

CAPÍTULO 9

UTILIZAR EL DATA MINING PARA UN MEJOR USO DE SU BASE DE DATOS

POR IVÁN GÓMEZ VILLAFANE

INTRODUCCIÓN

Este capítulo es un delineamiento general de un sencillo proceso de minería de datos para optimizar campañas de venta directa de un producto. Si bien un proceso completo es bastante más complejo y minucioso, entender uno básico le permitirá comprender, a grandes rasgos, de qué manera funciona esta técnica de extracción de conocimiento y por qué toda empresa que disponga de una base de datos debería aplicarla para sacar más provecho de la información de la que dispone.

Considere una empresa para la cual trabajamos que, todos los meses, hace llamados vía *telemarketing* ofreciendo un producto determinado a aproximadamente el 3% de su cartera de clientes. Antes de usar minería de datos, seleccionaba a este pequeño grupo a través de segmentaciones basadas en su propia experiencia y conocimiento del negocio, por ejemplo, considerando a quienes ya tuvieran otro producto similar o estuvieran dentro de cierto rango de edad e ingresos. Estas segmentaciones intentan **predecir** quiénes son sus clientes más propensos a aceptar la oferta del producto. Generalmente, esas campañas tenían una tasa de éxito de 4%. Nos pidieron hacer un **modelo predictivo**, para mejorar esa tasa, que se dirigiese a clientes más propensos

todavía. Cuando usaron nuestras predicciones obtuvieron una tasa de 8%, lo que generó la duplicación de sus ventas. ¿Cómo la minería de datos permitió esto?

La **minería de datos** consiste en, como su nombre lo indica, **minar datos** en búsqueda de información valiosa. Los **datos** que usamos son los registros digitales de la empresa, **bases de datos**, donde se guarda toda la información actual e histórica de sus clientes, lo que incluye quiénes ya fueron objetivos de la campaña de ventas en cuestión, qué características tenían en ese momento y si compraron el producto o no. A través de técnicas estadísticas y algoritmos matemáticos, lo que hicimos fue encontrar patrones que nos permitieron diferenciar a los que en el pasado aceptaron la oferta de los que no. Luego, usamos esos mismos patrones para dirigir mejor campañas futuras.

Para pensarlo de la manera más sencilla posible, considere la siguiente tabla de datos que contiene información de una campaña de ventas de seguros *premium* para autos:

CLIENTE	TIENE_AUTO	ACEPTO_SEGURO
1	SI	SI
2	SI	SI
3	NO	NO
4	NO	NO

A simple vista, podríamos decir que en el futuro le ofreceremos el seguro a quienes tengan auto. Muy lógico y sencillo. Pero amplíemos un poco más la misma información, agregando más filas y columnas.

CLIENTE	TIENE_AUTO	INGRESOS	ACEPTO_SEGURO
1	SI	MEDIOS	SI
2	SI	ALTOS	SI
3	NO	ALTOS	NO
4	NO	MEDIOS	NO
5	SI	ALTOS	SI
6	SI	BAJOS	NO
7	SI	MEDIOS	SI
8	NO	ALTOS	NO

Quizá le cueste un poco más, pero con algo de esfuerzo podrá volver a encontrar otra regla que le dará un 100% de efectividad para diferenciar a los que aceptaron el seguro de los que no: si tiene auto e ingresos medios o altos, aceptará el seguro; si no tiene auto o tiene ingresos bajos, no aceptará el seguro.

Seguramente usted estará pensando que en el mundo real no existen ni tales reglas de 100% de efectividad, ni campañas de ventas masivas de ocho casos, ni grandes empresas que solamente conozcan el nivel de ingresos del cliente y si tiene auto o no. La minería de datos se aplica justamente para poder lidiar con grandes volúmenes de información y encontrar complejos patrones, involucrando quizá decenas de variables (columnas) y decenas de miles de casos (filas). Un patrón más cercano a la realidad puede verse similar a la siguiente fórmula:

$$0.23 * VAR1 + 0.12 * VAR2 + 0.78 * VAR3 - 0.01 * VAR4 + 1.98 * VAR5 - 0.25 * VAR6 = PROPENSION$$

Aquí se involucran a seis variables, multiplicadas por un coeficiente determinado por un algoritmo matemático. Luego, se suman entre sí para dar un resultado final que indica la propensión a comprar el producto. Si bien suele ser más complejo, puede entenderse como si el algoritmo hubiera asignado un **peso** (coeficiente) a cada variable según su importancia en relación con la propensión a la aceptación de la oferta. Sumados todos los términos, a mayor resultado final, mayor propensión.

La minería de datos se encarga de lidiar con los grandes volúmenes de datos que manejan las grandes empresas de hoy en día, encontrando patrones complejos con fines predictivos.

Un departamento de Recursos Humanos para el cual trabajamos, consideraba que cierta variable era muy importante para determinar la propensión a darse de baja voluntariamente de sus colaboradores. Luego del análisis que realizamos, descubrimos que esa variable sí era importante, pero había otra que no solamente lo era mucho más sino que opacaba y dejaba redundante a la que ellos nos habían señalado.

Naturalmente, según la pregunta que hagamos, las variables que surgirán como relevantes serán distintas, o quizá sean las mismas pero con distinta interpretación. Considere un banco y la importancia del saldo disponible en cuenta con relación a colocar un plazo fijo u obtener un préstamo. En ambos casos seguramente la variable sea importante, pero en el primero probablemente la propensión tienda a aumentar con el saldo y en el segundo a disminuir con el mismo.

Una de las funciones más importantes es determinar sistemáticamente qué variables son relevantes para responder la pregunta que le hacemos a nuestra información.

COMPRENSIÓN DEL NEGOCIO

Como punto de partida, el proceso debe centrarse en el aspecto por mejorar dentro del negocio. De nada servirán miles de tablas y cientos de algoritmos y fórmulas probadas si en realidad no mejorarán el negocio de la empresa. Para nuestra campaña de ventas el objetivo es claro: si sabemos quiénes son nuestros clientes más propensos a aceptar la oferta, podemos dirigirnos a ellos y obviar a aquellos con los cuales solamente gastaremos tiempo y dinero. Algunas de las preguntas más comunes son:

- ¿Qué clientes van a rescindir su contrato con nosotros en los próximos tres meses?
- ¿Cuánto podré facturarle a cada cliente en los próximos tres meses?
- ¿Qué producto debo ofrecerle a cada cliente?
- ¿Qué clientes son más propensos a aceptar nuestra oferta?

Los interrogantes pueden ser sencillos, escritos en una línea. Lo que no será tan sencillo será trasladarlos a nuestros datos. Ahora veremos por qué.

