Navegar en los nuevos paradigmas

Del big data a la inteligencia artificial

Reflexiones sobre Gobierno Corporativo

ETHALO
Ethics & Algorithms

Imágenes realizadas por el autor con IA (Firefly)

© Ethalo Advisory, SL
Avenida Isabel de Farnesio (1512), local J
28660 Boadilla del Monte
www.ethalo.es

Serie Reflexiones sobre Gobierno Corporativo
Edición revisada: noviembre 2024
(Sobre el capítulo El big data como paradigma de gestión del libro Piensa como un Consejero)

Se permite su distribución sin finalidades comerciales, bajo la condición de que el material se ofrezca de manera gratuita y se cite debidamente al autor como la fuente original. La distribución de este trabajo debe hacerse sin modificaciones y en su forma completa, salvo autorización expresa. Se anima a los lectores a compartir este documento, siempre que se respeten estas condiciones.
Se permite el uso de fragmentos o citas de este libro en cualquier medio, siempre y cuando se reconozca debidamente la autoría, indicando el nombre del autor y la fuente original del texto. Este permiso está limitado a usos que no tengan finalidad comercial y que busquen enriquecer otros trabajos a través de la referencia adecuada.
Este documento está protegido bajo las leyes de derechos de autor. Cualquier uso adicional de este trabajo debe ser autorizado por escrito por el autor.

Cita del presente documento:
Patxi Barceló. Navegar en los nuevos paradigmas (del big data a la inteligencia artificial). Ethalo Advisory – Serie Reflexiones sobre gobierno corporativo. Noviembre 2024 (2ª ed.)

Contenidos

Introducción 4
¿Qué hay detrás del concepto de *big data*? 4
¿Qué es lo que permite que los modelos de negocio, apoyados en el *big data*, den un salto cuantitativo? 6
Volumen 7
Velocidad 9
Variedad 10
El corazón del big data: la modelización 10
Visión *end to end* de los procesos *data driven*. 12
Disponibilidad y captura de datos 12
Procesos de control, limpieza y depuración de datos 13
Identificación de la pregunta objetivo 14
Preparación de datos (sistematización) 16
Construcción del modelo 17
Validación 19
Integración en el proceso 19
Seguimiento 20
El gobierno del modelo y el riesgo de modelo 21
Roles clave 21
Negocio 22
Chief data officer (CDO) 22
Big data managers 23
Aspectos éticos del big data 23
Una propuesta de puntos a valorar 26

Introducción

Este documento está pensado para ayudar a consejeros y directivos a introducirse en los aspectos clave de las nuevas tecnologías relacionadas con la inteligencia artificial cuando este ámbito puede no estar en sus competencias principales.

Sin embargo, en el título hemos querido empezar por el principio, es decir, por el *big data*, pues la inteligencia artificial es fundamentalmente una herramienta para aprovechar de forma más eficiente y rápida ingentes volúmenes de datos, haciendo inferencias útiles para la toma de decisiones.

Aunque la inteligencia artificial introduce nuevas formas de abordar el tratamiento de datos, queremos aquí abordar una visión más amplia y que está detrás de cualquier objetivo empresarial. Ser capaces de construir modelos -algoritmos-, que hagan más eficiente y eficaz el resultado de nuestras organizaciones.

Ahora bien, esos modelos pueden estar construidos con muchas metodologías y la inteligencia artificial ser una parte pequeña o grande del proceso o ni siquiera estar presente. Por tanto, veamos la foto con todas sus posibilidades.

Este documento es una versión revisada y actualizada del capítulo El big data como paradigma de gestión del libro Piensa como un Consejero[1]. Hemos creído que era el momento de revisar su contenido, a la luz de los rápidos avances en la materia, si bien sin alterar su estructura de partida, dado que hemos ido publicando otros documentos que complementan a este que hace de introducción a la materia.

¿Qué hay detrás del concepto de *big data*?

Big data es un término que se ha convertido en habitual en nuestras conversaciones. Sin embargo, no es menos cierto que la emergencia de la inteligencia artificial parece haber dejado el término obsoleto, casi innecesario.

Son conceptos que, por su aparente vida efímera en el entorno empresarial, quedan desprovistos del armazón conceptual de otros vocablos, o sobre los que se asume una visión general, sin pararse en su trascendencia[2].

Como concepto, tiene matices que hacen que su forma pueda parecer muy diferente según la perspectiva desde la que se mira. Por ello, en esta primera parte, en lugar de intentar dar una definición general, vamos a abordar su alcance a partir de la revisión de las perspectivas más relevantes.

Una de las diferencias clave está en el concepto de la gestión basada en datos (*data driven organization*). Esta terminología puede confundirnos. No trata de que tengamos ahora unos datos que antes no teníamos (que, adicionalmente, podría suceder).

En las escuelas de negocios y la práctica empresarial siempre se ha defendido que las decisiones deben estar basadas en hechos y datos y no en meras intuiciones. Todos tenemos bien interiorizado el concepto de que los objetivos deben ser necesariamente medibles y

---

[1] Patxi Barceló. El *big data* como paradigma de gestión. Piensa como un consejero. Editorial Punto Rojo (2022)

[2] Nos referimos, como se indica, al ámbito empresarial, donde la practicidad y las prioridades del negocio dejan poco tiempo para la reflexión. Esto provoca que conceptos como el big data queden definidos de forma vaga entre los equipos no especialistas.

cuantificables. Sin embargo, existe un cambio de paradigma en el concepto del *big data* que todavía pocas organizaciones han hecho propio y que supone un paso de gigante en los procesos de gestión.

Este salto es resultado, fundamentalmente, de la combinación del incremento exponencial de la capacidad de proceso computacional y de las técnicas específicas para realizar dicho proceso, donde los algoritmos asociados al concepto de inteligencia artificial juegan un papel relevante.

Además, este doble avance ha permitido empezar a usar datos no estructurados de muy diversa naturaleza (incluyendo imágenes, sonidos y otros tipos de datos basados en sensores diversos), cuyo coste de explotación antes era enorme.

Lo que subyace a este cambio está en la reducción significativa de la incertidumbre, al poder usar modelos estadísticos para la toma de decisiones. En este nuevo paradigma, la mayor parte de los comportamientos posibles pueden ser capturados y modelizados en base a un volumen de datos muy elevado y continuamente actualizado. La precisión estadística cambia en varios órdenes de magnitud.

Pensemos en alguna de las últimas decisiones que hayamos podido tomar. Intentemos asignarle una probabilidad de éxito. Seguramente seamos capaces de decir, en términos generales, algo así como "estoy seguro en un 80% de que es la decisión correcta".

En una organización basada en datos, la respuesta podría ser: para los clientes del conjunto A, el retorno será del 12% con un margen de error de 3 puntos; para los clientes del conjunto B el retorno será del 6% con un margen de error de 5 puntos y para los clientes del grupo C existe una acción diferente que permitiría un retorno del 8% con un margen de error de 2 puntos.

El uso de modelos basados en big data permite una precisión en los resultados que supone un salto en varios órdenes de magnitud.

Y es relevante entender lo que subyace a esa respuesta aparentemente tan precisa. Seamos conscientes que el big data no nos permite decir si el cliente 1 o el cliente 2 van a hacer exactamente lo que hemos previsto, sino que, si aplicamos la acción de forma consistente a todos los clientes del grupo A, de media, el resultado será el previsto.

Pero debemos advertir de algo muy relevante. A veces el concepto de individualización de la oferta a los clientes nos puede llevar a pensar que seremos capaces de acertar exactamente con cada cliente lo que necesita. Esto no es cierto, ni somos autómatas ni somos capaces de manejar todos los datos necesarios para llegar a ese nivel de granularidad (lo que incluso podría poner en duda el libre albedrio, de lo que nos ocuparemos en otros documentos). El *big data* y la modelización subsiguiente siempre funciona sobre la base de tratar grupos suficientemente homogéneos de clientes, de un tamaño estadísticamente significativo, para lograr un resultado determinado con un margen de error muy preciso.

Por tanto, son dos los elementos mínimos que deben definir el buen uso de los datos: la validez estadística del modelo empleado (lo cual se puede medir) y la consistencia plena en el uso del modelo decidido (lo que depende plenamente de los responsables del negocio[3]).

[3] A lo largo de este capítulo usaremos la terminología genérica de los negocios o los responsables del negocio para referirnos al usuario final de un modelo basado en big data. En primera instancia, pensaremos en los modelos que afectan a las decisiones con clientes, pero de la misma forma otras funciones (control de gestión, recursos humanos, finanzas) pueden usar modelos y, en ese sentido, ser los responsables del modelo, en cuyo caso esa función sería la que denominamos negocio.

Esta precisión debe entenderse bien: no busca la individualización de decisiones, sino un elevado grado de acierto en determinados grupos homogéneos. Siempre habrá excepciones con un margen de error conocido.

El primer elemento necesita ser plenamente entendido por los gestores para ser conscientes de las posibilidades y limitaciones de cada modelo. No nos engañemos, la estadística nunca ha sido una disciplina bien tratada en los planes de estudio y su comprensión está alejada de la mayor parte de las personas. Eso no quiere decir que sea compleja ni especialmente difícil de asimilar. Simplemente requiere una dedicación que habitualmente no ha tenido.

