Tema 37: analítica predictiva I

Peter B. Mandeville*

In 1995, in Chicago, Bezos [Jeff Bezos, CEO Amazon.com] manned an Amazon booth at the annual conclave of the publishing industry …. Roger Doeren … was stopped short by Amazon's sign: "Earth's Biggest Bookstore." … Doeren considered this, then asked, "What's your business model?" Bezos said that Amazon intended to sell books as a way of gathering data on affluent, educated shoppers. The books would be priced close to cost, in order to increase sales volume. After collecting data on millions of customers, Amazon could figure out how to sell everything else dirt cheap on the Internet. … Before Google, and long before Facebook, Bezos had realized that the greatest value of an online company lay in the consumer data it collected.

G. Packer [1]

With your permission you give us more information about you, about your friends, and we can improve the quality of our searches. We don't need you to type at all. We know where you are. We know where you've been. We can more or less know what you're thinking about (Eric Schmidt, ex CEO Google).

D. Thompson[2]

As soon as we get them buying diapers from us, they're going to start buying everything else too (Andrew Pole, Sr. Group Manager-Marketing BI, Target).

F. Provost, T. Fawcett[3]

Cuando las organizaciones obtienen más datos y comienzan a resumirlos, tienden a manejarlos para mejorar las estimaciones, las previsiones, las decisiones y la eficiencia.[4]

El término análisis predictivo, *predictive analytics* (PA, por sus siglas en inglés), saltó a la fama en 2005, debido a la introducción de *Google Analytics*. El PA se basa en varias disciplinas: el aprendizaje automático, la inteligencia artificial, la cibernética, el análisis de datos, las redes neuronales, el reconocimiento de patrones, la estadística, el descubrimiento de conocimiento, la minería de datos y la ciencia de datos, entre otras.[4-7]

El PA abarca una variedad de técnicas estadísticas de modelado, aprendizaje automático y minería de datos que analizan los hechos históricos y actuales para predecir eventos, tendencias y patrones de comportamiento.[8]

Se ha utilizado el PA para:[9]

1. Predecir los resultados de las elecciones.
2. Identificar y filtrar los mensajes de spam de correo electrónico.
3. Prever la actividad delictiva.
4. Automatizar las señales de tráfico, de acuerdo a las condiciones del camino.
5. Producir estimaciones financieras de tormentas y desastres naturales.
6. Examinar los clientes perdidos.
7. Crear aviones de autopilotaje y coches de autoconducción.

8. Identificar a los individuos con la capacidad de hacer donativos.
9. Determinar tipos de consumidores susceptibles a publicidad.

En octubre de 2010, Andrew Pole, quien administra a los profesionales que ejecutan varios proyectos de PA en Target (la quinta tienda minorista en EUA), pronunció el discurso de apertura en *Predictive Analytics Mundial*: www.pawcon.com/Target. Hacia el final de su presentación, Pole describió un proyecto para predecir el embarazo del cliente.[5]

Los consumidores tienden a la inercia en sus hábitos de compra, y cambiarlos es muy difícil. Los tomadores de decisiones en Target sabían que con la llegada de un bebé a la familia la gente cambia sus hábitos de compra.[3] Pueden empezar a buscar a nuevas tiendas y desarrollar nuevas lealtades de marca.[10]

La mayoría de los minoristas sabe esto y compite entre sí, al tratar de vender productos relacionados con el bebé para los nuevos padres. La mayoría de los registros de nacimiento son públicos, así los minoristas obtienen esta información y envían ofertas especiales a los nuevos padres.[3]

Los vendedores de Target encargaron su división de análisis para verificar si había una manera de descubrir los embarazos de las clientas a través de sus patrones de compra.[10]

Normalmente se utiliza el PA para predecir el comportamiento de compra. Aquí, lo que se predijo no era una venta directa, sino más bien algo que podría ser en sí mismo un fuerte predictor de una amplia gama de compras.[5] Ellos estaban interesados en predecir qué personas esperan un bebé. Si pudieran hacerlo, ganarían una ventaja al realizar ofertas antes que sus competidores.[3]

El equipo de análisis examinó los historiales de compras de las mujeres que se inscribieron en su registro regalo-bebé. Se dieron cuenta de que estas mujeres compraron una gran cantidad de loción sin aroma en el tercer mes de embarazo, y que unas semanas más tarde tendían a comprar suplementos como el magnesio, calcio y zinc. El equipo finalmente descubrió alrededor de dos docenas de productos que permiten a la empresa calcular una puntuación de "predicción de embarazo" por cada cliente que paga con tarjeta de crédito, tarjeta de fidelidad o cupones por correo. Las correlaciones incluso permiten al minorista estimar la fecha de nacimiento dentro de un rango estrecho, por lo que podría enviar cupones correspondientes para cada etapa del embarazo.[3,10,11]

Con los modelos se calificó la probabilidad de que una clienta determinada esté embarazada. Los modelos incluyen los patrones de compra de los productos relacionados con el embarazo, la edad, el estado civil y los sitios web visitados. Las predicciones se basan en hechos que ocurrieron durante un periodo de tiempo, no en hechos aislados. Una vez identificadas como probables embarazadas, Target les podría entonces enviar cupones especializados para productos como lociones y pañales.[11]

Los detalles del modelo exacto que utilizó Target para predecir el embarazo de la clienta no están disponibles; sin embargo, una manera de construir un modelo de este tipo es utilizar la predicción basada en la clasificación. El procedimiento general se vería así:[11]

1. Recopilar datos sobre el pasado de clientes actuales o potenciales y sus actividades en el tiempo en el ciberespacio.
2. Recopilar datos de clientes que compran los productos que les interesan.
3. Seleccionar datos de entrenamiento que se utilizará para construir su modelo basado en la clasificación y dejar de lado algunos de los datos del pasado para su uso en pruebas de su modelo.
4. Probar el modelo hasta que sea validado y los resultados en datos históricos sea aceptable.
5. Desplegar su modelo. A medida que nuevos datos de entrada para una clienta llegan, el modelo la clasificará, ya sea como potencialmente embarazada o no.

