El **gran poder** de los megadatos

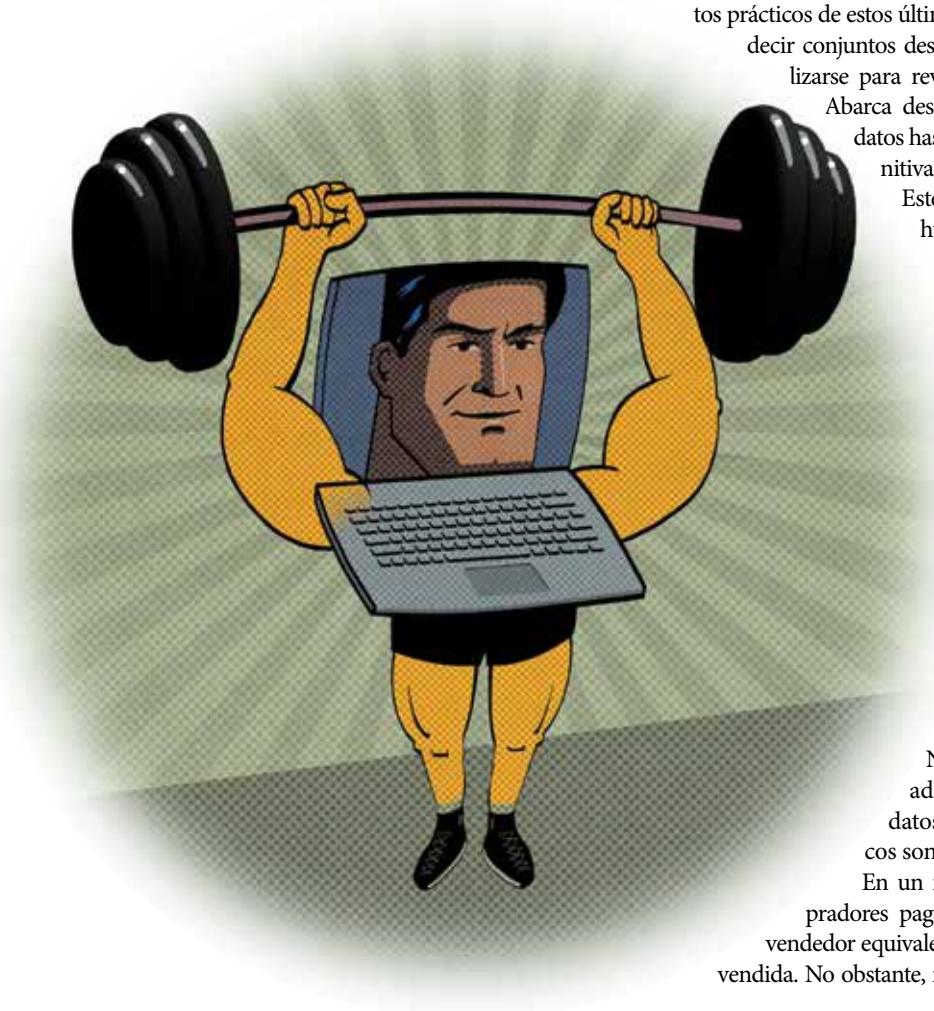
El poder informático impulsa el aprendizaje automático y transforma los negocios y las finanzas

Sanjiv Ranjan Das

EL PODER informático impulsa el aprendizaje automático y transforma los negocios y las finanzas

Hoy día el mundo tiene acceso a más datos que lo concebible apenas 10 años atrás. Las empresas están acumulando nuevos datos más rápido de lo que pueden organizarlos y entenderlos. Ahora deben descubrir cómo utilizar esta enorme cantidad de datos para tomar mejores decisiones y perfeccionar su desempeño.

El nuevo campo de la ciencia de datos procura extraer conocimientos prácticos de estos últimos, en particular de los megadatos, es decir conjuntos descomunales de datos que pueden analizarse para revelar patrones, tendencias y vínculos. Abarca desde la recopilación y organización de datos hasta su análisis y comprensión y, en definitiva, la aplicación práctica de lo aprendido. Este campo se enlaza con toda actividad humana, incluso la economía, las finanzas y los negocios.