Implica ser capaces de hacer preguntas relevantes y entender las respuestas ¿Cuál es el poder predictivo de un modelo? ¿Cuál es el margen de error esperable con un determinado nivel de confianza? ¿Para qué población es válido el modelo? ¿Cuáles son los supuestos sobre los que está construido el modelo? ¿Qué peso tienen las variables incluidas en el modelo y qué variables no se han considerado y por qué? Aspectos que desarrollaremos en mayor medida a lo largo del capítulo.

El segundo elemento, la consistencia en el uso de los modelos, requiere de una cultura específica que no es fácil de integrar en las organizaciones (salvo aquellas que nacen específicamente bajo este paradigma).

Significa, en primer lugar, que en las discusiones en los comités no deberían escucharse frases como “eso no va a funcionar”, sino más bien frases como “¿es eso lo que queremos conseguir?”. El matiz es muy importante. La primera afirmación lo que está indicando es que el directivo cree que su intuición es mejor que lo que dicen los datos (el modelo) y que el modelo está equivocado.

Si el modelo está bien construido, la validez estadística del modelo es conocida, así como su margen de error. Por tanto, la segunda pregunta es central, pues no pone en duda el resultado esperado, sino que pone la cuestión en si el objetivo que se va a lograr al tomar la decisión es el que se quiere realmente.

Sobre la importancia del objetivo volveremos más adelante.

Conocer el margen de error pone encima de la mesa la pregunta clave ¿es adecuado ese posible error al automatizar el proceso? ¿es la solución que queremos adoptar?

Y tiene otro significado que muchas veces no se quiere ver. El modelo está construido bajo determinados supuestos y, en esas condiciones, funcionará. Pero si se usa en otro entorno, el modelo puede provocar un error significativo. Forzar el uso del modelo (es decir, no ser consistente en su uso) es uno de los errores más habituales en las organizaciones, pues se prima la eficiencia bajo una ilusión de validez universal del modelo.

¿Qué es lo que permite que los modelos de negocio, apoyados en el *big data*, den un salto cuantitativo?

Hay tres elementos en los que la literatura sí está plenamente de acuerdo y constituyen la base para alcanzar este nivel de valor añadido: el volumen, la velocidad y la variedad[4].

[4] Posteriormente diversos autores han ido añadiendo otras características como la veracidad, la viabilidad, la visualización o el valor (hasta llegar a las 7v del big data). No obstante, creemos que estos elementos no son exclusivos del big data, aunque si necesarios.

Estas características no nacen de un cambio fundamental de concepto sobre el tratamiento tradicional de los datos, sino de las posibilidades que se han abierto con las mejoras de capacidad, velocidad y potencia de los sistemas informáticos, junto con la ampliación del universo de sensores de captación de datos.

Al mismo tiempo, no es menos cierto que se han desarrollado nuevas funcionalidades y herramientas que son capaces de maximizar el aprovechamiento de este crecimiento de los datos, tratando de crear estructuras que soporten los datos con menos restricciones y mayor flexibilidad. Aquí es donde aparecen conceptos como el de *data lake*, donde los datos fluyen, como innovación frente a las bases de datos estructuradas; lenguajes con características específicas para el tratamiento de datos de forma masiva, como Python; o ecosistemas como Hadoop para integrar los diferentes elementos y, por supuesto, la nube (*cloud*) como soporte de infraestructura flexible y escalable en un entorno donde el volumen de datos crece exponencialmente.

Volumen

Respecto a las tres características del fenómeno del big data, y empezando por el volumen, hoy en día la cantidad de datos que somos capaces de manejar ha crecido exponencialmente, fundamentalmente por tres razones:

La capacidad de aprovechar datos de fuentes antes no consideradas. Por ejemplo, los datos biométricos, los datos de voz, los datos de imágenes y videos, los datos de sensores que, aun existiendo desde hace tiempo, normalmente no estaban interconectados con otros sistemas y no generaban un flujo de datos continuo (pensemos, por ejemplo, en el flujo de datos de un scanner cerebral o de un sismógrafo).

También parece necesario analizar las posibles fuentes de datos externos, la cooperación con terceros y las fuentes de datos abiertas (*open data*).

Un ejemplo muy específico pero clarificador del aprovechamiento de estas nuevas fuentes de datos es el de los movimientos del ratón que hacemos al usar el ordenador[5]. Al capturar estos movimientos, se pueden crear sistemas de protección contra el fraude, al verificar que, al rellenar un formulario por internet, el autor es una persona y no un sistema informático preparado para rellenar miles de formularios de forma automática.

Este tipo de mecanismos son fundamentales, por ejemplo, para evitar el falso conteo en campañas de marketing digital en las que se satisface un importe por cada "*click*" conseguido, diferenciando los realizados por humanos frente a *bots*.

Esos mismos movimientos, junto con las cookies y los *clicks* realizados en una página, pueden ayudar significativamente a identificar aspectos relevantes de los intereses de una persona. ¿Cuántas veces usamos el ratón para señalar a otra algo que nos interesa en la pantalla?

En este mismo sentido, el volumen de imágenes almacenadas y las posibilidades de su tratamiento elevan exponencialmente la información disponible. Hoy, gran parte del planeta está fotografiado con un elevado nivel de detalle y perfectamente geolocalizado (uno de los

[5] Como referencia de los diversos trabajos sobre la correlación de los movimientos del ojo y el ratón, reseñamos el artículo de Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed y Alex Smola. Measurement and Modeling of Eye-mouse Behavior in the Presence of Nonlinear Page Layouts. (#60)

ejemplos más evidentes son las imágenes de Street View de Google, con fotografías de todas las calles y carreteras de gran parte de países).

A la hora de tratar de cuantificar el volumen de datos que hoy se podrían capturar, veamos algunos números significativos[6]: En 2019 teníamos en el planeta 4.100 millones de usuarios de internet, con un tiempo de conexión promedio de 6 horas y 42 minutos, sin olvidar que el 97% de la población está cubierta por una red móvil, al menos 2G. Hay 3.000 millones de usuarios de teléfonos inteligentes (es decir, capaces de capturar todo tipo de datos del usuario). En varias ciudades chinas el número de cámaras de vigilancia supera las 100 por cada 1.000 habitantes (más de una cámara por cada 10 personas); pero si nos vamos a Londres, ese número es también de 68 cámaras por cada 1.000 habitantes. Si todo esto lo convertimos en posibles datos, podemos entender la evolución exponencial de los datos disponibles incluso antes de entrar en la era del internet de las cosas (IoT).

Aplicaciones como Tinder usan elementos de geolocalización para identificar a los usuarios y el tiempo que pasan juntos, así como los lugares que frecuentan, lo que, unido a otros datos económicos o del perfil educativo, le permiten afinar su algoritmo. Debemos tener en cuenta que, en 2019, en Estados Unidos, más del 60% de las parejas homosexuales y casi el 40% de las heterosexuales afirman haberse conocido on line[7].

Preguntas para reflexionar:

- ¿El foco está puesto en capturar datos o en aprovecharlos?
- ¿Se usan fuentes externos para enriquecer, contrastar o complementar los modelos internos?
- ¿Existen las capacidades internas para aprovechar fuentes de datos no estructuradas?

Sólo esta aplicación de *dating* genera más de un millón de citas semanales. Una de sus funciones más populares es el proceso de *swiping* por el que sus usuarios van viendo fotos de posibles candidatos y los deslizan a izquierda (cruz) o derecha (*love*), según sus preferencias. La aplicación mide, además, el tiempo empleado en cada selección, permitiendo así generar un perfil de preferencias muy específico para cada usuario.

Como elementos adicionales de esta revolución, además de crecer la capacidad de captura de datos, se han incrementado las **capacidades de almacenamiento de esos datos y de su tratamiento masivo,** lo que permite hoy aprovechar volúmenes de datos que anteriormente no eran imaginables.

La estimación que hace IDC (International Data Corporation)[8] es que en 2020 se han producido 64 zettabytes[9] de datos, si bien sólo el 2% se han almacenado realmente (el resto es información usada, procesada y, posteriormente, desechada).

---

[6] El Atlas de la revolución digital. Editorial Clave Intelectual (2020)

[7] Michael J. Rosenfeld & Reuben J. Thomas & Sonia Hausen, 2019. "Disintermediating your friends: How online dating in the United States displaces other ways of meeting," Proceedings of the National Academy of Sciences, vol. 116(36), pages 17753-17758, September. (#841)

[8] Worldwide Global StorageSphere Forecast, 2020–2024: Continuing to Store More in the Core (IDC #US46224920) - https://www.idc.com/getdoc.jsp?containerId=US46224920. En informes posteriores se estima que en 2025 se producirán 175 zettabytes de información, con una tasa de crecimiento anual compuesta del 23%

[9] Un zettabyte son 1.000 millones de terabytes. Un terabyte es el tamaño de un disco relativamente grande de almacenamiento actual de un PC, donde se podrían almacenar algo más de 6,5 millones de páginas de documentos de texto. Por tanto, un zettabyte equivale a que cada habitante de la tierra escribiera un texto de casi un millón de páginas (o 2.000 libros de 500 páginas).