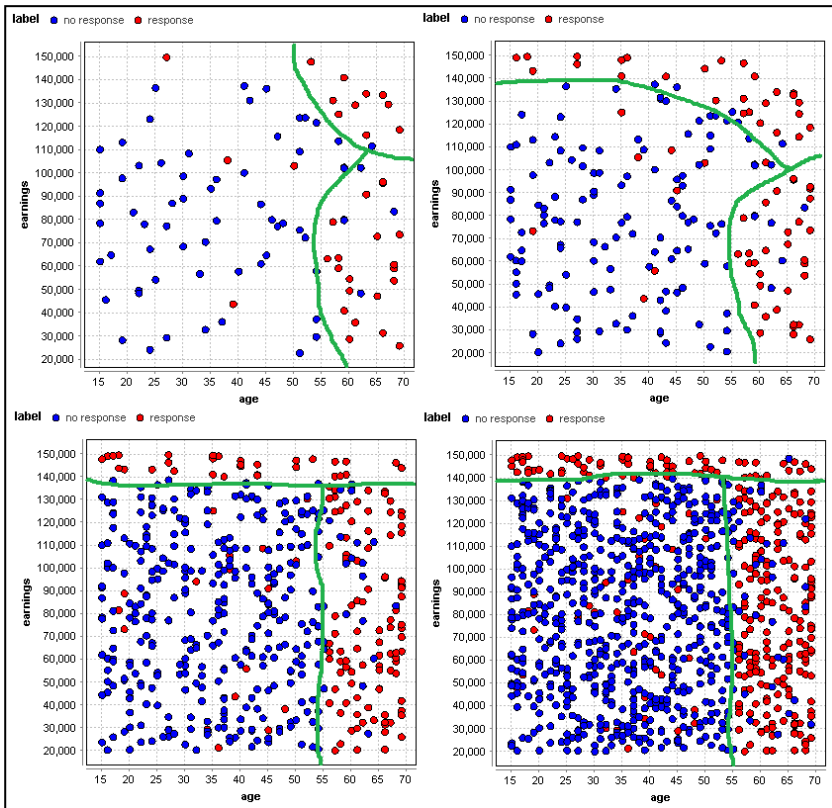
COMPRENSIÓN DE DATOS

Muchas veces este paso suele estar antes que el anterior, o en simultáneo, ya que en realidad, las preguntas que podremos responder estarán condicionadas por la información que tengamos. A continuación, enumeramos algunas de las características que deberán tener los datos con los cuales trabajaremos:

- **Existencia.** Puede parecer obvio, pero no podremos hacer nada si no hay ya una mínima base de datos. Si no la hubiera, se puede pensar y sugerir qué datos y en qué forma podrían ser útiles para nuestro objetivo, según los próximos puntos de esta lista. Cuanto más grande sea la organización y cuanta más información tenga, más efectividad tenderá a tener la minería de datos.
- **Valor predictivo.** Debemos contar con información que intuitivamente estará correlacionada con lo que queremos investigar. Podemos tener cincuenta años de historia climática de Hungría en una base de datos

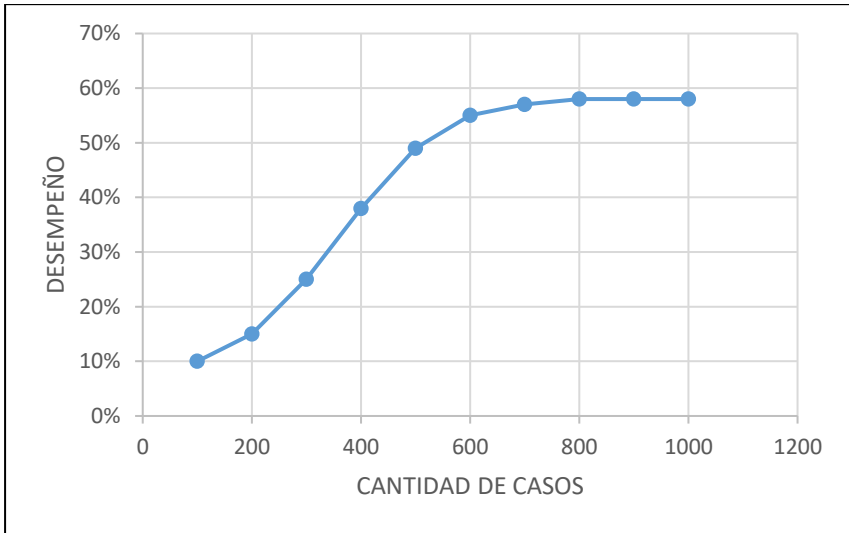
perfectamente organizada, pero de nada nos servirá para predecir la aceptación de una oferta de determinado producto.

- **Veracidad.** Supongamos que contamos con una serie de variables demográficas, sociales y laborales de nuestros prospectos, las cuales esperamos que puedan contribuir a nuestro modelo predictivo. Deberíamos intentar asegurarnos de que esa información sea real, ya que si tuviera muchos datos falsos o erróneos no servirá a los algoritmos.
- **Velocidad.** La información que se ha de utilizar tiene que estar disponible con cierta rapidez, según cada caso y el poder predictivo que esperemos tener. Si quisiéramos predecir la propensión a la compra según datos del prospecto de hace tres años atrás, vamos a tener mucho menos éxito que si lo hiciéramos con datos del mes pasado. El detalle es que cada sistema de datos tiene su propio esquema de procesamiento, por lo tanto, debemos asegurarnos de que las ventanas de tiempo sean razonables.
- **Volumen.** Para encontrar patrones estadísticos en nuestros datos, necesitamos que los mismos tengan cierto volumen. Anteriormente hemos visto una fórmula matemática que asignaba pesos a una serie de variables para determinar la propensión a aceptar la oferta. Veamos ahora cómo dos variables explican la respuesta a la oferta:



En el eje X tenemos la edad del individuo y en el eje Y, sus ingresos. En color azul se representan los casos que no respondieron a la oferta y en color rojo, los que sí. La línea verde indica las separaciones que un modelo podría hacer para diferenciar entre ambos casos. De izquierda a derecha y de arriba abajo, el primer gráfico contiene 100 ejemplos, el segundo 200, el tercero 500 y el cuarto 1000. Observe cómo del primer gráfico al tercero la línea verde se va modificando, mientras que del tercero al cuarto permanece casi idéntica. Si hiciéramos un modelo con 100 o 200 ejemplos solamente, no obtendríamos los **verdaderos** patrones que describen a nuestra población, los cuales, según este ejemplo, encontraríamos a partir de 500 casos.

