

Powerhouse: Data Mining usando Teoría de la información

Marcelo R. Ferreyra
mferreyra@dataxplore.com.ar

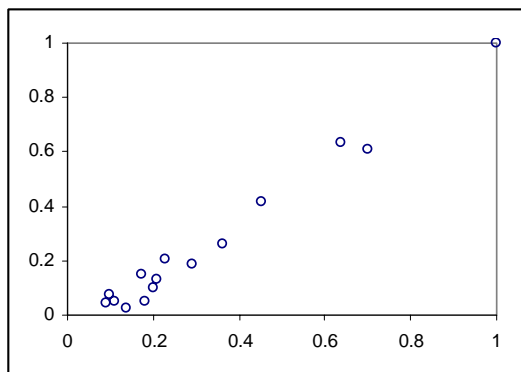
Cualquier proyecto de Data Mining que busque predecir una variable y/o explicar qué contienen los datos necesita resolver tres tareas muy importantes para lograr el éxito:

1. Preparar los datos
2. Seleccionar las variables más importantes
3. Construir un modelo simple de entender

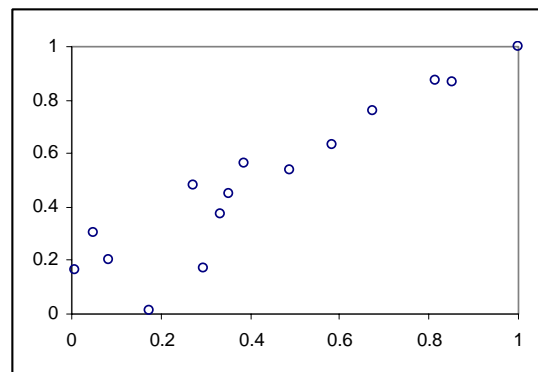
Preparación de datos

La preparación de datos es una tarea compleja y lleva mucho tiempo (se estima que podría llevar hasta un 70% del tiempo total del proyecto). La distribución de las variables numéricas podría necesitar ajustes, las variables categóricas podrían necesitar ser transformadas en numéricas o las numéricas en categóricas. Cualquiera de estas transformaciones debe realizarse evitando romper la estructura interna de los datos y tratando de exponer al máximo la información que cada variable podría contener.

Los siguientes ejemplos muestran dos transformaciones, la primera realizada sobre variables numéricas. En el gráfico de la izquierda se representa la relación con los valores originales de las variables. A la derecha se muestra la misma relación pero con las variables transformadas. Se puede notar que la transformación expandió los valores bajos y comprimió los altos, logrando una representación más uniforme de la relación entre estas dos variables.



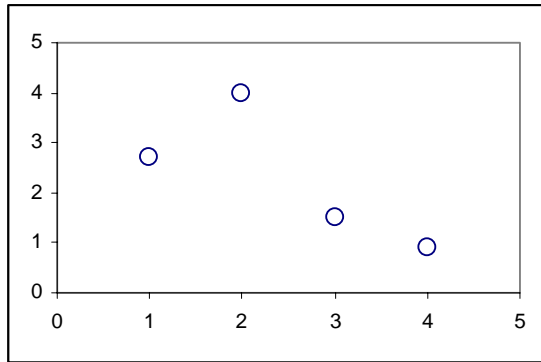
Variables sin transformar



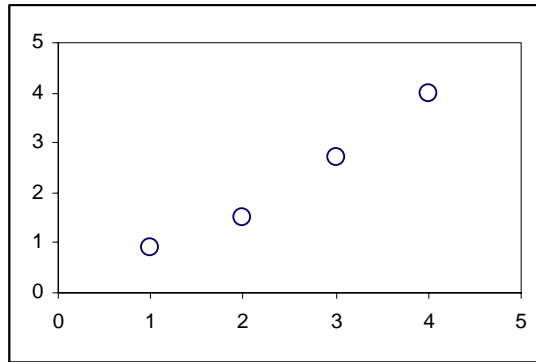
Variables transformadas

En el caso de las variables categóricas, es muy común asignarles valores numéricos a cada categoría en forma arbitraria. Esta práctica desarticula la posible relación haciéndola más compleja y más difícil de modelar. Una buena numeración de variables categóricas debería respetar las relaciones. A la izquierda aparece una relación entre una variable categórica y otra numérica. Las categorías están numeradas en forma arbitraria. A la derecha, la misma relación