Actualmente, la capacidad de almacenamiento existente apenas llega al 10% de la información generada y no es suficiente para todos los datos que se producen[10]. Esta diferencia se seguirá incrementando con el crecimiento de los datos procedentes del IoT (*Internet of Things* o Internet de las Cosas), a pesar de que la capacidad de almacenamiento se duplica aproximadamente cada 4 años y se estima que llegará casi a los 9 zettabytes en 2024.

En 2020 aproximadamente el 65% de la información se ubica en discos instalados en PC, lo que implica una elevada fragmentación del acceso a gran parte de esta. Sin embargo, se estima que en 2025 casi el 60% de la capacidad de almacenamiento estará en la nube.

Dado que el volumen de datos no para de crecer, uno de los aspectos más relevantes es la arquitectura para almacenar esos datos, de forma que estos puedan estar ubicados en múltiples espacios distribuidos y con un modelo de interconexión altamente eficiente entre los diferentes conjuntos de almacenamiento.

Este es uno de los aspectos críticos a valorar en el desarrollo de la infraestructura de datos de una empresa. Los servicios en la nube son escalables y más eficientes en términos de coste, pero comportan riesgos de accesibilidad que deben ser adecuadamente previstos y disponer de planes de contingencia al respecto. Para las pequeñas y medianas empresas, la nube permite acceder a servicios antes inimaginables a bajo coste, por lo que es imprescindible que sean capaces de acercase a esta realidad.

Preguntas para reflexionar:

- ¿Existe una estrategia sobre datos en la organización?
- ¿La organización tiene una infraestructura de datos común, compartida y accesible?
- ¿Existe un área responsable de facilitar el acceso a los datos, su identificación correcta, ayudar en su interpretación y utilización adecuada?

Velocidad

La velocidad de datos surge como una combinación del almacenamiento de datos y su propias limitaciones, de la conexión permanente de dispositivos en tiempo real y la capacidad y velocidad de procesamiento de toda esa información.

Uno de los grandes cambios de las herramientas de *big data* es la posibilidad de usar flujos de datos en movimiento, no necesariamente integrados en bases de datos tradicionales.

Esto implica el uso de datos según se producen y sin necesidad de almacenamiento a largo plazo. El dato es integrado en el proceso en tiempo real, en el momento en que se produce. Incluso en los procesos más sofisticados, con la capacidad de aprender de dicho flujo de datos para adaptar progresivamente la respuesta del sistema.

Lo vivimos diariamente, por ejemplo, cuando mientras leemos la prensa en internet nos aparecen anuncios relacionados con las últimas búsquedas que hemos realizado de determinados productos o cuando nos llegan ofertas al móvil de un establecimiento que está cercano de donde nos encontramos en ese momento.

---

[10] Otra forma de tratar de dimensionar estos volúmenes de información es decir que grabar en alta definición (HD) la historia completa de la Tierra desde su formación ocuparía aproximadamente 125 Zettabytes (https://computerhoy.com/noticias/internet/que-demonios-es-zettabyte-quetienes-que-saberlo-63492)

Son datos que no tienen que ser almacenados necesariamente, pero que fluyen en tiempo real, alimentando procesos que adaptan la información que nos llega a lo que estamos haciendo, dónde estamos ubicados o incluso con quién estamos en ese momento.

Debe tenerse en cuenta que, si se toman decisiones mediante ese flujo de datos en tiempo real, hay que asegurar la captura de "fotos instantáneas" de esos datos para asegurar la trazabilidad de la decisión y la propia supervisión del modelo. Además, los algoritmos basados en este tipo de datos deben estar preparados para detectar posibles incoherencias en el flujo de datos y establecer medidas preventivas.

Variedad

En apartados anteriores, especialmente al hablar del volumen de información, ya hemos visto la diversidad de acciones (incluso no acciones) que pueden transformarse en datos. Imágenes, movimientos del ratón, tiempos de respuesta o la forma de llegar a un determinado artículo de prensa y el tiempo de lectura de este como elementos que configuran parte de nuestra huella digital.

Otro ejemplo es el uso de las grabaciones de sonar de los barcos para, mediante procesos de aprendizaje, identificar con precisión los posibles objetos que se hallan en el fondo del mar, logrando una precisión imposible de alcanzar por los mejores expertos humanos. Esto permite distinguir con claridad una mina submarina de otros muchos posibles objetos[11].

Los sonidos pueden generar otra información sensible más poderosa, como puede ser el estado de ánimo de una persona o su nivel de nerviosismo. La combinación del lenguaje empleado, la tonalidad, el tiempo, el volumen y otros atributos puede llegar a usarse para tener ventaja en procesos de negociación telefónica.

Como siempre, no debemos confundir la elevada certeza estadística que puede lograrse mediante estas técnicas frente a la individualización de cada caso.

El corazón del big data: la modelización

Hemos ido viendo que entre las características del big data está el uso de datos en grandes volúmenes, con todo tipo de datos (variedad) y hasta usando datos en tiempo real (velocidad) e iniciamos este documento indicando que el concepto de big data no era fácil de definir y que nos aproximaríamos a través de sus atributos.

Ahora tenemos que añadir que el propio nombre genera confusión, pues parece referirse fundamentalmente a estos atributos (datos en grandes cantidades) y no menciona la razón de ser de estos, que no es otra que la capacidad de construir modelos de alta precisión sobre el comportamiento general de las cosas o las personas, es decir, ser capaces de clasificar, predecir e incluso promover comportamientos[12].

---

[11] Estos modelos permiten incluso detectar el tipo de peces de los bancos que se identifican mediante el sonar. Véase, por ejemplo, https://www.azti.es/nuevo-modelo-de-inteligencia-artificial-identificar-bancos-de-peces-pelagicos/

[12] Hablamos de comportamientos y, en general, usaremos ejemplos relacionados con aspectos del comportamiento humano, pero de la misma manera pude ser aplicado a objetos y su comportamiento, por ejemplo, modelos para clasificar la calidad de piezas (modelos descriptivos), predecir la rotura por uso de esa pieza (modelo predictivo) o incluso generar acciones de mantenimiento preventivo para evitar la rotura (modelo prescriptivo), sobre lo que se profundiza a continuación.

Los datos en sí mismos no tienen más valor que una enciclopedia. El objetivo es poder sacar conclusiones de los datos. Y estas conclusiones se centran, de forma simplificada, en tres aspectos:

• Agrupar objetos o personas en conjuntos suficientemente homogéneos y describir su comportamiento (modelos descriptivos).

• Predecir el resultado o comportamiento de cada grupo ante diferentes eventos (modelos predictivos).

• Identificar acciones que permitirán lograr un determinado comportamiento de cada uno de esos grupos (modelos prescriptivos).

Por tanto, el proceso de desarrollar modelos sobre los datos (lo que denominaremos modelización) es el corazón del sistema.

Los datos en sí mismos no tienen más valor que una enciclopedia. El objetivo es poder sacar conclusiones de los datos, **mediante modelos** que ayudan a clasificar, predecir resultados o identificar acciones que faciliten cumplir determinados objetivos.

Es la capacidad de extraer el valor de los datos y ponerlos al servicio de un objetivo concreto[13]. Sin este proceso fundamental, estaríamos ante un posible diagnóstico de Diógenes de los datos, acumular datos sin obtener valor a cambio, a modo de basura moderna.

Es aquí donde vuelve con intensidad el cambio de paradigma en la gestión que supone el *big data*, pues implica un proceso de desarrollo de modelos sobre bases estadísticas sólidas para alcanzar niveles de precisión antes inalcanzables en términos de clasificación, predicción y promoción de comportamientos, con lo que esto implica en términos tanto económicos como de control social.

Quizás en este punto podamos pensar que la modelización es un estado del arte que no hemos alcanzado en la mayor parte de las organizaciones. Pero toda empresa toma decisiones sobre los precios que aplica, los descuentos que ofrece, las condiciones de pago, las compras que necesita hacer, los salarios de los empleados, el nivel de mano de obra necesaria, el coste financiero asumible para los pasivos o los proveedores de determinados servicios.

La cuestión relevante es si estamos obviando que todas esas decisiones, como prácticamente cualquier otra, las asumimos en base a modelos. En ocasiones, el modelo es algo tan poco formalizado como el que tiene en la cabeza un directivo y está basado en su experiencia. En otros casos es una hoja de cálculo sencilla. En otros, podemos estar hablando de un modelo estadístico de valoración del nivel de riesgo a asumir con un cliente o del precio que le ofrecemos en base a un sofisticado algoritmo.

Toda empresa utiliza modelos para tomar decisiones. La cuestión es valorar la posibilidad de lograr que esos modelos multipliquen significativamente la capacidad de la empresa de tomar decisiones acertadas.

Es aquí donde reiteramos que nuestro nivel de seguridad y, en consecuencia, de acierto en la toma de decisiones se verá fuertemente impactado por el tipo de modelo que usamos. El

[13] La Data Science Association en su propuesta de código de conducta define Data Science como el estudio científico de la creación, manipulación y transformación de datos para dotarlos de sentido. Es esta finalidad la que aporta la razón de ser del big data; los datos son la materia bruta.

aprovechamiento del *big data* y el nivel de exigencia que nos impongamos sobre los diferentes elementos que componen este viaje serán el combustible que nos permitirá llegar más o menos lejos.

Visión *end to end* de los procesos *data driven*.