Precisamente, cuál es el volumen requerido va a variar según cada problema, en función de la complejidad de los patrones subyacentes que respondan el interrogante planteado. En cualquier caso, bajo un correcto esquema de evaluación, siempre obtendremos un gráfico similar al colocar sobre el eje X la cantidad de casos y sobre el eje Y el desempeño del modelo:



Inicialmente, nos beneficiaremos de agregar más y más casos, hasta que en cierto punto no habrá nada nuevo bajo el sol y más casos solamente implicarán más tiempo de procesamiento para encontrar exactamente los mismos patrones.

HISTORIA DEL EVENTO

Con **evento** nos referimos a lo que estamos intentando predecir. Este punto tiene algunas consideraciones y aproximaciones indirectas en el caso de que no estuviera el dato preciso. Supongamos que aún no se hubieran realizado campañas de ventas del producto en cuestión. No tendremos forma de responder **directamente** la pregunta planteada, pero sí podremos atacarla **indirectamente** con varias hipótesis.

Predecir la tenencia de ese producto en función del resto de las características del cliente puede ser una opción. De esta forma, obtendríamos patrones que nos indicarían que algunos clientes son muy similares a quienes ya tienen el producto, con la única diferencia de que no lo tienen. Debemos tener cuidado al incluir variables relacionadas con el producto: usar el monto total de plazos fijos para predecir la tenencia o no de los mismos nos dará un **modelo perfecto** que en la práctica no nos servirá para nada.

Observar la compra espontánea es otro camino estrechamente relacionado con el anterior. ¿Cuál es la diferencia? En el caso anterior, consideramos la tenencia del producto independientemente de la antigüedad del mismo. En éste, buscamos los patrones que se presentan exactamente en el momento de la compra. Considere una persona cuyo tiempo de viaje hasta su trabajo, usando

transporte público, sea muy elevado y haya decidido comprar un automóvil para mejorar su calidad de vida diaria. En el momento de la compra quizá haya tenido mucho dinero en el banco, ya sea a la vista o en inversiones, el cual usó y fue un factor determinante para la compra. Luego de la misma, ese dinero ya no estaría más en su cuenta. Observar la compra espontánea consideraría este factor, mientras que observar la tenencia no.

Usar campañas de otro producto similar también es razonable, por ejemplo, ofreciéndoles un seguro al perfil que haya sido propenso a aceptar otro en el pasado. Este tipo de traslados también pueden suceder con distintos tipos de campañas del mismo producto, por ejemplo, cuando usamos un historial de campañas a través de *telemarketing* para encontrar patrones para campañas a través de SMS.

¿Cuál es el mejor camino por seguir? Las tres son hipótesis (¿se les ocurren otras?) y muchas veces lo mejor que se puede hacer es probar las tres. Si el primer mes de la campaña enviara 1000 piezas, puede repartirlas entre los candidatos de las tres hipótesis en función de qué tan fuerte se considere cada una y el poder predictivo de cada modelo. Luego de observar los primeros resultados, no solo se debe dar más cuota a la hipótesis que mejor desempeño haya tenido sino también comenzar a retroalimentar el proceso con un nuevo modelo, usando los nuevos datos disponibles que ahora sí atacarán directamente la pregunta planteada.

ANÁLISIS DE DATOS

A partir de ahora, usaré el software **RapidMiner**. Lo invito a descargarlo y probarlo. Tiene una interfaz muy intuitiva y varios tutoriales interactivos. Si más adelante quisiera experimentar con nuestro ejemplo, usaremos el operador **Generate Direct Mailing Data** con 2000 ejemplos y *random seed = 1*. Esto generará un conjunto de datos genérico que contiene información de clientes y si aceptó o no una oferta propuesta. En esta muestra de los primeros diez casos, puede observar que tenemos nueve variables.

ExampleSet (1000 examples, 1 special attribute, 8 regular attributes)									
Row No.	label	name	age	lifestyle	zip code	family status	car	sports	earnings
1	no response	naVzUges	38	active	96569	married	expensive	soccer	115562
2	no response	oiSav4lw	24	cozily	76959	single	expensive	soccer	31302
3	response	P4lUHrXK	65	cozily	19415	married	expensive	soccer	21477
4	response	oin8PhBY	50	cozily	58229	married	practical	soccer	146151
5	no response	IVkEWk3q	34	active	62615	married	practical	soccer	37192
6	no response	xzdYfICT	59	cozily	64695	single	practical	soccer	98526
7	no response	PFnn0Gkq	34	active	32494	married	practical	soccer	132903
8	no response	VTmprLK3	35	active	87569	married	practical	athletics	138076
9	no response	e63heXGX	30	active	99295	married	expensive	soccer	114016
10	response	6MULo1mv	25	active	91463	single	expensive	athletics	142097

Label: marca respuesta favorable a la oferta o no. Los valores son *response/no response* y nos referiremos a los mismos como positivo/negativo, caso de interés/negativo, variable de salida, variable por predecir, etcétera.

Name: el nombre de la persona casi nunca tendrá valor predictivo e, incluso, en esta tabla figura anonimizado como una serie de letras y números aleatorios. Es conveniente no extraer nombres de bases de datos y manejarse siempre a través de un ID numérico, principalmente para proteger la confidencialidad de los individuos.

Age: la edad.

Lifestyle: el estilo de vida.

Zip code: esta columna puede ser útil para localizar geográficamente al individuo. En nuestro caso, no la usaremos y, al igual que *Name*, la dejaremos afuera en una preselección de variables.

Family status: estado civil.

Car: tipo de auto que posee.

Sports: su deporte preferido.

Earnings: sus ingresos anuales.

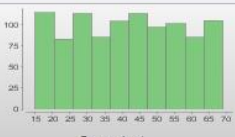
Como mencionamos anteriormente, siempre es necesario corroborar la **veracidad** de los datos con los cuales trabajaremos. A simple vista, en la tabla no vemos errores, pero es muy común encontrarse con valores perdidos o erróneos si hacemos un análisis a fondo. Es importante saber diferenciar entre ambos:

Valor perdido/blanco/nulo: ocurre cuando el espacio que corresponde al valor se encuentra en blanco. Este caso es fácil de detectar. Piense en cuando un operador ingresa los datos del cliente del sistema, y como no dispone de su edad, no coloca ningún número.

Valor erróneo: ocurre cuando la variable tiene un valor, pero este no es real o es directamente incoherente con la información que se supone que contiene o con

el resto de la tabla. Este caso puede ser más o menos difícil de detectar. Piense en el mismo operador encontrándose con que el sistema le solicita sí o sí colocar un dato. Probablemente ingrese 0, o una X, o quizá se le ocurra colocar 999.

