sin estudiar	Tiempo transcurrido desde sus últimos estudios	7%
Idiomas	Número de idiomas que habla	7%
Ratio cargos/años	Número de cargos / número de años de trayectoria profesional	4%

Fig. 28. Curva ROC del módulo social del scoring crediticio.

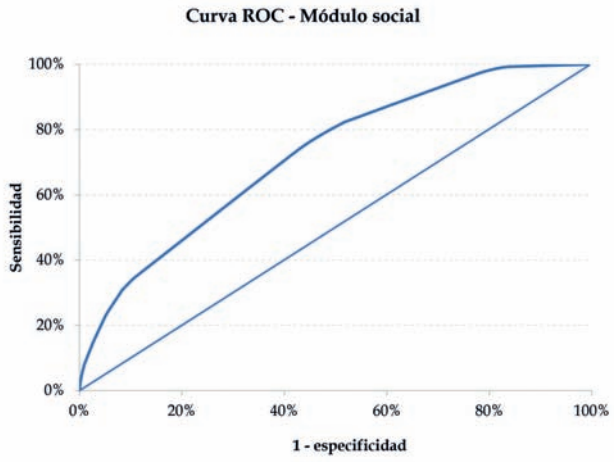
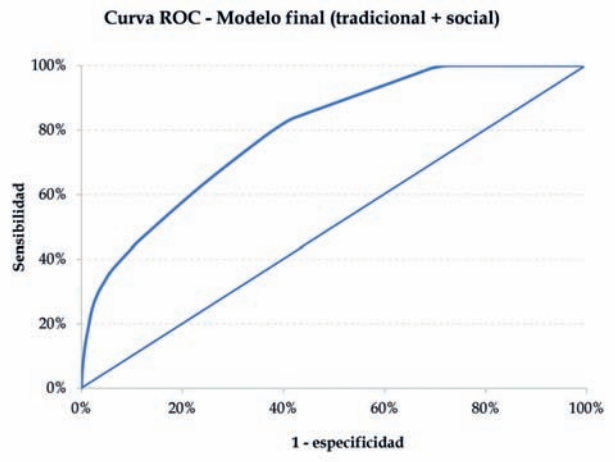


Fig. 29. Curva ROC del modelo final, que integra los módulos tradicional y social.



Bibliografía

Basel Committee on Banking Supervision 239 (2013). *Principios para una eficaz agregación de datos sobre riesgos y presentación de informes de riesgos.*

BBVA Research (2014). *Situación Latinoamérica. Cuarto trimestre de 2014.*

Berkeley (2015). http://datascience.berkeley.edu/about/what-is-data-science/

Bloomberg (2013). *HSBC Judge Approves $1.9B Drug-Money Laundering Accord.*

DeZyre (2014). *Hadoop in Financial Sector.*

Dhar, V. (2013). *Data Science and Prediction, Association for Computer Machinery.*

Digital Leadership GmbH (2014). *What banks will have to work on over the next couple of years.*

EFMA (2013). *World Banking Report 2013.*

European Banking Authority (2014). *Guidelines on common procedures and methodologies for the supervisory review and evaluation process (SREP).*

European Central Bank (2014). *Guía de supervisión bancaria.*

European Commission (2014). *EU Bank Recovery and Resolution Directive (BRRD): Frequently Asked Questions.*

Evans, P. (2014). *How data will transform business. TED.*

Federal Big Data Commission (2014). *Demistifying big data, a practical guide to transforming the business of Government.*

Federal Reserve (2014). *Consumer and Mobile Financial Services 2014.*

Fernald, J. (2014). P*roductivity and Potential Output before, during, and after the Great Recession. NBER Working Paper 20248,* National Bureau of Economic Research, Cambridge, Massachusetts.

Financial Conduct Authority (2015). fca.org.uk/firms/being-regulated/enforcement/fines

Gartner (2013). *Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020.*

Gigaom Research (2014). *Sector RoadMapTM: data discovery in 2014.*

Goldman, R. (2014). *Big Data, Risk Management, and Full-Fidelity Analytic.* Cloudera.

Gordon, R. (2014). *The Demise of U.S. Economic Growth: Restatement, Rebuttal, and Reflections.* NBER Working Paper 19895, National Bureau of Economic Research, Cambridge, Massachusetts.

Hall, R. (2014). *Quantifying the Lasting Harm to the U.S. Economy from the Financial Crisis.* NBER Working Paper 20183, National Bureau of Economic Research, Cambridge, Massachusetts.

Harvard Business Review (2012). *Data Scientist: The Sexiest Job of the 21st Century.*

Harvard (2014). *CS109 Data Science.*

KPCB (2014). I*nternet trends 2014.*

IBM (2014a). *Demystifying Big Data: Decoding The Big Data Commission Report.*

IBM (2014b). *What is a Data Scientist.*

IBM & University of Oxford (2015). *The real world of Big Data.*

Inetco (2015). *Driving Banking Engagement with Customer Analytics*.

International Monetary Fund (2014). *World Economic Outlook, oct. 2014*.