Hemos comentado previamente que los datos son la materia prima para un objetivo: la predicción y promoción de comportamientos.

Esto lo logramos a través de la construcción de modelos de base estadística que permiten tomar decisiones de forma automatizada y que se integran en los procesos del negocio.

Por tanto, hablaremos de procesos guiados por los datos (*data driven processes*) para referirnos a aquellos procesos de la organización que tienen incorporado un modelo de decisión basado en el análisis de datos, normalmente con métodos estadísticos[14].

Hablamos de modelo de decisión en cuanto que las decisiones son tomadas, ya sea de forma automática o con intervención humana, usando los resultados del modelo. Incluso puede existir un subproceso que permita tomar una decisión concreta en discrepancia con el modelo, pero este hecho está a su vez regulado en cuanto a la forma y el número de veces en que es posible hacerlo.

Desde esta perspectiva, cuando hablamos de la visión *end to end* nos referimos a todos los elementos que permiten construir, utilizar y controlar un proceso guiado por los datos (*data driven process*) dentro de una organización.

Una organización con una fuerte cultura del dato promueve el desarrollo de modelos para la mejora de la toma de decisiones.

Finalmente, podemos hablar de una organización guiada por los datos (*data driven organization*) cuando la cultura organizativa promueve la adaptación de los procesos de decisión de la empresa a procesos *data driven* y estos cumplen con los requisitos de construcción, uso y control adecuados.

Por ello, en este apartado, desarrollaremos cada uno de estos elementos y sus requisitos de alto nivel.

El proceso no debe ser entendido como una secuencia lineal. Si bien en su primera iteración posiblemente deba seguir un orden natural, es un proceso que puede ser objeto de retroalimentación en cualquier punto y, en consecuencia, iterar en diferentes pasos según sea necesario.

Además, el proceso debe ser entendido como una estructura, de forma que cambios en algunos de sus componentes requieren revisar el conjunto mediante el ajuste de otros elementos de la propia estructura hasta llegar a un punto de equilibrio.

Disponibilidad y captura de datos

La captura de datos y su puesta a disposición es un proceso permanente que no debe estar limitado a los datos necesarios para dar respuesta a las preguntas del negocio. Hoy la captura y almacenamiento son relativamente baratos y el *data mining* (analítica de datos)

[14] Existen modelos desarrollados sobre otros procesos no necesariamente de naturaleza estadística, pero para los objetivos de este capítulo, no es relevante la distinción.

puede obtener información relevante de su análisis, independientemente de que exista una finalidad específica.

No obstante, existen algunas cuestiones sobre las que es necesario profundizar. La primera de ellas es la cultura de silos, esto es, la captura de datos de forma descentralizada e incoherente que suele ser habitual en las organizaciones que no nacen desde una cultura del dato. Los *legacy* (herencias) tecnológicas y de procesos, la propia estructura interna y los miedos a perder el control de los datos son obstáculos serios para lograr una gestión eficiente.

Lo habitual en las organizaciones con cierta historia es la existencia de diversas estructuras de datos, en muchas ocasiones independientes, que se alimentan de sistemas diversos según la aplicación que dio origen al proceso. Esto provoca incluso que el mismo dato esté recogido en varios repositorios, desde fuentes diferentes, con las correspondientes incoherencias internas[15].

Esto viene agravado por el hecho de que, al aplicar estrategias de centralización, estas suelen tratar de abordar el problema en su conjunto: fuentes de datos, repositorios y su explotación, lo que suele generar muchas reticencias en los usuarios, dado que no sólo pierden la propiedad del dato y su repositorio, sino que además ven reducida su capacidad de explotación de los datos.

No abogamos aquí por un modelo u otro, solo señalamos los riesgos y motivos de fracaso habituales en estos planes. Sí queremos destacar que, la existencia de una disponibilidad plena de los datos para todos los usuarios, con accesos y herramientas adaptados a sus necesidades (datos brutos, datos elaborados o directamente datos de síntesis) es una clara ayuda.

El otro catalizador de esta evolución es la existencia de un equipo volcado en facilitar el conocimiento del dato. Suele hablarse del diccionario de datos como la herramienta clave para ello. Como en otras cuestiones, no sólo es importante el instrumento, sino el equipo que lo administra, dado que su función debe ser entender las necesidades de los negocios y ayudarles a encontrar el dato adecuado, al tiempo que promueven la captura amplia y ordenada de datos para maximizar su disponibilidad.

> Los modelos parten del uso de datos. La democratización ordenada del dato es una actividad clave para maximizar su aprovechamiento.

Procesos de control, limpieza y depuración de datos

Una cadena es tan frágil como el más débil de sus eslabones. Los datos son, habitualmente, el eslabón más débil de los procesos *data driven*. No por su carencia, sino por su calidad y consistencia. Enumeramos aquí una serie de casuísticas habituales que remarcan la importancia de mantener procesos continuos de control, limpieza y depuración de datos.

Los datos se acumulan y acaban teniendo un tamaño inabordable. Los pequeños errores acaban por convertirse en enormes monstruos. No abordar los problemas de control y

[15] En muchas organizaciones se reproduce lo que es un hecho en internet: la duplicidad de información en múltiples repositorios. Hoy se estima que en internet el 90% de la información es copia de una información base que no alcanza más que el 10% de los contenidos totales (Fuente: https://firstsiteguide.com/big-data-stats/). Las empresas más grandes deberían abordar esta cuestión mediante algoritmos de identificación de ficheros y datos duplicados y procesos de alerta y acción temprana.

limpieza en origen acaban generando costes desproporcionados, de forma que o bien se renuncia al dato o, si este es imprescindible, en ocasiones incluso por cuestiones regulatorias, supone un elevado gasto sobrevenido.

Un mismo dato puede cambiar su significado a lo largo del tiempo. Rara vez se tienen en cuenta estos cambios de criterio al interpretar los datos y es una fuente de importantes errores.

Los criterios cambian con el tiempo. Los datos, en ocasiones, tienen procesos de captura manual, sujetos a la interpretación de la persona. Otras veces se dan instrucciones precisas que cambian con el tiempo. En otras, se cambia el instrumento de medida o el algoritmo de captura. En resumen, la comparación de los datos en el tiempo no siempre es coherente, por lo que es necesario mantener un registro claro de estos cambios que permitan al analista de datos interpretar correctamente el dato y homogeneizarlo.

Los flujos en tiempo real no siempre funcionan. Cuando queremos tener decisiones *on line*, hay que tener en cuenta que no todas las fuentes de entrada tienen el mismo nivel de estabilidad, por lo que hay que establecer alternativas para los posibles errores (datos faltantes o fuera de rango o que llegan con la latencia inadecuada), de forma que el modelo no tome decisiones equivocadas por ese motivo.

La interpretación del dato no siempre es evidente. El analista de datos debe conocer con precisión el significado del dato. Las denominaciones no siempre son claras y, en ocasiones, el conocimiento no es directo. Un buen diccionario de datos ayuda en estos casos, pero son raras las organizaciones que lo tienen actualizado y de acceso sencillo.

Existen datos que no son datos. Parece absurdo, pero muchos procesos acaban poniendo un cero, una fecha del 1 de enero de 1990 o incluso una media cuando el dato no llega de la fuente origen. Esos datos no son datos[16], pero esta casuística no siempre se identifica adecuadamente.

No pretendemos hacer una enumeración exhaustiva de casos, sino enfatizar la importancia de la sistemática en la vigilancia y limpieza de los datos para poder llegar a buen puerto.

Identificación de la pregunta objetivo

Este elemento es obvio, pero a la vez requiere especial atención. Un modelo trata de determinar la respuesta más adecuada a una pregunta concreta en cada caso que se le presenta.

No seamos ilusos, es necesario saber lo que buscamos; el disponer de muchos datos (*big data*) y hacer análisis exhaustivos de los datos (*data analytics*) no nos una solución sin conocer el problema que queremos resolver[17]. Por tanto, el modelo nace para dar respuesta a una pregunta. De una forma muy simplificada, podemos decir que un modelo se construye al revés. Se evalúan las posibles respuestas a la pregunta y las características de los datos

[16] Se pueden convertir en datos cuando esas decisiones de sustitución están debidamente razonadas y documentadas y el proceso de construcción del modelo lo tiene en cuenta para hacer los ajustes de parametrización convenientes en cada caso.

[17] El análisis de datos en bruto nos puede proporcionar intuiciones (*insights*) de cuestiones que se nos podían estar pasando por alto; relaciones relevantes en nuestro modelo de negocio. Pero estos resultados son metadatos (datos que nacen de la explicación de los datos), todavía no resuelven un problema concreto.

de que disponemos que permiten identificar en cada caso como se relacionan esas características con las posibles respuestas[18].

La utilidad del modelo dependerá en gran medida de la pregunta empleada. Esta debe ser relevante para el negocio, debe ser clara, sin dudas en cuanto a su interpretación y debemos hacerla lo más precisa posible en relación con el objetivo perseguido. El nivel de precisión marcará en gran medida su aplicabilidad. Esto no quiere decir que, a mayor precisión más útil, sino que precisamente encontrar el nivel de precisión necesario en cada caso es un proceso muy relevante.