La ciencia de datos aporta las herramientas del aprendizaje automático, un tipo de inteligencia artificial que permite que las computadoras aprendan sin programación explícita (Samuel, 1959). Al aplicar estas herramientas a grandes cantidades de datos, la gestión empresarial y el análisis de la política económica podrían cambiar radicalmente.

Algunos de estos cambios son muy promisorios.

Perfil del consumidor

No sorprende que las empresas estén adoptando cada vez más la ciencia de datos debido a que sus aspectos económicos son convincentes.

En un mercado competitivo, todos los compradores pagan el mismo precio y el ingreso del vendedor equivale al precio multiplicado por la cantidad vendida. No obstante, muchos compradores están dispuestos

a pagar más que el precio de equilibrio, y retienen el superávit del consumidor que puede obtenerse utilizando megadatos del perfil del consumidor.

Al aplicar distintos precios a los consumidores en base al análisis de sus perfiles, las empresas pueden cobrar el precio más alto que estos estén dispuestos a pagar por un producto dado. Optimizar la discriminación de precios o la segmentación del mercado empleando megadatos es sumamente rentable. Esta práctica era la norma en ciertos sectores como el de las aerolíneas, pero ahora se está extendiendo a todo el espectro de la producción.

Además, las ventajas de los objetivos de precios permiten que las empresas ofrezcan descuentos a los consumidores que de otro modo no podrían costear el precio de equilibrio, y así aumentar sus ingresos, el número de clientes y, tal vez, el bienestar social. La categorización del consumidor utilizando megadatos es una importante razón para la valoración de empresas como Facebook, Google y Acxiom, que ofrecen productos y servicios en función de los datos de sus clientes.

Si bien los megadatos pueden emplearse para explotar a los consumidores, también están cambiando las prácticas empresariales para ayudar a los propios consumidores. Las firmas están utilizando los datos generados por la interacción de la gente en las redes sociales para entender mejor su comportamiento crediticio. Vincular el historial crediticio de la gente con su presencia en las redes sociales permite mejorar los sistemas de calificación crediticia y conceder préstamos a personas que de otro modo podrían ser rechazadas.

En particular, los megadatos eliminan los prejuicios que surgen cuando la gente toma decisiones en base a una información limitada. Esta falta de datos individuales detallados dio lugar a la discriminación en las solicitudes de crédito, una práctica que data de la década de 1930. Los prestamistas hipotecarios marcaban en los mapas las zonas en las que no concederían préstamos debido a la composición racial o étnica. Esta práctica estereotipada negaba el crédito a sectores enteros de la sociedad.

Sin embargo, los megadatos eliminan los estereotipos. Los datos subjetivos brutos ahora pueden reemplazarse por datos más refinados e individualizados. Las empresas de calificación crediticia pueden aprovechar la heterogeneidad identificable en las interacciones de la gente en las redes sociales, sus flujos de mensajes, microblogs, tendencias en el uso de tarjetas de crédito y datos de categorización, además de los datos demográficos típicos, tales como ingreso, edad y localización (Wei *et al.,* 2014). El uso de datos más detallados permite clasificar mejor la calidad crediticia de cada persona.

Pronósticos y análisis del riesgo

Los pronósticos económicos han cambiado notablemente con los métodos de la ciencia de datos. En los pronósticos tradicionales, las estadísticas clave sobre la economía, como el informe trimestral sobre el PIB, se publican con considerable retraso. La ciencia de datos puede evitar esa demora utilizando información publicada con más frecuencia, como cifras sobre desempleo, pedidos industriales, o incluso opiniones sobre las noticias, para predecir las variables publicadas con menos frecuencia.

El conjunto de técnicas aplicadas en esta actividad se denomina *nowcasting* o "predicción del presente", pero se comprende mejor como pronóstico en tiempo real (véase "La reina de los números", en la edición de marzo de 2014 de *F&D).*

La ciencia de datos también está avanzando en lo que se refiere al análisis del riesgo financiero sistémico. El mundo está más interconectado que nunca, y al poder medir los vínculos sistémicos se obtiene nueva información que facilita la toma de decisiones económicas.