International Telecommunication Union (2014). *The World in 2014, facts and figures*.

Kurzweil, R. (2014). *The accelerating power of technology*.

Management Solutions (2012). *Riesgo de liquidez: marco normativo e impacto en la gestión*.

Management Solutions (2013). *Análisis de impacto de las pruebas de resistencia del Sistema financiero*.

Management Solutions (2014). *Model Risk Management: aspectos cuantitativos y cualitativos de la gestión del riesgo de modelo*.

McCallum (2014). *Disk Drive Prices (1955-2014),* jcmit.com.

McKinsey (2011). *Big data: The next frontier for innovation, competition and productivity*.

Moore, G. (1965). *Cramming more components onto integrated circuits*. Electronics Magazine. p. 4.

Office of the Comptroller of the Currency y Board of Governors of the Federal Reserve System (2011-12). *Supervisory Guidance on Model Risk Management*.

Office of the Comptroller of the Currency (2014). *OCC Guidelines Establishing Heightened Standards for Certain Large Insured National Banks, Insured Federal Savings Associations, and Insured Federal Branches; Integration of Regulations*.

O'Neil, C. y Schutt, R. (2013). *Doing Data Science*. O'Reilly.

Pearn, J. (2012). *What is Google's total computational capacity?*

Pethuru, R. (2014). *Handbook of Research on Cloud Infrastructures for Big Data Analytics*. IGI Global.

Pingdom: royal.pingdom.com

Portio Research (2013). *Portio Research Mobile Factbook 2013*.

SiliconAngle (2014). *When Will the World Reach 8 Zettabytes of Stored Data?*

Wired (2015). *White House Names DJ Patil as the First US Chief Data Scientist*.

World Bank, Sabbata, S. y Graham, M. (2013). *Internet Population 2011* – DeSabbata Graham OII.

Glosario

Bail-in: rescate de una institución con cargo a sus accionistas y acreedores.

Bail-out: rescate de una institución con cargo a fondos públicos.

Buffer de capital: recargo de capital, cuyo objetivo es garantizar que una entidad sea capaz de absorber las pérdidas derivadas de su actividad en periodos de estrés.

Comité de Supervisión Bancaria de Basilea (BCBS): organismo supranacional para la regulación prudencial de los bancos. Su objetivo es mejorar la calidad y promover la homogeneización de la supervisión del sistema financiero.

COREP (Common Reporting): marco normativo de reporting definido por la EBA que estandariza la presentación de informes de solvencia.

EBA (European Banking Authority): autoridad independiente de la Unión Europea, cuyo objetivo principal es mantener la estabilidad financiera dentro de la Unión y salvaguardar la integridad, eficiencia y funcionamiento del sector bancario. Se estableció el 1 de enero de 2011 como parte del Sistema Europeo para la Supervisión Financiera (ESFS) y absorbió al anterior Comité Europeo de Supervisores Bancarios (CEBS).

FATCA (Foreign Account Tax Compliance Act): ley federal que requiere a las entidades financieras de todo el mundo que reporten a la agencia fiscal de Estados Unidos las cuentas en el extranjero de las personas estadounidenses. Su objetivo es promover la transparencia fiscal.

Fed (Federal Reserve System): banco central de Estados Unidos, fundado en 1913 con el objetivo de proveer a la nación de un sistema monetario y financiero más seguro, flexible y estable. Con los años, su papel en el sector bancario y económico se ha expandido, incluyendo actividades como dirigir la política monetaria nacional, supervisar y regular las instituciones bancarias o proveer de servicios financieros a entidades depositarias.

Federal Big Data Commission: comisión federal cuyo objetivo es proporcionar asesoramiento al Gobierno de Estados Unidos sobre cómo utilizar los datos para incrementar su eficiencia y reducir sus costes.

Financial Stability Board (FSB): organismo supranacional que pretende incrementar la estabilidad del sistema financiero global a través de una mayor coordinación entre las autoridades financieras nacionales.

FINREP (Financial Reporting): marco normativo de reporte definido por la EBA que estandariza la presentación de los estados financieros.

IAS 39 e IFRS 9: normas relativas a la contabilidad de instrumentos financieros y que, entre otras medidas, requieren el cálculo de provisiones mediante modelos internos.

ICAAP (Internal Capital Adequacy Assessment Process): proceso interno de autoevaluación de la adecuación del capital en el sector bancario.

ILAAP (Internal Liquidity Adequacy Assessment Process): proceso interno de autoevaluación de la adecuación de la liquidez en el sector bancario.

Internet de las cosas: interconexión de los objetos de uso cotidiano a través de Internet. Según Gartner, en 2020 habrá en el mundo 26.000 millones de objetos conectados.

IRB (Internal Rating Based): método avanzado de estimación de capital regulatorio basado en modelos de rating internos. Para acceder a él, las entidades deben cumplir un conjunto de requisitos y obtener autorización del supervisor.

KYC (Know Your Customer): información relevante de clientes obtenida con diversos objetivos, como el cumplimiento regulatorio respecto a fraude, blanqueo de capitales, financiación del terrorismo o corrupción.