Veamos un caso en el que queremos determinar la probabilidad de que un cliente que entra en una tienda efectúe una compra, o incluso determinar posibles acciones prescriptivas, como modificar un determinado anuncio en una pantalla de la propia tienda o que se acerque un dependiente a atenderle.

Esto podemos pensarlo tanto para una tienda física como para una tienda virtual.

La pregunta la podemos dejar más o menos abierta. Podría referirse a la posibilidad de que compre determinado artículo, a que haga cualquier compra o a que realice una compra superior a 100 euros.

En el primer caso puede ocurrir incluso que, de ese artículo concreto, el volumen de ventas sea muy pequeño y no se puedan extraer conclusiones significativas desde el punto de vista estadístico. O puede ser que, para compras inferiores a 100 euros, si trabajamos en un modelo de margen muy reducido, no nos compense hacer ninguna acción y lo consideremos una compra residual.

Podemos verlo con otro ejemplo muy diferente. Si queremos determinar la probabilidad de que uno de nuestros clientes no nos pague (un pago aplazado, un préstamo, una factura), la pregunta requiere identificar un plazo de impago. Por ejemplo, que no pague en la fecha prevista, 3 días después o incluso 30 días después.

En los dos primeros casos, podemos tener muchos impagos simplemente por cuestiones técnicas (fechas de transferencias) o incluso por error involuntario del cliente y que se subsanen a los pocos días. Esto implica que el modelo posiblemente nos dé una alta tasa de impagos con un buen nivel de predicción, pero realicemos muchas acciones de recobro innecesarias, dado que se solucionarían incluso sin acción previa.

En el tercer caso, al usar un plazo más largo (30 días), evitaremos estos aciertos en la predicción de bajo valor, pero si el plazo es excesivo, quizás lleguemos tarde con nuestras acciones de recobro y nos sea más difícil lograr el objetivo.

Es por esta razón que la determinación de la pregunta es muy relevante para la utilidad real del modelo y requiere de una fuerte interacción con el responsable del negocio para

> La clave de la construcción de un buen modelo es saber hacer la pregunta; los datos nos darán una respuesta más adecuada cuanto mejor formulada esté la pregunta.

[18] En ocasiones se hace análisis de datos para buscar relaciones entre los mismos, sin un objetivo predeterminado. No obstante, sólo cuando decidamos la relación que nos interesa y porqué dar prioridad a esa relación sobre otras, estaremos hablando de que desarrollamos un modelo. Exagerando la cuestión, puede que sea muy fácil saber el estilo de música que le gusta a nuestros clientes, pero quizás no nos sea muy útil si nuestro negocio es de reparación de calzado.

asegurar un adecuado equilibrio entre la capacidad predictiva o prescriptiva del modelo y su eficacia en el proceso de negocio correspondiente.

En muchas ocasiones, del análisis de los datos, de su disponibilidad y de las conclusiones obtenidas se llegará a la conclusión de que es mejor cambiar la pregunta objetivo para lograr respuestas más precisas o con mayores posibilidades de generar acciones resultantes.

Preparación de datos (sistematización)

Es el proceso de selección de los datos, objetos y relaciones relevantes para dar respuesta a la pregunta objetivo. Las premisas de este paso, más allá del conocimiento técnico necesario, las podríamos resumir en:

• No despreciar datos a priori. No es necesario inundarnos de datos, pero las capacidades actuales de proceso nos permiten buscar pautas y relaciones entre volúmenes increíblemente grandes de información.

• Pensar en la posibilidad de incluir datos externos, incluso de fuentes públicas.

• Mantener un diálogo intenso con los responsables del negocio para identificar posibles relaciones a indagar, la interpretación correcta de los datos y fuentes de datos que pueden no haberse considerado por ser muy específicas o incluso desestructuradas (documentos en papel que pueden llegar a digitalizarse y leerse de forma automatizada para capturar datos históricos).

Es en este momento cuando empiezan a introducirse sesgos sobre el modelo. Es verdad que los datos ya pueden estar sesgados, pero aquí comienza la intervención humana. Se decide despreciar determinados datos, sustituir unos datos por otros, corregir determinados datos faltantes o usar fuentes alternativas. Todas ellas son decisiones legítimas, pero implican, valga la redundancia, decisiones, es decir, tomar un camino entre varios posibles. Y esto implica que estamos introduciendo alteraciones en los datos.

> En el proceso de manipulación de los datos se introducen con facilidad numerosos sesgos que es fundamental comprender y gobernar.

Es imprescindible, para el buen fin del modelo, que estas decisiones sean comprendidas y compartidas por todas las partes implicadas, incluyendo al responsable de negocio.

La comprensión implica saber que determinadas decisiones suponen ciertas limitaciones para el modelo y que, por tanto, en el momento de su integración y uso real, habrá que considerar esas limitaciones.

Y decimos decisiones compartidas por ser fundamental la participación tanto de los técnicos de análisis de datos como del negocio. En ocasiones, serán cuestiones técnicas las que deban prevalecer; en otras ocasiones cuestiones de negocio. Pero es fundamental que se entienda la problemática desde todas las perspectivas, los impactos que tienen las diferentes alternativas y se decida la mejor solución de forma compartida. En este punto no puede darse nada por supuesto. En el fondo sería como entregar un coche de formula 1 al piloto sin explicarle el significado de los mandos que tiene a su disposición y que efecto tienen en el coche.

Construcción del modelo

Un modelo matemático es una relación entre ciertos objetos matemáticos y sus conexiones por un lado, y por el otro, una situación o fenómeno de naturaleza no matemática[19].

Cuando hablamos de modelos, es importante tener en consideración algunas cuestiones propias de la naturaleza de un modelo:

• Un modelo es una simplificación de la realidad y, por tanto, no pude abordar todos los casos de forma individualizada. Sería como intentar tener un mapa de escala 1:1. Su precisión sería incuestionable, al igual que su inutilidad, pues para llegar de un sitio a otro habría que desplegarlo en el terreno, ocupando el mismo espacio que el terreno mismo (es decir, el mapa sería una copia de la realidad y dejaría de ser útil para lo que fue construido).

• El modelo nos muestra relaciones estadísticas entre diferentes componentes. Algunas de esas relaciones nos sorprenderán y parecerán contradecir el sentido común. Ante esto, se deben aceptar dos consecuencias: dichas relaciones existen (siempre que el modelo esté bien construido) y eso no implica que debamos confundir la existencia de la correlación con la causalidad[20]. Este aspecto de la causalidad tiene enorme importancia, pues se corre el riesgo de acabar afirmando que, si algo funciona para el 95% de las personas, entonces es correcto[21].

• Los modelos, como ya hemos insistido anteriormente, no sirven para determinar cada caso concreto (no son leyes, sino probabilidades estadísticas). Es muy habitual usar nuestro sesgo académico y considerar un modelo como erróneo por no acertar en un caso concreto (o en varios).

Sin embargo, el modelo seguirá siendo válido, en tanto que sea capaz de decir en qué porcentaje de los casos sucederá lo previsto y con qué margen de confianza.

La construcción de modelos sigue una metodología bien conocida por los científicos de datos y con un fuerte soporte académico[22]. No obstante, el proceso incorpora varios pasos en los que se deciden determinadas preferencias de construcción que afectan sensiblemente al resultado final. Por ello, vamos a pararnos brevemente

Hay dos aspectos conceptuales clave para entender un modelo: la diferenciación entre correlación y causalidad y la naturaleza estadística (probabilística) del resultado.

[19] Blomhoj, M. (2004) Mathematical modelling - A theory for practice. En Clarke, B.; Clarke, D. Emanuelsson, G.; Johnansson, B.; Lambdin, D.; Lester, F. Walby, A. & Walby, K. (Eds.) International Perspectives on Learning and Teaching Mathematics. National Center for Mathematics Education. Suecia, p. 145-159 (#059)

[20] Para un análisis sobre el concepto de causalidad a lo largo del pensamiento filosófico véase Gustavo R. Cañadas, Maria M. Gea, J.M. Contreras y R. Roa (2015). La “Causalidad” y su relación con la correlación y asociación. JAEM 2015. (#175c). En https://17jaem.semrm.com/aportaciones/n73.pdf y el divertido estudio sobre la relación entre el consumo de chocolate y los premios nobel en Franz H. Messerli (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. The new Englando Jounal of Medicine. (#175b).

[21] Véase la opinión, que generó bastante polémica, de el redactor jefe de la revista Wired, Chris Anderson (2008). The End of Theory: The Data Deluge Makes Scientific Method Obsolete. Wired. (#175d). En https://www.wired.com/2008/06/pb-theory/

[22] En determinados ámbitos, este proceso se hace mediante algoritmos de construcción de modelos cuyo proceso no es posible desmenuzar en sus componentes de forma clara. En ocasiones se usa la terminología de lineales, logísticas u otras, afectan al modelo, por lo que es importante que el responsable sea capaz de explicitar las decisiones adoptadas y que puedan contrastarse con otras alternativas en el proceso de validación.

en algunas de estas elecciones. El responsable del modelo debe ser capaz de explicar con claridad a un tercero estas decisiones, los motivos que han llevado a ellas y qué impacto podría tener sobre otras alternativas.

a. **La segmentación**. Los modelos requieren identificar y cuantificar los atributos de los objetos que se evalúan. Así, por ejemplo, un atributo puede ser la situación laboral de una persona, lo que daría un conjunto limitado y manejable de posibilidades para crear agrupaciones. En otras ocasiones, el atributo puede ser más simple incluso, como una característica cuya respuesta solo pueda ser si o no. Sin embargo, en otros casos, podemos estar ante valores continuos, como pueden ser los ingresos medios de la unidad familiar. En estos casos es necesario establecer grupos de niveles de ingresos para poder establecer una segmentación. Incluso en variables más sencillas, puede interesar en ocasiones agrupar algunas.