pero con las categorías numeradas respetando la estructura interna de los datos. Es evidente que la relación representada a la derecha es más fácil de modelar que la de la izquierda



Variables transformadas en forma arbitraria



Variables transformadas respetando la estructura interna

Otro tema a considerar es qué hacer con los outliers y los valores nulos. Desechar los outliers y reemplazar los nulos con el valor promedio de la variable es una pésima decisión.

Por último, podría suceder que la variable a predecir sea dicotómica y una de las categorías esté muy poco representada. En ese caso podría ser necesario balancear los datos si es que el algoritmo para modelar lo requiere.

Cada uno de estos temas está excelentemente tratado en el libro *Data Preparation for Data Mining* de Dorian Pyle.

Selección de variables

Cuando existen muchas variables disponibles hay varias razones para seleccionar un grupo de variables. Cuando el modelo se ponga en producción, tener disponibles muchas variables podría ser demasiado costoso. Entender un modelo que usa muchas variables puede ser muy difícil y hasta imposible. Por último, cuando se usan muchas variables se necesitan muchos más ejemplos (filas) para que la muestra con la que se construye el modelo sea representativa de la población.

El inconveniente es que cuando hay muchas variables disponibles, seleccionar las *mejores* no es una tarea simple. Usar la fuerza bruta es imposible. Por ejemplo, con 10 variables disponibles, hay 1023 posibles combinaciones. Si existieran 100 variables y se deseara probar todas las combinaciones de 6 variables, se necesitarían hacer más de 1.000 millones de pruebas.

Por supuesto, existen distintos métodos que evitan la fuerza bruta, pero la mayoría de ellos no son óptimos.

Ya que no es posible encontrar la mejor selección de variables, al menos debería ser posible encontrar una selección óptima, que lleve la mayor cantidad posible de información que los datos puedan proporcionar y la menor cantidad de ruido posible. Además estos datos deberían ser representativos de la población.

Modelos simples de entender

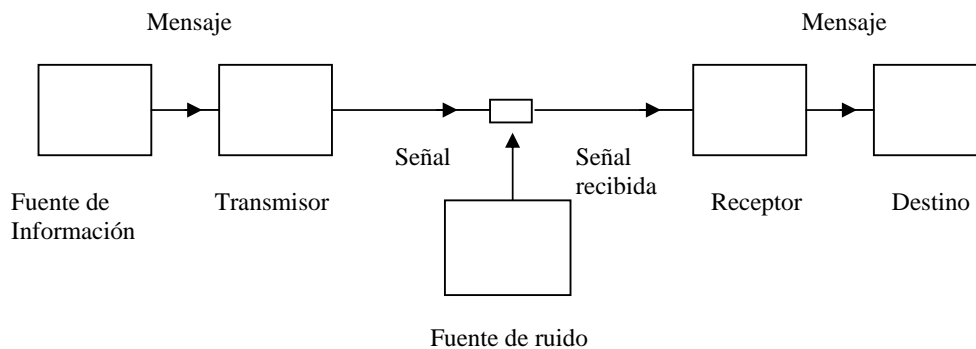
En muchas ocasiones es conveniente entender cómo funciona el modelo. En otros casos es absolutamente imprescindible que el modelo sea simple de entender, aún con el riesgo de perder exactitud y precisión. Si un gerente de marketing no entiende cómo funciona el modelo, le costará mucho confiar en sus predicciones. En ámbitos de negocio las cajas negras no tienen demasiado lugar.