Target juntó los datos de entrenamiento mediante la fusión de los datos del registro del bebé con otros datos de las clientas, y generó un modelo predictivo "bastante exacto". El registro del bebé no sólo revela el embarazo, sino también la fecha de nacimiento. La tienda puede aplicar el modelo a las clientas que no se han registrado como embarazadas, y éste ha identificado 30% más de clientas a Target para contactar.[5]

Siegel relata que, unos meses después de la presentación de Pole, fue entrevistado por el periodista Charles Duhigg, del *New York Times*, quien le pidió

que describiera algunos hallazgos interesantes surgidos de PA. Siegel mencionó varios, incluyendo la predicción del embarazo, y proporcionó la URL del video de la presentación de Pole. En febrero de 2012, Duhigg publicó un artículo en el *New York Times Magazine* con extractos de su libro.[5,12,13]

Duhigg relata que, un día, un hombre enojado irrumpió en una tienda Target en Minnesota para ver al gerente. "Mi hija recibió esto en el correo!, gritó, ella todavía está en la escuela secundaria, ¿y ustedes están enviando sus cupones para ropa de bebé y cunas?¿Estás tratando de animarla para quedar embarazada?". Cuando el gerente llamó al hombre, unos días más tarde, para disculparse, la voz del otro lado de la línea comentó: "Tuve una charla con mi hija, dijo, resulta que ha habido algunas actividades en mi casa y no me las había informado. Ella está esperando para agosto. Te debo una disculpa".[10,11]

Peng dice que el método de investigación y el análisis no dependen sólo de las escalas de medición de las variables, sino de los objetivos de la investigación que pueden incluir:[14]

1. Descriptivo - toda una población.
2. Exploratorio - una muestra aleatoria con muchas variables medidas.
3. Inferencial - la población adecuada, muestras aleatorias.
4. Predictivo - un entrenamiento y prueba de datos establecidos de la misma población.
5. Causales - datos de un estudio aleatorizado.
6. Mecánico - datos sobre todos los componentes del sistema.

En el caso del PA, el objetivo es la predicción. Debe hacerse notar que la correlación no implica causalidad.[5] El descubrimiento de una relación predictiva entre A y B no significa que uno causa al otro.[5] Se sabe el qué, pero no se sabe el por qué y a menudo no importa necesariamente. El objetivo es predecir y no entender el mundo, y averiguar qué lo hace funcionar.

Referencias

1. Packer G. Amazon is good for customers. But is it good for books? The New Yorker. February 17, 2014. Available from: http://www.newyorker.com/magazine/2014/02/17/cheap-words .
2. Thompson D. Google's CEO: 'The Laws Are Written by Lobbyists'. The Atlantic. October 1, 2010. Available from: http://www.theatlantic.com/technology/archive/2010/10/googles-ceo-the-laws-are-written-by-lobbyists/63908/ .
3. Provost F, Fawcett T. Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. Sebastopol, CA: O'Reilly Media; 2013.
4. Abbott D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. New York: John Wiley & Sons; 2014.
5. Siegel E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. New York: John Wiley & Sons; 2013.
6. Wikipedia contributors. Predictive analytics. Wikipedia, The Free Encyclopedia. August 27, 2014, 07:16 UTC. Available from: http://en.wikipedia.org/w/index.php?title=Predictive_an alytics&oldid=622990178. Accessed August 31, 2014.
7. Wikipedia contributors. Predictive modelling. Wikipedia, The Free Encyclopedia. August 26, 2014, 17:06 UTC. Available from: http://en.wikipedia.org/w/index.php?title=Predictive _modelling&oldid=622908302. Accessed August 31, 2014.
8. Wikipedia contributors. Machine learning. Wikipedia, The Free Encyclopedia. August 22, 2014, 06:59 UTC. Available from: http://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=622303354. Accessed August 31, 2014.
9. Lantz B, Machine Learning with R. Birmingham, UK: Packt Publishing Ltd; 2013.
10. Mayer-Schonberger V, Cukier K. Big Data. New York: Houghton Mifflin Harcourt Publishing; 2013.
11. Bari A, Chaouchi M, Jung T. Predictive Analytics for Dummies. New York: John Wiley & Sons; 2014.
12. Charles Duhigg. The Power of Habit. New York: Random House Publishing Grou 13. p. 2014.
14. Duhigg C. How Companies Learn Your Secrets. New York Times Magazine. February 16, 2012. Available from: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted =all&_r=0
15. Peng RD. Structure of a Data Analysis Part 1. Reproducible Research week 1. Johns Hopkins Bloomberg School of Public Health. May 5, 2014.