Una de las formas de analizar datos es usando mediciones estadísticas y gráficos. Por ejemplo, podemos observar el rango de cada variable, su valor promedio y graficar un histograma para examinar su distribución.

Name	Type	Miss.	Statistics		
label	Nominal	0	Least response (313)	Most no response (687)	
age	Integer	0			
lifestyle	Nominal	0	Least healthy (320)	Most active (354)	
family status	Nominal	0	Least single (479)	Most married (521)	
car	Nominal	0	Least expensive (498)	Most practical (502)	
sports	Nominal	0	Least soccer (322)	Most athletics (340)	
earnings	Integer	0	Min 20155	Max 149962	Average 84710.103

Los datos que estamos usando como ejemplo no tienen ninguna complicación, ya que no hay casos perdidos (*Miss.*) y las distribuciones de las variables no presentan ningún valor extraño. Observe la columna *Type*, que indica de qué tipo es la variable. Las *integer* están compuestas por números enteros y las *nominal* por dos o más categorías con valores determinados.

¿Qué se podría haber hecho si contáramos con valores perdidos o erróneos? Si fueran muy pocos casos, y nuestra cantidad de casos totales lo suficientemente generosa, podríamos simplemente descartarlos. Si quisiéramos conservar los casos y no tuviéramos la edad, podríamos reemplazar el valor por el promedio. Lo mismo para los ingresos anuales. Para las variables nominales es razonable agregar otro valor llamado “perdido”, que indique no se dispone de ese dato.

PREPARACIÓN DE DATOS

Ya hemos visto cómo un modelo trabaja sobre números y recién contamos que hay variables, en este caso la mayoría, de tipo nominal. ¿Cómo podemos aplicar un coeficiente a una palabra? Ya que cada variable nominal tiene un conjunto determinado de N valores posibles, lo que haremos será transformar esa única variable en otras N variables, una para cada posible valor de la variable original. Observe la siguiente tabla:

ingresos	ingresos = bajos	ingresos = medios	ingresos = altos
bajos	1	0	0
medios	0	1	0
altos	0	0	1

Como se puede apreciar, la conversión de categorías a números es sencilla y no tiene mucho misterio. Sin embargo, disponemos también de la posibilidad de convertir la variable a una escala numérica, si es que ésta tuviera cierto orden natural. En este caso, podríamos transformar la variable de la siguiente forma:

ingresos	ingresos escala
bajos	1
medios	2
altos	3

En el caso de estudio generado en RapidMiner, convertiremos todas las variables nominales en numéricas sin ninguna escala. Nos quedaremos con la variable *family status = single* y descartaremos *family status = married*, ya que *family status* cuenta solamente con dos valores y hacer dos variables se torna redundante (cuando una vale 0 la otra vale 1, y viceversa). A esta altura, nuestra tabla inicial ha cambiado un poco y se ve de la siguiente manera:

ExampleSet (1000 examples, 1 special attribute, 11 regular attributes)												
Row No.	label	lifestyle = h...	lifestyle = a...	lifestyle = c...	family statu...	car = practic...	car = expen...	sports = so...	sports = ba...	sports = ath...	age	earnings
1	no response	0	1	0	1	0	1	1	0	0	38	115562
2	no response	0	0	1	0	0	1	1	0	0	24	31302
3	response	0	0	1	1	0	1	1	0	0	65	21477
4	response	0	0	1	1	1	0	1	0	0	50	148151
5	no response	0	1	0	1	1	0	1	0	0	34	37192
6	no response	0	0	1	0	1	0	1	0	0	59	98526
7	no response	0	1	0	1	1	0	1	0	0	34	132903
8	no response	0	1	0	1	1	0	0	0	1	35	138076
9	no response	0	1	0	1	0	1	1	0	0	30	114016
10	response	0	1	0	0	0	1	0	0	1	25	142097

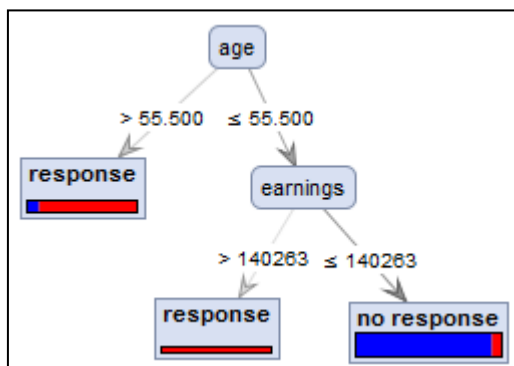
Resta solamente diseñar un esquema de evaluación. La práctica más habitual consiste en dejar una cantidad de casos completamente aislados del entrenamiento del modelo predictivo, para asegurarnos de no estar encontrando supuestos patrones que luego no se cumplirán al trasladarlos a un conjunto de datos nuevos.

Determinar qué cantidad de datos se dejarán a un lado dependerá del volumen total de información con el cual contemos. En este caso, haremos una división 50/50 entrenamiento/evaluación. Si contáramos con pocos casos, puede ser necesario usar más para el entrenamiento. En situaciones extremas, se pueden utilizar todos los datos para generar el modelo y evaluarlo con un esquema conocido como **Cross-Validation o validación cruzada**. Este método consiste en dividir la muestra en *k-folds* o *k-partes*, que generalmente son 10. A continuación, se generan *k* modelos distintos, dejando para cada uno una de las muestras a un lado y usándola para evaluar ese mismo modelo, promediando luego el resultado de las *k* evaluaciones.

EXPLORACIÓN DE ALGORITMOS

Hasta ahora hemos descartado dos variables que no usaremos, transformado a números las variables predictivas, y hemos separado a nuestros datos en un conjunto de entrenamiento y otro de evaluación. Estamos listos para experimentar con varios algoritmos y evaluar cuál nos dará mejores resultados. Si bien hay cientos de técnicas diferentes, nos concentraremos en comprender mínimamente y probar dos de ellas.