Un modelo representa y resume las relaciones más importantes que contienen los datos. Una manera de confiar en un modelo es reconocer que las relaciones que estén representadas en el mismo tengan sentido. El problema es que a veces, cuando un modelo es simple de entender, su desempeño no es tan bueno. Por el contrario, un buen desempeño podría ser el resultado de un modelo muy difícil o imposible de entender.

La Teoría de la Información nos brinda una base con sólidos fundamentos matemáticos para resolver estas tres tareas de un modo simple, directo y rápido.

Introducción a la Teoría de la Información

La Teoría de la Información fue desarrollada en 1948 por Claude Shannon y nació como un modelo matemático para usar como base en el análisis de un sistema de comunicación como el siguiente



The Mathematical Theory of Communication, Shannon & Weaver, Pg. 7

Este esquema contiene una *fuentes de información* de la que se seleccionarán los *mensajes* a transmitir. El *transmisor codifica* el mensaje para convertirlo en una *señal* que será transmitida por medio de un *canal de comunicación*. La señal arribará al *receptor* intacta o algo cambiada si es que una fuente de ruido está presente. Finalmente se lo *decodifica* para convertirlo nuevamente en el mensaje que recibirá el *destinatario*.

Este sistema es lo suficientemente general como para adaptarlo a cualquier proceso de comunicación. Por ejemplo, la fuente de información podría ser el mercado de valores de Buenos Aires, el mensaje el precio del Merval, el transmisor una imprenta que transforma el precio en una serie de símbolos (números) a imprimir, el canal un diario, y el receptor la persona que lo lee y convierte los símbolos en el mensaje original. Una mancha de tinta podría ser la fuente de ruido que dificulta su lectura.

Shannon se ocupó de definir exactamente lo que significa *información* y cómo medirla y gracias a esto fue capaz de estimar la máxima cantidad de información libre de ruido que podía ser enviada a través de un determinado canal de comunicaciones. Además explicó cómo debían codificarse los mensajes para un mejor aprovechamiento del canal.

La información está relacionada con la cantidad de mensajes disponibles en la fuente de información y con la mayor o menor libertad de elección de los mismos. Por ejemplo, supongamos que nuestra fuente de información es el lanzamiento de una moneda. Hay dos mensajes distintos que puede entregarnos esta fuente: cara o cruz. Si la moneda no está cargada, ambos mensajes tienen igual probabilidad de aparecer y en este caso se dice que la fuente contiene 1 bit de información.

La información se mide por medio de la *entropía*. La entropía es una medida usada en termodinámica para medir el grado de desorden de un sistema. En Teoría de la Información, la entropía es usada para medir la información como una función del grado de libertad de elección de un mensaje y se define por medio de la siguiente ecuación

$$H = - \sum_i^n p_i * \log_2(p_i)$$

H representa la entropía, medida en bits, de una fuente de información con n mensajes. Cada mensaje tiene una probabilidad p de ser elegido. El signo menos es sólo para evitar que la entropía sea negativa, ya que los logaritmos de números menores que 1 son negativos. Calculemos la entropía asociada al lanzamiento de una moneda

Lanzamiento de una moneda			
Mensaje	p	$-\log_2(p)$	$-p * \log_2(p)$
Cara	0.5	1	0.5
Cruz	0.5	1	0.5

H	1
---	---

La entropía, que mide la cantidad de información presente en la fuente, es 1 bit. Ahora supongamos que la moneda está cargada y las probabilidades son 30% para Cara y 70% para Cruz. La entropía es ahora

Lanzamiento de una moneda			
Mensaje	p	$\log_2(p)$	$p * \log_2(p)$
Cara	0.3	1.737	0.521
Cruz	0.7	0.515	0.360

H	0.881
---	-------

En este caso, la fuente de información privilegia un mensaje sobre otro. La incertidumbre ha bajado. En el caso anterior no teníamos ninguna idea de cuál podía ser el mensaje que recibiríamos, ahora sabemos que es más probable que aparezca Cruz. Dicho de otro modo, si el mensaje es Cruz estaremos menos sorprendidos que si es Cara. Es más, si tuviéramos una moneda que siempre sale Cruz, la incertidumbre sería nula, como también lo serían la entropía y la información que recibiríamos de la fuente.