Observar el riesgo sistémico a través de las redes es una técnica poderosa. Los expertos en datos ahora emplean abundantes datos para elaborar panoramas de las interacciones entre bancos, empresas de seguros, corredores de bolsa, etc. Es obvia la utilidad de saber qué bancos están más conectados que otros, al igual que cuáles son los más influyentes, lo que se calcula usando un método basado en valores propios. Una vez armadas estas redes, los expertos en datos pueden medir el grado de riesgo de un sistema financiero, así como la proporción de cada institución financiera en el riesgo global, lo cual brinda a los reguladores una nueva forma de analizar y, en definitiva, gestionar el riesgo sistémico. Véanse Espinosa-Vega y Solé (2010); FMI (2010); Burdick *et al.* (2011), y Das (2016).

Estas técnicas recurren en gran medida a las matemáticas de las redes sociales desarrolladas en sociología, y están implementadas en redes enormes que emplean modelos informáticos avanzados, y culminan en una provechosa fusión de diversas disciplinas académicas.

Más que palabras

El análisis de texto es un campo de la ciencia de datos que crece velozmente y un interesante complemento de los datos cuantitativos en el área de las finanzas y la economía (véase "Las dos caras del cambio", en esta edición de *F&D*). Abundan las aplicaciones comerciales de explotación de textos: empresas como iSentium extraen opiniones de corto y largo plazo de las redes sociales que utilizan Twitter; StockTwits brinda indicadores de opinión a través de una aplicación web autoadaptable a dispositivos móviles.

> Los megadatos eliminan los prejuicios que surgen cuando la gente toma decisiones en base a una información limitada.

Ahora es posible clasificar a una empresa por las utilidades trimestrales declaradas en el formulario 10-K, un informe anual sobre el desempeño financiero de la firma, que esta presenta a la Comisión de Valores y Bolsa de Estados Unidos (SEC, por sus siglas en inglés). El conteo de palabras vinculadas con riesgos halladas en estos informes constituye un sistema de clasificación exacto para pronosticar utilidades. Las empresas cuyos informes trimestrales son más difíciles de leer tienden a tener peores utilidades, probablemente porque intentan informar malas noticias empleando un lenguaje confuso (véase Loughran y McDonald, 2014). Empleando el Índice Gunning Fog, una vieja medida de la legibilidad, es fácil calificar los informes financieros en este aspecto, y entidades reguladoras

como la Oficina para la Protección Financiera del Consumidor están considerando establecer normas de legibilidad.

En la literatura se ha concluido que la mera longitud del informe trimestral basta para detectar malas noticias (informes más largos presagian una disminución de utilidades) debido a la correlación entre confusión y verborrea; en último caso, el mero tamaño del archivo de declaración que cada empresa sube al sitio web de la SEC indica los resultados trimestrales. Por su rápida evolución, esta esfera de trabajo despierta grandes expectativas.

Una nueva disciplina denominada "análisis de noticias" extrae datos de las noticias. Empresas como RavenPack brindan cada vez más servicios, que abarcan desde la clasificación de opiniones y el análisis predictivo para el comercio hasta la elaboración de pronósticos macroeconómicos. RavenPack examina grandes cantidades de datos desestructurados de las noticias y redes sociales y los convierte en datos e indicadores granulares para apoyar a las empresas en la gestión de activos, creación de mercados, gestión del riesgo y cumplimiento.

Dentro de esta categoría, el análisis del flujo de noticias tiene especial interés. Los fondos de inversión especulativos examinan a diario miles de fuentes de noticias para extraer los cinco o diez temas más importantes y luego rastrean la evolución diaria de la proporción de temas para detectar cambios en las tendencias del mercado. Un análisis similar sería útil para autoridades y entidades reguladoras, como los bancos centrales. Por ejemplo, tal vez sea oportuno revisar la política de tasas de interés cuando la proporción de ciertos temas tratados en las noticias (como inflación, tipos de cambio o crecimiento) cambia abruptamente.

El análisis de temas comienza con la construcción de un cuadro gigante de frecuencia de palabras, conocido como "matriz término-documento", donde se catalogan miles de artículos de noticias. Los términos (palabras) son las filas del cuadro y cada artículo de noticias es una columna. Esta gran matriz puede descubrir temas mediante el análisis matemático de la correlación entre palabras y entre documentos. Los conglomerados de palabras se indexan y los temas se detectan mediante el uso de aprendizaje automático como la indexación semántica latente y la asignación de Dirichlet latente (LDA, por sus siglas en inglés). El análisis de LDA genera un conjunto de temas y listas de palabras que aparecen en estos temas.