ÁRBOL DE DECISIÓN



Árbol 1

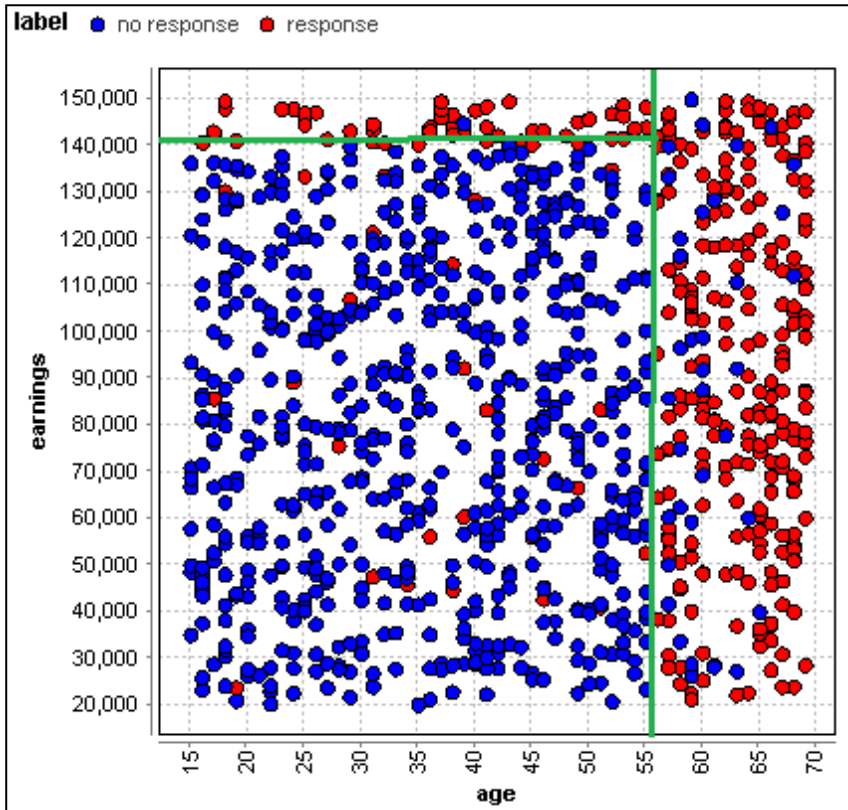
Un **árbol de decisión** consiste en una serie de reglas, una detrás de otra, usando una variable a la vez según cuál sea la que en ese segmento del conjunto de datos de entrenamiento permita diferenciar lo más posible entre las dos categorías de la variable que se ha de predecir. En el **Árbol 1**, las reglas también se pueden escribir de la siguiente manera:

```

1) age > 55.500: response {no response=31, response=229}
2) age ≤ 55.500
  2a) | earnings > 140263: response {no response=1, response=58}
  2b) | earnings ≤ 140263: no response {no response=639, response=42}
  
```

1. Si es mayor de 55 años, es tomador de la oferta en un 88,76% de los casos.
2. Si tiene 55 años o menos:
 - a. y tiene un ingreso mayor de 140.263, es tomador en un 98,31% de los casos.
 - b. y tiene un ingreso menor o igual de 140.263, es tomador en un 6,17% de los casos.

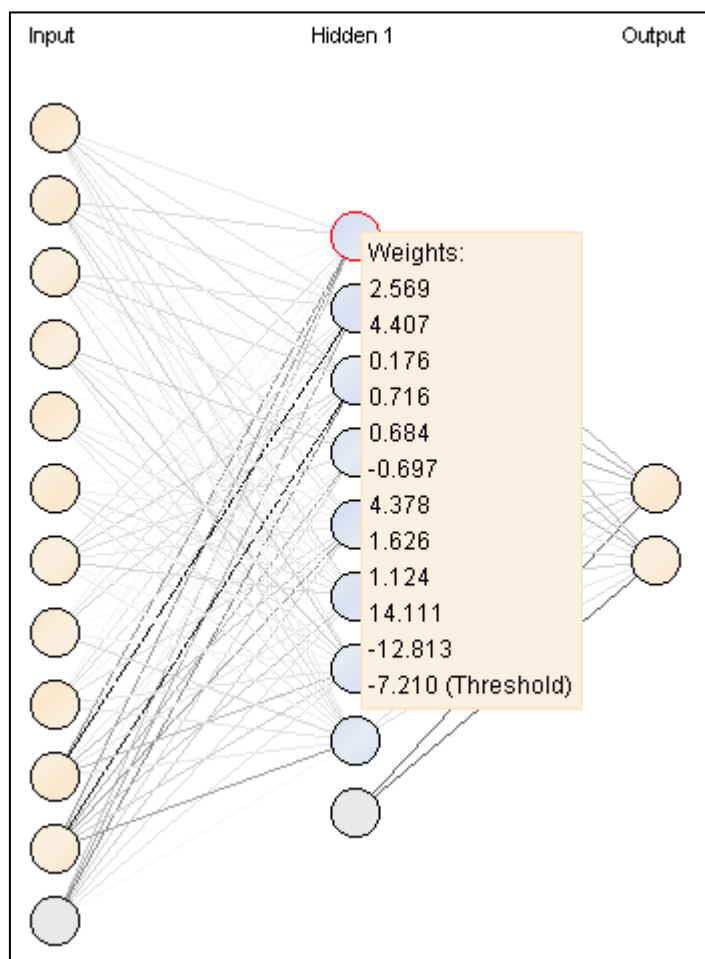
El resultado de este árbol serían 3 grupos: el (2a), con la mayor propensión a aceptar la oferta, seguido por el (1) y, finalmente, por el (2b). Observe el gráfico de los datos según la edad y los ingresos:



Marcamos con verde las líneas de separación del árbol de decisión. Podrá notar que este gráfico ya lo ha visto. ¡Es el mismo de la etapa de comprensión de datos! ¿Observa la diferencia? El algoritmo del árbol de decisión determinó primero dividir el espacio según la edad y luego según el nivel de ingresos, mientras que nosotros, a ojo, lo dividimos al revés. No nos queda ninguna duda de que la división automática hecha por el algoritmo es mejor que la nuestra. Aquí vale la pena señalar que:

La minería de datos automáticamente explora variables y decide, con la mayor precisión posible, de qué forma cada una se relaciona con el evento que se ha de predecir, superando cualquier análisis manual.

RED NEURONAL



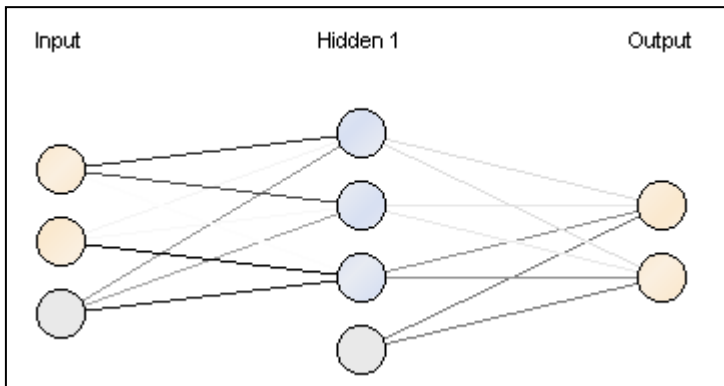
Red Neuronal 1

Se complica un poco, ¿verdad? En este caso, lo que tenemos son las 11 predictoras en la primera columna de nodos (los círculos), llamada *Input*. El último círculo gris es un nodo adicional (*Threshold*), que consiste en un número fijo (no variable). La segunda columna de nodos, la capa oculta (*Hidden 1*), tiene 8 nodos que se calculan a partir de los nodos de la capa anterior. En la imagen está seleccionado el primer nodo de la capa intermedia. Los valores que se listan debajo de *Weights* son los coeficientes por los cuales se multiplica a cada