Estas decisiones tienen un importante efecto en el resultado final del modelo y, en ocasiones, vienen motivadas por lograr un mayor poder predictivo del modelo. En todo caso, deben seguir una lógica adecuada al negocio y no forzar agrupaciones espurias o que, aunque adecuadas desde el punto de vista de negocio, llevan a modelos de peor rendimiento.

b. **Selección de funciones y/o algoritmos**. Como se indicaba en la definición de modelo, lo que se busca es encontrar una relación matemática, lo que implica elegir el tipo de relación más adecuada para explicar el fenómeno que estamos tratando.

Decidir si determinadas variables siguen una distribución normal o de otro tipo o si las relaciones entre variables son “modelos de caja negra”, guiadas por un algoritmo de redes neuronales, son cuestiones clave.

Es habitual probar varios modelos y escoger el de mayor poder predictivo, pero este criterio no debería ser el único. Aunque modelos más complejos pueden dar resultados excelentes, su utilización requiere de un análisis muy específico, dados los mayores riesgos asociados. Esta distinción no entra en el alcance de este documento, dada su extensión.

No obstante, si creemos que cualquier consejero que se tenga que enfrentar a este tipo de soluciones, debe poner especial énfasis en entender sus riesgos y supuestos de base.

c. **Interpretación de las variables del modelo y selección**. En el proceso de construcción, normalmente puede llegarse a diversos modelos con un poder predictivo semejante, pero que usan variables diferentes. Decidir qué variables finalmente serán consideradas en el modelo y su lógica de negocio es, posiblemente, el punto más relevante del proceso, no tanto por su relevancia matemática, sino porque esta decisión normalmente genera una mayor confianza o desconfianza hacia el modelo por parte de los usuarios.

> Los sistemas condicionan la conducta de los usuarios. Las decisiones sobre el modelo alterarán de forma paralela las prioridades del negocio y los intereses de los gestores.

Además, esta selección puede condicionar de forma significativa los indicadores de rendimiento del negocio e incluso, dependiendo de su importancia, en las propias prioridades de los responsables.

d. **Gestión de la incertidumbre del modelo**. En el proceso de construcción, tanto en el uso de los datos como en el desarrollo matemático, se realizan determinados supuestos, entre los que se encuentran los anteriores, que suponen añadir incertidumbre al modelo. Es relevante conocer todas las fuentes de incertidumbre de un modelo y decidir, en

consecuencia, si es necesario añadir algún margen de seguridad en los resultados del mismo.

Imaginemos que nuestro modelo trata de predecir la probabilidad de que una pieza se rompa y la conveniencia de su sustitución temprana. Dependiendo de los criterios de construcción del modelo, podemos haber introducido sesgos que implican un cierto error del modelo (que puede ser en una dirección u otra). Esos sesgos se deben tratar de cuantificar con los instrumentos estadísticos habituales y establecer si el resultado del modelo debe ajustarse por ese sesgo y en qué magnitud.

Por último, el proceso de construcción no puede concluir sin calcular las métricas de calidad del modelo, que se refieren fundamentalmente a su poder predictivo y capacidad discriminante.

El consejo debe ser informado con claridad de las métricas empleadas, los estándares aceptados y el nivel de cumplimiento de los modelos relevantes.

Validación

Todo proceso relevante de nuestra organización debe tener un control, proporcionado a los riesgos que se asumen. Esto implica, por sí mismo, que nuestros procesos deben tener una evaluación de riesgos y una decisión consciente sobre el nivel de control requerido.

Cuando el riesgo asumido es bajo o incluso medio, deben existir controles, pero seguramente puedan automatizarse y delegarse en el propio gestor del proceso. Sin embargo, cuando el riesgo es alto, es una práctica necesaria que el control lo realice un tercero independiente.

Definimos así la validación como el proceso por el que un tercero verifica que el modelo que empleamos cumple con los estándares internos (y externos si los hubiera[23]). Estos estándares deben incluir como mínimo elementos de calidad de los datos, de construcción del modelo, de seguimiento y control de este y de su integración en los procesos de negocio.

Integración en el proceso

Este paso es posiblemente el más sensible de todo el proceso y, en muchas ocasiones, no recibe la atención necesaria. El modelo se considera un elemento técnico, construido por un equipo experto, bajo unas premisas matemáticas que le dan una elevada seguridad aparente (con todos los *caveat* que hemos ido valorando en los apartados anteriores).

Sin embargo, es en este momento cuando el modelo se integra en un proceso real para la toma de decisiones y se incorporan elementos adicionales que, aun con la mejor de las intenciones, pueden hacer que el modelo tome decisiones totalmente equivocadas.

Un ejemplo concreto podría ser el caso de un modelo que evalúa la calidad crediticia de un cliente para decidir si se le otorga un préstamo, y donde se introducen determinadas reglas que llevan a una decisión con independencia de la respuesta del modelo.

---

[23] Por ejemplo, el Reglamento sobre inteligencia artificial de la Unión Europea exige para los modelos de alto riesgo una serie de requisitos que deben ser cumplidos y que cualquier organización debería verificar mediante una figura independiente del constructor del modelo. No entramos en este documento a realizar un análisis de dicha normativa, pero es relevante que la definición que da el citado reglamento a la inteligencia artificial se equipara a casi cualquier modelo de decisión apoyado en software, independientemente del algoritmo subyacente (aunque pueda ser una simple regresión).

Entre estas reglas podría encontrarse alguna que establezca que los clientes de banca privada se decidirán en un comité especializado y sin tener en cuenta el resultado del modelo, o que, si un cliente dejó una deuda impagada en los últimos cinco años, no se le concede el préstamo, aunque el modelo lo califique como un buen cliente, o que, si el cliente tiene más de cierta edad, el plazo del préstamo no puede ser superior a 5 años.

La diferencia entre el modelo y lo que aquí denominamos reglas especiales es que, en el primer caso, son variables que se han incorporado al modelo al ser predictivas, mientras que, en el segundo caso, son reglas establecidas por el negocio en base a criterios que, aunque pueden tener todo el sentido, no tienen un soporte estadístico en el propio modelo.

Las razones pueden ser muy variadas y legítimas. En algunos casos se incorporan porque se trata de variables de las que no existen datos para incorporarlas en el modelo; o bien son casos poco habituales y por tanto no existe capacidad para discriminar ese grupo del resto de decisiones; en otros casos pueden ser políticas de empresa, independientes del objetivo del modelo o incluso estrategias de negocio.

En todos los casos hay que verificar la idoneidad de las reglas, pero es en esta última situación cuando hay que tener especial cuidado, pues en base a esa supuesta estrategia de negocio puede acabar obteniéndose un resultado contrario al esperado y puede terminar incluso desprestigiando el modelo cuando, en realidad, las decisiones no las ha tomado el modelo, pues las reglas impuestas violentaban la propia lógica del modelo.

Además, estas reglas suelen tener una tendencia a crecer, pues a medida que se producen casos que chocan con las expectativas del negocio, se van añadiendo excepciones y matices adicionales que pueden llegar a hacer realmente complejo el proceso de decisión y, como decíamos antes, incluso llegando a tomar decisiones contrarias al modelo en un volumen de operaciones significativo.

La integración de reglas de negocio sobre las decisiones de los modelos supone una complejidad adicional que debe ser debidamente gobernada y revisada de forma recurrente.

Si recordamos el inicio de este capítulo, precisamente esta era una de las dos condiciones para poder tener una organización orientada a la gestión en base a datos: la validez estadística y la consistencia en el uso del modelo. Sin lo segundo, el modelo será acusado de ser erróneo, pero lo realmente inadecuado será el uso que se hace del mismo al imponer un corsé de reglas no basadas en datos y estadística.

Seguimiento

Cualquier modelo en uso requiere un seguimiento continuo. La razón está inserta en sus propias características y su uso. Cuando tomamos decisiones en base al modelo, estamos interviniendo en el proceso natural que trata de describir el modelo, de forma que generaremos consecuencias que, a su vez, implicarán un cambio de comportamiento. Por tanto, es necesario captar estos cambios a lo largo del tiempo e introducirlos en el propio modelo, pues de otra forma perdería precisión de forma cada vez más acelerada (lo que implica que es importante establecer un tiempo de revisión de los modelos que equilibre el coste de su mantenimiento con su curva exponencial de deterioro).

Del mismo modo, el propio comportamiento, especialmente cuando hablamos de personas, evoluciona con el tiempo e incluso la población objetivo puede ser diferente, lo que llevaría a la necesidad de recalibrar el modelo.

Esto debe llevarnos a establecer un calendario de revisiones periódicas del modelo, cuyas conclusiones deben ser expuestas con claridad y provocar acciones concretas si no se cumplen los estándares necesarios.