Estos métodos de modelaje son demasiado técnicos para explicarlos aquí, pero en realidad son solo técnicas de estadística que descubren los principales grupos de palabras en un conjunto de documentos (por ejemplo, de noticias). Es probable que estas pistas de lenguaje sean ampliamente utilizadas por las autoridades económicas, así como en la toma de decisiones de políticas, por ejemplo, para redefinir el mensaje de una campaña política.

La inteligencia artificial y el futuro

Las computadoras son más poderosas que nunca y su capacidad para procesar grandes cantidades de datos ha estimulado el campo de la inteligencia artificial. Una nueva clase de algoritmos denominados "redes de aprendizaje profundo", inspirados en las redes neuronales biológicas, han demostrado su inmenso poder para imitar el funcionamiento del cerebro y constituyen numerosos ejemplos exitosos de inteligencia artificial.

El aprendizaje profundo es una metodología estadística que emplea redes neuronales artificiales para mapear un gran número de variables de entrada hacia variables de salida, es decir, para identificar patrones. La información se analiza a través de una red de neuronas basadas en silicio y *software*. Los datos se utilizan para reforzar las conexiones entre estas neuronas, del mismo modo en que los humanos aprenden de la experiencia con el paso del tiempo. El asombroso éxito del aprendizaje profundo obedece a dos razones: la disponibilidad de enormes cantidades de datos de los que las máquinas aprenden y el crecimiento exponencial del poder de cálculo gracias al desarrollo de chips de computadora especiales para aplicaciones de aprendizaje profundo.

El aprendizaje profundo impulsa buena parte de la tecnología moderna que el mundo está empezando a dar por hecha, como la traducción automática, los vehículos autónomos y el reconocimiento y rotulación de imágenes. Es probable que esta clase de tecnología muy pronto cambie la economía y la política. Las calificadoras de riesgo ya la utilizan para generar informes sin intervención humana. Grandes redes neuronales pronto podrán elaborar pronósticos e identificar relaciones entre variables económicas de mejor manera que los métodos estadísticos estándar.

Es difícil predecir qué dominios de la ciencia funesta experimentarán el mayor crecimiento en el uso del aprendizaje automático, pero ya estamos definitivamente en esta nueva era. En palabras del famoso escritor de ciencia ficción William Gibson, "El futuro ya está aquí, solo que no está distribuido en forma equitativa". ■

Sanjiv Ranjan Das es profesor en la Escuela de Negocios Leavy de la Universidad de Santa Clara.

Referencias:

Billio, Monica, Mila Getmansky, Andrew W. Lo y Loriana Pelizzon, 2012, "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors", Journal of Financial Economics, *vol. 104, No. 3, págs. 535–59.*

Burdick, Douglas, Mauricio A. Hernandez, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana Stanoi, Shivakumar Vaithyanathan y Sanjiv Das, 2011, "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study", IEEE Data Engineering Bulletin, *vol. 34, No. 3, págs. 60–7.*

Das, Sanjiv, 2016, "Matrix Metrics: Network-Based Systemic Risk Scoring", Journal of Alternative Investments, *vol. 18, No. 4, págs. 33–51.*

Espinosa-Vega, Marco A., y Juan Solé, 2010, "Cross-Border Financial Surveillance: A Network Perspective", IMF Working Paper 10/105 (Washington: Fondo Monetario Internacional).

Fondo Monetario Internacional (FMI), 2010, "Systemic Risk and the Redesign of Financial Regulation", Global Financial Stability Report, *capítulo 2 (Washington, abril).*

Lin, Mingfen, Nagpurnanand Prabhala y Siva Viswanathan, 2013, "Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending", Management Science, *vol. 59, No. 1, págs. 17–35.*

Loughran, Tim, y Bill McDonald, 2014, "Measuring Readability in Financial Disclosures", Journal of Finance, *vol. 69, No. 4, págs. 1643–71.*

Samuel, A.L., 1959, "Some Studies in Machine Learning Using the Game of Checkers", IBM Journal of Research and Development, *vol. 3, No. 3, págs. 210–29.*

Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte y Chrysanthos Dellarocas, 2015, "Credit Scoring with Social Data", Marketing Science, *vol. 352, págs. 234–58.*