variable de la capa anterior, para luego sumarse todos los valores y dar como resultado el valor de ese nodo en particular de la capa intermedia. Con pesos distintos para cada nodo, se calcula toda esta capa. Finalmente, se aplican otros coeficientes a la capa intermedia para dar como resultado los dos valores de la capa final. Con una función que convierte el resultado en una puntuación entre 0 y 1, cada número representa la propensión relativa a aceptar la oferta o no. ¡Cuidado! No se trata de una probabilidad **real**. Que un caso tenga puntuación 0,2 significa que viene antes que el 0,1 y después que el 0,3, no que tiene un 20% de probabilidad de comprar el producto.

Su fórmula resultaría muchísimo más larga que el puñado de reglas del árbol de decisión. Parece muy complicado, pero lo más importante es saber que este tipo de arquitecturas permiten separar el espacio sin limitarse a divisiones lineales perpendiculares a los ejes de coordenadas, como sucede con los conjuntos de reglas de los árboles de decisión. Sin embargo, a diferencia de la mayoría de los algoritmos del árbol de decisión, no tienen un mecanismo incorporado de selección de variables: simplemente intentan encontrar los mejores pesos para todos los nodos que se le entreguen. Si le entregara 2000 variables, intentaría encontrar todos los pesos de esas 2000 variables y los de su capa intermedia, incluso si 1990 de esas variables fueran completamente inútiles.

¿Cómo solucionamos este problema? Hay muchas formas de seleccionar variables; una puede ser, por ejemplo, utilizando las variables seleccionadas por algún algoritmo que sí tenga un sistema de selección. Podemos entonces utilizar las variables *age* y *earnings* que seleccionó el árbol, aplicarlas a una red neuronal y comparar los resultados.



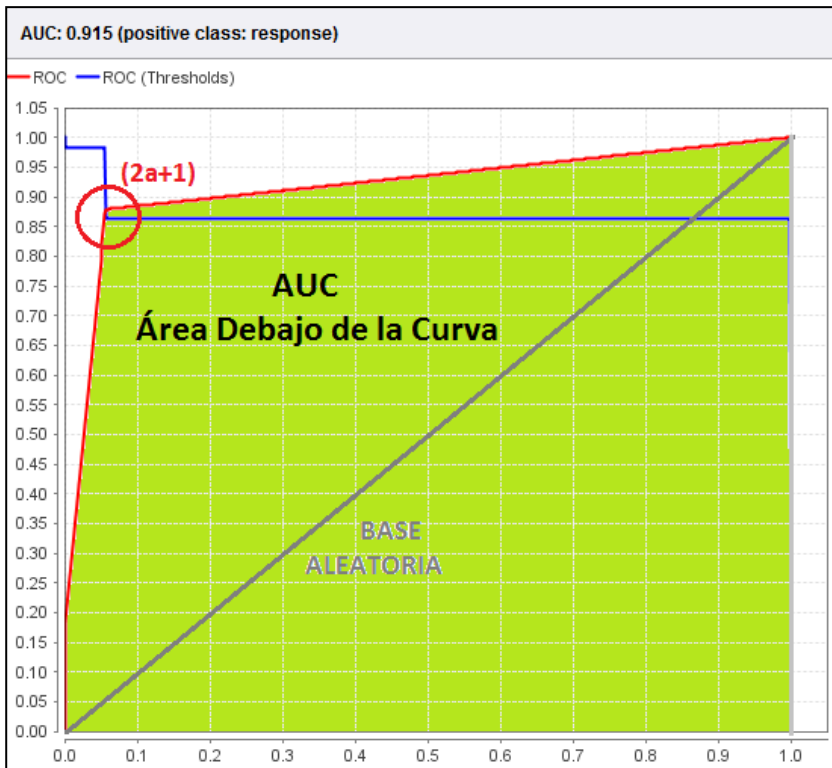
Red Neuronal 2

La red neuronal resultante es ahora mucho más sencilla y utiliza dos variables que, por un lado, un método de selección marcó como importantes y, por otro, pudimos ver que verdaderamente sí lo eran. Hasta ahora tenemos un árbol de decisión y dos redes neuronales. ¿Cómo sabemos cuál es la mejor solución?

EVALUACIÓN DE MODELOS PREDICTIVOS

Probablemente piense que la mejor manera de evaluar un modelo es observando la precisión. Por lo tanto, simplemente deberíamos aplicar los modelos al conjunto de prueba, hasta ahora separado, y examinar qué porcentaje de los casos hemos acertado en clasificarlos como tomadores o no de la oferta. Con esta lógica, un analista que nos presente que su modelo tiene una precisión de 96% sería casi perfecto. Pero hay una trampa.

Considere la campaña de ventas que le comentamos al comienzo del capítulo, en la cual solamente el 4% tomaba el producto. ¿Ya lo descubrió? Si yo le **predijera** que nadie tomará el producto, tendría ese sorprendente 96% de acierto. En problemas en los que interesa descubrir casos de interés poco frecuentes, es necesaria otra medida de evaluación conocida como **AUC** (*Area Under the Curve* o, en español, área debajo de la curva). ¿Qué curva?



Técnicamente, este gráfico se llama ROC. Quizá este sea el momento más intrincado del capítulo. ¡Coraje! Lean atentamente y estamos seguros de que lo entenderán. Les ahorramos preocuparse por la línea azul, olvídenla. En el eje X, está marcado el porcentaje de no tomadores de la oferta (*no response*), del total de no tomadores que aparecieron en toda la campaña. A este porcentaje se lo conoce como **specificity** o **especificidad**. En el eje Y, figura lo mismo para los tomadores de la oferta (*response*), conocido como **sensitivity** o **sensibilidad**. Consideremos nuestra muestra de entrenamiento de 1000 casos:

	TOTAL DE MUESTRA	BASE ALEATORIA 10%	MODELO EN (2a+1)
RESPONSE	329	33	287
NO RESPONSE	671	67	32
SENSIBILIDAD	100%	10%	87,2%
ESPECIFICIDAD	100%	10%	4,8%

La línea que va de (0,0) a (1,1) representa la base resultante de tomar un subconjunto aleatorio de los datos. Sea cual fuere la cantidad, por estadística, siempre encontraremos más o menos la misma cantidad porcentual del total de negativos y positivos de toda la campaña. Si tomamos aleatoriamente el 10% de la misma, obtendremos 33 positivos y 67 negativos, ambos representando al 10% de su categoría.