Mecanismo Único de Supervisión (SSM): mecanismo creado en 2014 que asume las competencias de supervisión de las entidades financieras europeas. Está formado por el Banco Central Europeo y las autoridades nacionales competentes de supervisión de los países de la zona euro. Sus principales objetivos son asegurar la solidez del sistema bancario europeo y aumentar la integración y la seguridad financieras en Europa. Realiza la supervisión directa de las 120 entidades más significativas y la indirecta de las aproximadamente 3.000 menos significativas.

Modelo de scoring crediticio (credit scoring): sistema de calificación automática del nivel de riesgo de un crédito. Se emplea, entre otros usos, para el cálculo de su probabilidad de incumplimiento y para decidir de forma automática sobre su concesión.

MREL (Minimum Requirement for Own Funds and Eligible Liabilities): requerimiento mínimo de fondos propios y pasivos elegibles para el bail-in.

NFC (Near Field Communication): tecnología inalámbrica que permite enviar y recibir datos a alta velocidad y corta distancia. Se emplea, entre otros usos, para realizar pagos con el teléfono móvil.

NLP (Natural Language Processing): procesamiento del lenguaje natural; estudio de las interacciones entre máquinas y el lenguaje humano a través del análisis de las construcciones sintácticas y el nivel léxico entre otros elementos.

OCC (Office of the Comptroller of the Currency): agencia federal estadounidense que se encarga de la regulación y supervisión de bancos nacionales, oficinas federales y agencias de bancos extranjeros. Tiene como objetivo principal garantizar que operen de forma segura y sólida, así como el cumplimiento regulatorio, incluyendo el tratamiento justo e imparcial de clientes y su acceso al mercado financiero.

PPNR (Pre-Provision Net Revenue): ingreso neto previo al ajuste de dotaciones de provisiones.

Ring-fencing: división financiera de los activos de una empresa hecha generalmente por motivos fiscales, normativos o de seguridad. En el sector financiero, alude a la separación jurídica entre las actividades mayoristas y la banca tradicional, como medida de protección de los depositantes.

Curva ROC (Receiver Operating Characteristic): curva empleada para analizar el poder predictivo de un modelo de salida binaria. Representa la relación entre el error de tipo 1 (clasificar incorrectamente sucesos adversos) y el error de tipo 2 (clasificar incorrectamente sucesos favorables).

Single Resolution Board: autoridad del mecanismo único de resolución, operativa desde el 1 de enero de 2015, encargada de la toma de medidas ante la inviabilidad de una entidad de crédito.

SREP (Supervisory Review and Evaluation Process): proceso de revisión y evaluación supervisora. Su objetivo es asegurar que las entidades financieras cuentan con los procesos, el capital y la liquidez adecuados para garantizar una gestión sólida de los riesgos y una adecuada cobertura de los mismos.

Stress test: técnica de simulación utilizada para determinar la resistencia de una entidad ante una situación financiera adversa. En un sentido más amplio, se refiere a cualquier técnica para evaluar la capacidad para soportar condiciones extremas, y es aplicable a entidades, carteras, modelos, etc.

TLAC (Total Loss Absorbing Capacity): requerimiento de capacidad total de absorción de pérdidas, cuyo objetivo es garantizar que las entidades de importancia sistémica internacional (G-SIBs) tengan la capacidad necesaria para asegurar que, en caso de resolución e inmediatamente después, las funciones críticas se mantengan sin poner en riesgo los fondos de los contribuyentes ni la estabilidad financiera.

Nuestro objetivo es superar las expectativas de nuestros clientes convirtiéndonos en socios de confianza

Management Solutions es una firma internacional de servicios de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización y procesos, tanto en sus componentes funcionales como en la implantación de sus tecnologías relacionadas.

Con un equipo multidisciplinar (funcionales, matemáticos, técnicos, etc.) de más de 1.400 profesionales, Management Solutions desarrolla su actividad a través de 18 oficinas (9 en Europa, 8 en América y 1 en Asia).

Para dar cobertura a las necesidades de sus clientes, Management Solutions tiene estructuradas sus prácticas por industrias (Entidades Financieras, Energía y Telecomunicaciones) y por líneas de actividad (FCRC, RBC, NT) que agrupan una amplia gama de competencias -Estrategia, Gestión Comercial y Marketing, Organización y Procesos, Gestión y Control de Riesgos, Información de Gestión y Financiera, y Tecnologías Aplicadas-.

En la industria financiera, Management Solutions presta servicios a todo tipo de sociedades -bancos, entidades aseguradoras, sociedades de inversión, financieras, etc.- tanto organizaciones globales como entidades locales y organismos públicos.

Luis Lamas
Socio de Management Solutions
luis.lamas@msspain.com

Javier Calvo
Director de I+D de Management Solutions
javier.calvo.martin@msspain.com

Marta Herrero
Supervisora de Management Solutions
marta.herrero.martin@msspain.com

Diseño y Maquetación
Dpto. Marketing y Comunicación
Management Solutions - España