El gobierno del modelo y el riesgo de modelo

Como hemos expuesto a lo largo de todo el documento, el uso de modelos incrementa de forma exponencial la capacidad predictiva e incluso prescriptiva de las empresas. En contraprestación, los errores en la construcción o el uso de los modelos pueden llevar al desastre si no existen los mecanismos de control y adaptación adecuados. En este sentido, defendemos que es responsabilidad del consejo, como órgano responsable del control de los riesgos de cualquier empresa, asegurar la cobertura de determinados aspectos de gobierno:

El consejo de administración es responsable último de la diligencia debida en la gestión del riesgo de modelo, como parte de los riesgos de las tecnologías de información.

• La existencia de unos estándares internos claros sobre la construcción y el uso de modelos en la toma de decisiones. Este proceso conlleva definir el apetito al riesgo de modelo o, dicho de otra forma, la tolerancia al error y los márgenes de conservadurismo adecuados para minimizar las pérdidas por decisiones equivocadas.

• La identificación de los órganos de decisión al primer nivel que autorizan el uso de los modelos en los procesos de negocio y aquellos otros que controlan su implementación, con la adecuada presencia de todos los *stakeholders* relevantes.

• Los procedimientos que determinan los controles necesarios y qué funciones ejercen ese control, así como los mecanismos de escalado y de contingencia ante la identificación de posibles fallos.

La responsabilidad del consejo en estas materias es creciente. El reglamento DORA (resiliencia operativa) hace responsable directamente al consejo de la resiliencia de los sistemas de información[24]. Aunque este reglamento es aplicable a entidades financieras, también incluye a todos sus proveedores. Adicionalmente, otras normativas como NIS2[25] y el reglamento de inteligencia artificial imponen obligaciones al consejo cada vez más directas y que afectan a cualquier tipo de empresas[26].

Roles clave

Como toda actividad colectiva, la gestión del *big data* en las organizaciones requiere la existencia de determinadas figuras clave. Aunque aquí nos centraremos en sus responsabilidades, siempre tendremos que considerar que las funciones las desarrollan personas y que su estilo de gestión y forma de afrontar las diferentes situaciones serán

[24] En concreto, "todo lo relacionado con la capacidad de los sistemas de redes y de información de resistir, con un nivel determinado de fiabilidad, cualquier hecho que pueda comprometer la disponibilidad, autenticidad, integridad o confidencialidad de los datos almacenados, transmitidos o tratados, o de los servicios ofrecidos por tales sistemas de redes y de información o accesibles a través de ellos"

[25] Directiva (UE) 2022/2555 de 14 de diciembre de 2022 relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión (NIS 2)

[26] Para más información sobre estas responsabilidades del consejo ver Patxi Barceló. DORA (Resiliencia Operativa): Implicaciones en la gobernanza. Ethalo Advisory – Serie Reflexiones sobre gobierno corporativo. Abril 2024. Disponible en https://patxibarcelo.es/inicio/publicaciones/

realmente el motor de la organización. Insistimos en este aspecto para evitar que se asuma que este ámbito es eminentemente técnico y requiere de perfiles donde este componente predomine sobre otros. El éxito de los procesos de transformación viene de integrar lo técnico en el sentido común del negocio que no en transformar el negocio en algo técnico.

Negocio

La palabra clave es la involucración. El negocio[27] no puede ser un demandante de soluciones técnicas sin mayor aportación. La captura de datos es una responsabilidad clave y, en la mayor parte de las ocasiones, depende del negocio, del usuario del modelo. Establecer el origen correcto del dato, su interpretación o incluso su segmentación son pasos que requieren de su participación.

Posteriormente, la validación de ciertos supuestos y decisiones a la hora de construir el modelo, así como definir las variables a incorporar finalmente al modelo de entre todas las posibles, son actividades que no se pueden hacer sólo con la visión del constructor del modelo. Tanto por motivos técnicos (distinguir correlación y causalidad), como por sentido común e, incluso, por las implicaciones éticas en el uso del modelo, que veremos al final de este capítulo, es necesario que el negocio se involucre en el proceso.

Y finalmente, la integración en los procesos de la empresa del modelo es competencia del usuario del modelo, que a su vez deberá participar activamente en su gobierno y control.

Todo ello puede parecer obvio. Quién usa el modelo es el responsable final del mismo. Sin embargo, el sesgo técnico e incluso tecnológico que se suele dar a los modelos hace que muchas organizaciones miren más, o, incluso exclusivamente, al equipo que desarrolla los modelos como responsables del mismo. Es importante que la alta dirección considere a los negocios como responsables de los modelos, siendo quienes deben tener la capacidad de explicar los modelos y sus riesgos, de forma que el consejo pueda establecer el apetito de riesgo y tener una clara visión de que este se respeta en toda la organización.

Para ello, es imprescindible un modelo de gobierno y control robusto, evitando que, ante cualquier incidencia, unos se miren a otros pensando que la responsabilidad no es suya.

> El usuario del modelo (el negocio) debe sentirse el responsable directos de los modelos y de su uso.

Chief data officer (CDO)

Su mejor apodo debería ser “el facilitador”. Es una figura todavía poco comprendida en las organizaciones. En unos casos se la considera un administrador, encargado de mantener el orden dentro de los lagos de datos que mantienen las organizaciones. En otros casos, se le otorga un papel de control, siendo el responsable de garantizar que se cumplen las exigencias internas y externas respecto a los datos.

No hay duda de que estas tareas son importantes e imprescindibles en cualquier empresa y de que su ubicación en esta función es más que razonable. Sin embargo, más allá de esas asignaciones, es fundamental su papel como catalizador del uso de los datos en la organización.

Debe asegurar su disponibilidad y, sobre todo, su accesibilidad de forma sencilla y poco burocrática. Debe democratizar los datos en la organización. Pensemos en organizaciones realmente grandes, con más de 50.000 empleados y decenas de departamentos. Su objetivo

[27] Recordamos aquí que el concepto de negocio lo empleamos en este capítulo como cualquier departamento o función de una empresa que quiere usar un modelo basado en datos.

debe ser que toda la organización entienda el valor de hacer una gestión compartida de los datos. Los ahorros significativos, en costes y en tiempo, de hacer capturas únicas y correctas de los datos, en que estos se almacenen de forma coherente y con un significado claro para toda la organización, que se mantenga la consistencia y control de todos estos insumos y que encontrar los datos que cualquier departamento necesite, entendiendo su origen y significado preciso, es algo sencillo en todos los sentidos.

En las mid-caps, la eficiencia en el uso de datos y modelos es crucial.

Pero en una empresa mediana, donde los costes son un factor crucial, su papel adquiere aun más relevancia, pues sacar valor añadido a los datos requiere un esfuerzo coordinado y eficiente.

Debemos buscar los incentivos adecuados para asegurar esto en todas las organizaciones.

Big data managers

Es un *deja vu*. Ya lo hemos vivido en la interacción entre los directivos y los equipos de tecnología. Contratamos universitarios expertos en *data science*, pero tienen lenguajes diferentes que no son comprendidos por los directivos, bases de conocimiento sin espacios comunes y visiones parciales.

Los científicos de datos y los directivos están en la misma situación que los informáticos y los responsables de negocio, como si hubiéramos vuelto 30 años atrás en el tiempo.

Existe una amplia terminología técnica que nos dificulta visualizar el conjunto. Sin embargo, aunque es verdad que existe un cambio de dimensión en la gestión de los datos, no es menos cierto que el núcleo central sigue siendo el mismo de cualquier problema empresarial tradicional: hacer las preguntas adecuadas, disponer, seleccionar y tratar los datos necesarios para obtener la respuesta óptima y tener una cultura donde las evidencias predominen sobre las intuiciones.

Alrededor de ello, es cierto, tenemos datos, software para tratarlo, hardware para soportarlo y algoritmos matemáticos de elevada complejidad para tratar de resolver de forma óptima las preguntas que nos hacemos, pero estos elementos no son más que los instrumentos para llegar al núcleo.

Hay necesidad de perfiles que sea capaces de unir la visión de negocio con la de los científicos de datos.

Por tanto, son imprescindibles perfiles que sean capaces de transformar esta complejidad en conceptos inteligibles para toda la organización. Fomentar el diálogo estableciendo una *lingua franca* entre las partes, lo que requiere perfiles específicos para esa función.

Normalmente es un aspecto poco valorado y que supone un elemento diferencial en la dinamización de las organizaciones.

Aspectos éticos del big data

Los algoritmos tienen sesgos. Por su propia construcción. Si los datos de partida tienen sesgos, el algoritmo que resulta amplificará este sesgo y su uso continuado generará un efecto de profecía cumplida[28]. La diferencia está en que los algoritmos generan esos sesgos de forma consciente, en tanto que los humanos los tratamos de evitar de forma consciente.

---

[28] Algunas referencia muy didácticas sobre el sesgo de los algoritmos las podemos encontrar en El País del 22 de septiembre de 2017, en el artículo de Javier Salas “Si está en la cocina, es una mujer:

Cuando decimos que los algoritmos generan esos sesgos de forma consciente, simplemente queremos decir que, al ser su objetivo buscar la mayor precisión en el resultado, no les importa el significado de la variable que usan, simplemente es la mejor para predecir y conscientemente la incluyen en el algoritmo.