Por otro lado, la línea roja representa la mejora que puede hacer el modelo sobre esa selección aleatoria de base. Por ejemplo, en el punto marcado (2a+1) captamos al 87,2% del total de positivos con apenas 4,8% de negativos. La tasa de respuesta de este segmento es de 91,7%, muy superior al 32,9% de la campaña. ¿Cómo se compone ese punto (2a+1)? Recuerde el árbol de decisión encontrado:

```
1) age > 55.500: response {no response=31, response=229}
2) age ≤ 55.500
2a) | earnings > 140263: response {no response=1,
response=58}
2b) | earnings ≤ 140263: no response {no response=639,
response=42}
```

Pasemos los segmentos a una tabla:

SEGMENTO	2a	1	2b	TOTAL	2a+1
RESPONSE	58	229	42	329	287
NO RESPONSE	1	31	639	671	32
TOTAL SEGMENTO	59	254	681	1000	313
TASA DE RESPUESTA	98,3%	90,1%	6,1%	X	91,7%

SENSIBILIDAD	17,6%	69,6%	12,7%	100%	87,2%
ESPECIFICIDAD	0,2%	4,6%	95,2%	100%	4,8%

Ordenamos a los segmentos según su tasa de respuesta que contienen. El más importante es el (2a), con 98,3%, le siguen el (1) y el (2b). En un escenario de implementación, podríamos dirigirnos al primer y segundo segmento (2a+1), evitando el tercero por tener una tasa de respuesta muy por debajo de la media (32,9%). Si observa el gráfico, el punto (2a+1) está marcado en sensibilidad con 87,2% y especificidad con 4,8%, tal como se desprende de la tabla. El punto del segmento (2a) está casi pegado al eje Y, ya que la especificidad es un bajísimo 0,2%.

El área debajo de la curva no es más que el área verde marcada en el gráfico. Mientras mayor sea, mejor será nuestro modelo predictivo. Veamos los valores de AUC para los tres modelos entrenados:

MODELO	SELECCIÓN DE VARIABLES	AUC ENTRENAMIENTO	AUC PRUEBA	DIFERENCIA
Árbol de Decisión	Automática, usa dos.	0,915	0,936	+0,021
Red Neuronal 1	Ninguna, usa todas.	0,931	0,911	-0,02
Red Neuronal 2	Tomadas del árbol.	0,925	0,944	+0,019

Aquí hay varias cosas para observar. En primer lugar, fíjese qué hubiera sucedido si solamente hubiésemos tomado la evaluación en entrenamiento para seleccionar a nuestro modelo: creeríamos que la **Red Neuronal 1** es la mejor cuando, en realidad, en la evaluación en prueba resulta ser la peor de todas. ¿Por qué? Probablemente, debido al uso forzado de variables que en realidad no tenían una clara relación con la respuesta. El algoritmo se las rebuscó para encontrar relaciones que en un conjunto de datos diferente no existen, por lo tanto, el desempeño cae.

En segundo lugar, observe que la **Red Neuronal 2**, usando exactamente las mismas variables que el **Árbol de Decisión**, obtiene mejores resultados tanto en

entrenamiento como en prueba. Si bien uno podría caer en la trampa de decir que las redes neuronales siempre serán mejores que los árboles de decisión, recuerde que gracias al árbol pudimos mejorar el desempeño de la red que usaba todas las variables disponibles. En **minería de datos** usualmente se explora la mayor cantidad de soluciones posibles, nunca se sabe de antemano cuál será la mejor.

Finalmente, considere que la diferencia entre la evaluación y el entrenamiento es aceptable. Diferencias muy grandes indicarían que los patrones encontrados en entrenamiento explican el espacio de forma demasiado diferente en la base de prueba, sugiriendo que tales patrones no son muy reales o robustos. Para solucionar esta diferencia, se puede o bien disminuir la complejidad de los algoritmos a través de los parámetros de su configuración (tema que no veremos aquí) o bien sumar datos.

IMPLEMENTACIÓN

Una vez seleccionada la **Red Neuronal 2** como el **modelo campeón**, resta ahora llevarla a un conjunto de datos nuevos para otorgarles una puntuación según su propensión a aceptar la oferta. La estructura será igual a la base de entrenamiento, solamente que faltará la variable que se ha de predecir. Como vimos gráficamente, los más puntuados serán aquellos que tengan mayor edad e ingresos:

Row No.	prediction(label)	confidence(no response)	confidence(response)	age	earnings
1	no response	0.966	0.034	38	115562
2	no response	1.000	0.000	24	31302
3	response	0.000	1.000	65	21477
4	response	0.001	0.999	50	146151
5	response	0.024	0.976	59	98526
6	no response	0.997	0.003	34	132903
7	no response	0.662	0.338	35	138076
8	no response	0.751	0.249	25	142097
9	no response	1.000	0.000	32	115888
10	response	0.000	1.000	67	70979

Los casos 3, 4, 5 y 10 son los de más alta puntuación (*confidence*) positiva. Los casos 7 y 8 tienen una puntuación media y los restantes una puntuación baja. Con toda nuestra cartera de clientes puntuada, lo único que tenemos que hacer es ordenarla de mayor a menor puntuación y dirigirnos concretamente a los que figuren arriba de todo.

GRUPOS DE CONTROL

Es conveniente adoptar un esquema especial para poder evaluar el impacto del algoritmo predictivo. Si adoptásemos una nueva puntuación y, al mismo tiempo, las campañas de ventas se tornaran muchísimo más promocionales para los clientes, gracias a una nueva oferta de *marketing*, podríamos estar adjudicándole a la minería de datos mucho más mérito del que se merece. Por otro lado, si el producto se volviese en general mucho menos atractivo para el mercado, por el avance de una competencia muy fuerte, podríamos menospreciar un pequeño incremento o, incluso, alarmarnos al enterarnos de que las nuevas campañas con minería de datos estarían dando peores resultados.

La única forma de medir apropiadamente el desempeño es a través de **grupos de control**. Si cada mes enviamos 1000 piezas a comercializar, el mes que comencemos a utilizar minería de datos, podemos enviar 500 con y 500 sin, asegurándonos de que el único sesgo entre un grupo y este esté dado por la puntuación. Obtendríamos, por ejemplo, los siguientes resultados:

	TASA DE EFECTIVIDAD POR MES			
SEGMENTACIÓN	1/15	2/15	3/15	4/15
HABITUAL	10%	9%	11%	15%
DATA MINING	X	X	X	30%
LIFT	X	X	X	2

Una vez aplicada la minería de datos para el mes 4/15, obtuvimos una tasa de efectividad del 30%, contra el 15% del grupo de control, duplicando así las ventas. El *lift* es simplemente el resultante de 30/15, cuyo valor correcto es 2 y no 3 (que hubiera resultado de promediar los últimos tres meses en lugar de usar grupos de control en el mes 4/15). Como ya señalamos, lo mismo puede suceder a la inversa:

	TASA DE EFECTIVIDAD POR MES			
SEGMENTACIÓN	1/15	2/15	3/15	4/15
HABITUAL	10%	9%	11%	5%
DATA MINING	X	X	X	10%
LIFT	X	X	X	2

Sin grupos de control, en este caso, podríamos pensar que nuestros algoritmos no sirvieron para nada, cuando en realidad salvaron a un mes pobre en el que, si

todo hubiese seguido como siempre, las ventas hubiesen caído a la mitad. El *lift* es exactamente el mismo.

Incluso, a medida que la minería de datos vaya ganando confiabilidad y se utilice cada vez más dentro de la empresa, sigue siendo necesario realizar, aunque sea, pequeños grupos de control para seguir demostrando la efectividad de la inteligencia comercial para segmentar. También, se pueden seguir utilizando para probar nuevos modelos que se vayan generando o retroalimentando con los resultados obtenidos.

AMPLIANDO A OTROS OBJETIVOS

Como hemos visto, una parte importante del proceso de minería de datos consiste en preparar una serie de variables de entrada que se utilizarán para predecir una variable de salida determinada. ¿Qué podemos tener como salida? Prácticamente lo que a usted se le ocurra. La técnica se ha utilizado en miles de campos e implementaciones distintas, la optimización de campañas de ventas es solamente un minúsculo ejemplo.

Los autos que conducen solos, por ejemplo, trabajan con minería de datos. Toman la información captada por los sensores colocados en el vehículo y a todo momento **predicen** qué es cada señal que reciben, no solamente formando una imagen tridimensional de sus alrededores sino, a la vez, también prediciendo cómo cambiará. ¿El peatón cruzará? ¿El auto que tengo delante frenará? ¿Puedo cambiar al carril izquierdo? Por supuesto, el sistema detrás de un automóvil inteligente tiene una complejidad imposible de describir en pocos párrafos, pero en el fondo **simplemente** se trata de varios algoritmos predictivos que, en su conjunto, le permiten a una computadora predictiva analizar su entorno y tomar decisiones mucho mejor que un conductor humano.

Volviendo al mundo empresarial y el CRM, uno de los objetivos más comunes también es la prevención de abandono. En lugar de predecir el resultado de una campaña, podemos predecir si un cliente rescindirá su contrato o no. Si pensamos en el sistema del auto inteligente, a nuestro CRM podemos incorporar una predicción de valor futuro de cada cliente. Este valor también tendrá peso a la hora de decidir en qué clientes debemos concentrarnos para retenerlos: quizá sea preferible esforzarse más en conservar a un cliente con la mitad de probabilidad de irse que otro cuya cuenta con nosotros sea diez veces menor.

AMPLIANDO A BIG DATA

Como habrá imaginado, en la realidad se suele trabajar con muchas más que unas diez variables predictivas y mil casos. En el mundo bancario, por ejemplo, es

usual contar con más de 500 predictoras posibles y cientos de miles de casos. En los últimos años ha cobrado relevancia el término **Big Data**, el cual hace referencia, en parte, a utilizar aún volúmenes de información más grandes que los que se venían usando hasta ahora para hacer minería de datos.

Si tomamos una base de datos bancaria, por ejemplo, podríamos primero contar con un puñado de datos demográficos generales y marcas de tenencias de productos: cuántas cuentas, tarjetas de crédito, seguros, inversiones, etcétera. Luego, podríamos agregarle información detallada de cada producto, contando con saldos de cuentas, consumos de tarjetas, montos de inversiones, primas de seguros y demás. ¿Y qué hay del detalle de los consumos de tarjetas? Es posible agrupar por rubros y obtener promedios. ¿Y si pensáramos en calcular también el promedio de cada variable, en su valor para cada uno de los últimos seis meses? ¿Y el máximo y el mínimo? Como puede apreciar, muy fácilmente la cantidad de variables predictivas se puede multiplicar varias veces, buscando cada una un poco más de valor en la gran montaña de datos de la cual disponemos.

El proceso general que hemos descripto se mantendrá más o menos igual. Naturalmente, la etapa inicial de comprensión y preparación de datos será cada vez más intrincada y extensa. Los tiempos de procesamiento de los algoritmos se extenderán. Pero las técnicas serán más o menos las mismas. La evaluación se hará exactamente igual y lo que podremos hacer es ver el incremento en el desempeño gracias al agregado de nuevas fuentes y cantidades de datos.

PALABRAS FINALES

El objetivo de esta introducción a la **minería de datos** se habrá cumplido no solamente si le ha servido para comprender de qué se trata la técnica, sino también si ha entendido su gran utilidad y la necesidad de aplicarla cuanto antes en su organización.

Si todavía no hace **Data Mining**, ¡no espere más! Intente, aunque sea, confeccionar un rudimentario Árbol de Decisión y cuando comience a ver cómo los algoritmos descubren información valiosa que usted quizá ni sabía que poseía, inmediatamente querrá explotar al máximo su base de datos y dejará atrás los análisis manuales del pasado para concentrarse en la automatización de la predicción del futuro.

BIBLIOGRAFÍA

Gareth, J. et al. *An Introduction to Statistical Learning*, Nueva York, Springer, 2013.

Hofmann, M.; Klinkenberg, R. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, CRC Press, 2014.

Kuhn, M.; Johnson, K. *Applied Predictive Modeling*, Nueva York: Springer, 2013.

Provost, F.; Fawcett, T. *Data Science for Business*, O'Reilly, 2013.

IVÁN GÓMEZ VILLAFÑE

Es Técnico en Comercialización de la Universidad de Belgrano, especializado en Minería de Datos/*Data Mining*. Desarrolló su *expertise* trabajando en *Data Mining* para Inteligencia Comercial del Banco Santander Río Sucursal Argentina, para el cual, entre otros trabajos, desarrolló varios modelos de prevención de churn y optimización de campañas de ventas. Uno de los modelos más complejos decidía qué producto venderle a cada cliente y a través de qué canal.

También trabajó con el Departamento de Recursos Humanos, analizando encuestas, usando *Text Mining* y realizando modelos de predicción de comportamiento de colaboradores.