Por el contrario, los humanos tienden a evitar usar ese tipo de variables en sus predicciones. Aunque el ser humano puede manejar pocas variables y su capacidad de generar respuestas con alto poder predictivo es baja, sin embargo, conscientemente, tratará de evitar los sesgos (aunque por su limitada capacidad matemática, en muchas ocasiones no logrará evitar que el sesgo se cuele en su modelo).

Por ello, ambos mundos se necesitan. Los algoritmos por si solos prefieren fiarse de la correlación y no tienen en cuenta la causalidad. Las personas, por el contrario, solemos tener una creencia sobre la causalidad que defendemos sin verificar si los datos la corroboran.

Sin embargo, el problema no acaba aquí. No debemos olvidar que los algoritmos son, en su génesis, una construcción basada en las instrucciones de búsqueda que le damos a una máquina, incluyendo muchos supuestos. Del mismo modo, y como ya hemos indicado anteriormente, el uso del modelo y su integración en los procesos de negocio conlleva determinados ajustes que, de nuevo, vuelven a poner en riesgo este difícil equilibrio.

Dos sentencias en el ámbito europeo ya ponen el dedo en la llaga, haciendo visible la problemática en torno al uso de algoritmos en la toma de decisiones.

La primera[29], del 31 de diciembre de 2020, es del tribunal italiano de Bologna y declara discriminatorio un algoritmo automático, condenando a la empresa que lo usa al pago de una multa de 50.000€ y a publicar la sentencia en un diario nacional.

Frank, el modelo organizativo de la plataforma Deliveroo para los *riders*, está basado en la reputación digital y penaliza a aquel que, habiendo reservado una sesión de trabajo, la cancela posteriormente, aunque sea con antelación, sin tener en cuenta si esto se debe a motivos de salud, de asistencia a familiares o por su adhesión a iniciativas sindicales de huelga.

Según la sentencia, el algoritmo parte de una elección de la empresa, que consciente e insensiblemente ha ignorado o no desea tomar en consideración las diferentes razones que podrían justificar la cancelación de la reserva por parte del *rider*.

La segunda sentencia[30], de mayor importancia, si cabe, al afectar a un organismo público, es del 5 de febrero de 2020 de la Corte de Distrito de La Haya. Anula el algoritmo que usa el Ministerio de Asuntos Sociales y Empleo de Holanda para detectar fraudes en la seguridad

---

cómo los algoritmos refuerzan los prejuicios" (https://elpais.com/elpais/2017/09/19/ciencia/1505818015_847097 .html) (#062); en el libro de Cathy O'Neil: Weapons of Math Destruction: How big data increases inequality and threatens democracy (Random House LCC US. 2016) o en el libro de Paloma Llaneza: Datanomics (Deusto. 2019), especialmente en su capítulo 5 sobre Casandra. Véase también el artículo de Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez y Kai-Wei Chang: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (#064).

[29] Texto completo de la sentencia del Tribunal Ordinario de Bolonia 31 diciembre 2020, n. 29491 en http://www.bollettinoadapt.it/wp-content/uploads/2021/01/Ordinanza-Bologna.pdf (#061)

[30] Texto completo en https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878 para la versión en inglés. (#063)

social (SyRI acrónimo de System Risk Indication), que emplea cientos de datos en poder del estado para determinar un score de los ciudadanos en relación con su probabilidad de cometer fraude.

El Tribunal reconoce que evitar un delito es un objetivo legítimo y de gran relevancia social pero que el modelo de score usado puede tener efectos no deseados, como estigmatizar y discriminar a la ciudadanía en función de determinados datos personales. De hecho, la sentencia se basa en que la práctica empleada supone una vulneración de la Declaración Europea de Derechos Humanos y, en concreto, contra su artículo 8 sobre el derecho a la vida privada y familiar.

En su defensa, el Ministerio indicó que el sistema solo usa datos que el estado ya posee, y dibuja con ellos un árbol de decisión (un modelo de predicción), de modo que no se usa la IA ni tampoco algoritmos de autoaprendizaje. Sin embargo, el tribunal señala que dicha afirmación no se puede comprobar por falta de transparencia, dado que no habían recibido suficiente información sobre el sistema, ya que el Estado mantiene en secreto la forma de calcular la posibilidad de fraude.

Este último punto impacta en otro importante hecho, como es el secretismo que rodea a gran parte de los algoritmos. Si bien puede entenderse el argumento de que estos mecanismos forman parte de la ventaja competitiva de las empresas, es fundamental impulsar procesos de auditoría que validen estos sistemas y den la tranquilidad, al consejo en primer término, a los usuarios y a la sociedad en general, de que están construidos en base a determinados estándares éticos.

También en los Estados Unidos existen ya sentencias al respecto.

Por ejemplo, la administración usa un sistema de evaluación mediante un algoritmo para determinar la probabilidad de reincidencia de los delincuentes (Correctional Offender Management Profiling for Alternative Sanctions –COMPAS–)[31].

En el caso State v. Loomis, 2016, la Corte Suprema de Wisconsin revisó el empleo de dicho algoritmo y señala que, al usarlo, se desconocían los factores de riesgo utilizados, que los criterios para identificar grupos de riesgos no son válidos para su asignación a individuos concretos, que en su base de origen podrían existir prejuicios raciales y que no había sido validado en concreto para ciudadanos de Wisconsin.

Como apunte final, reflexionemos incluso sobre algoritmos muy sencillos, como es la asignación de los turnos de atención en un banco. Cada vez es más habitual que debamos introducir nuestro NIF en una máquina que nos asignará el turno. ¿Es razonable que el sistema nos asigne un puesto anterior a otro por ser cliente o tener un mayor saldo en la cuenta? ¿Es razonable que no considere darle prioridad a una persona de avanza edad y que tenga dificultades motoras? En este último caso, el error del algoritmo puede ser por omisión, lo que eleva el nivel de exigencia necesario a la hora de evaluar los estándares de los modelos de decisión. Quizás la pregunta clave sea saber si, de hacerse público el algoritmo, dañaría la reputación de la empresa.

---

[31] Este tipo de algoritmos no son exclusivos de Estados Unidos. De acuerdo con información de prensa, en el sistema penitenciario de Cataluña -que tiene cedidas las competencias en la materia- se usa un algoritmo para las decisiones sobre permisos penitenciarios. https://www.elespanol.com/invertia/disruptores-innovadores/politica-digital/organizaciones/20210221/ alejandro-huergo-permisos-penitenciarios-cataluna-algoritmo-predictivo/559694961_0.html

Una propuesta de puntos a valorar

Terminamos este capítulo aportando un resumen de lo expuesto en forma de preguntas. Son cuestiones que consideramos relevantes a la hora de acompañar el proceso de evolución de las organizaciones en el entorno del big data.

• ¿Existe una cultura en la empresa que identifica los datos y los modelos como elementos indispensables para mejorar la toma de decisiones y se asume que su uso sistemático es necesario para que sean eficaces?

• ¿Existe un conocimiento, por parte de todos los departamentos, de los datos existentes, su disponibilidad y posibilidades de uso? No es difícil encontrar ejemplos de directivos que sostienen que determinados datos no existen en la organización y, sin embargo, otro los está capturando o, incluso, usando habitualmente.

• ¿Hay una estructura tecnológica, tanto de personas como de recursos, que permite impulsar el almacenamiento consistente de datos desde todos los procesos de la organización y su puesta a disposición sencilla y eficiente?

• ¿Existe un cuadro de mando sobre la calidad de los datos que se usan en la organización? ¿Hay estándares de calidad establecidos y planes de remediación cuando no se alcanzan los mínimos?

• ¿Dispone la organización de un catálogo de modelos, entendido como una visión de alto nivel de todos los procesos que tienen integrados decisiones basadas en modelos? ¿Conoce el consejo este catálogo y recibe explicaciones periódicas sobre aquellos modelos que son especialmente sensibles o relevantes?

• ¿Existe una declaración, completa y sencilla de entender, de los riesgos que se asumen al usar un modelo de decisión en la organización? ¿Hay, en consonancia con ese análisis, un nivel de control establecido para cada modelo?

• Los responsables de negocio ¿conocen bien los modelos que afectan a su ámbito de responsabilidad? ¿existe un control específico sobre el uso final de modelo, del impacto de las decisiones contra el modelo y de las reglas de decisión que operan sobre el modelo?

• ¿Existen mecanismos de gobierno adecuados para la implantación y control de los modelos, con participación de los responsables de los negocios, las funciones de control y la involucración de la alta dirección? ¿Se explican los modelos con claridad suficiente para que los riesgos puedan ser comprendidos y valorados por todos?

• ¿Los modelos de decisión que utiliza la empresa han sido evaluados desde una perspectiva ética, de no discriminación e, incluso, de no generar efectos indeseados en las personas y la sociedad en su conjunto?

• ¿Están adecuadamente equilibrados los perfiles de las personas dedicadas a la transformación digital de la empresa? ¿Existen las capacidades técnicas suficientes, pero acompañadas de personas que son capaces de hacer de nexo de unión con el negocio y generar confianza en el proceso, haciendo inteligible el cambio y su razón de ser para toda la organización?

ETHALO

Ethics & Algorithms

Boosting a healthful data driven society

www.ethalo.es

info@ethalo.es