A menor probabilidad de que un mensaje aparezca, mayor es la cantidad de información que lleva. En el caso de la moneda cargada, la información que lleva cada mensaje en forma individual está dada por el logaritmo en base 2 de la probabilidad. El mensaje Cara lleva 1.737 bits de información, mientras que Cruz lleva sólo 0.515 bits. Dicho de otro modo, será una mayor sorpresa recibir Cara que Cruz.

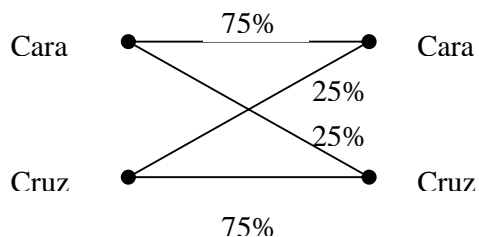
La información total que contiene la fuente puede ser interpretada como un promedio de la información que lleva cada mensaje ponderado por las probabilidades de cada uno de ellos.

La información que contiene el lanzamiento de un dado es mayor que la del lanzamiento de una moneda (en ambos casos consideramos monedas y dados no cargados). La razón es que existen más mensajes disponibles en un dado (6) que en una moneda (2). El lanzamiento de un dado contiene 2.585 bits de información.

Ruido y Equivocación

Un punto clave es cómo el ruido afecta la información que se transmite. Para analizarlo continuaremos con el lanzamiento de la moneda. Supongamos que el transmisor es una persona que ve el mensaje y lo codifica con una palabra (cara o cruz) que emitirá en voz lo suficientemente alta para superar el ruido de ambiente. A varios metros se encuentra el receptor que recibirá el mensaje.

Consideremos que el ruido es de tal nivel que un 25% de las veces el mensaje que arriba al destinatario es el equivocado. Un diagrama representa esta situación



Se presentan dos interrogantes:

1. Cuál es la probabilidad de que el mensaje enviado haya sido Cara dado que se haya recibido Cara
2. Cuál es la probabilidad de que el mensaje recibido haya sido Cara dado que se haya enviado Cara

Exactamente las mismas preguntas aparecerían si en vez de Cara hubiéramos usado Cruz. Lo importante no es el mensaje recibido o enviado, sino dado un mensaje cualquiera que se haya recibido, cuál será la probabilidad de que ese mismo mensaje haya sido enviado, y dado un mensaje enviado, cuál será la probabilidad de que ese mismo mensaje haya sido recibido. La respuesta tiene que ver con la entropía relativa entre los mensajes enviados y las señales recibidas, o en forma más general, entre dos conjuntos de señales o mensajes. Estos dos interrogantes planteado en forma general son:

1. Cuál es la incertidumbre del mensaje enviado cuando se conoce la señal recibida
2. Cuál es la incertidumbre de la señal recibida cuando se conoce el mensaje enviado

El primer caso lleva el nombre de *equivocación* y se representa como $H(X|Y)$ y el segundo caso es el *ruido* y se representa como $H(Y|X)$.

La información transmitida neta puede calcularse como la información enviada $H(X)$ menos la equivocación $H(X|Y)$, o como la información recibida $H(Y)$ menos el ruido $H(Y|X)$.

$$T = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

La Teoría de la Información como base del Data Mining

Ahora que tenemos algunos conceptos más claros acerca de la Teoría de la Información veremos cómo podemos usarla como marco de referencia para el análisis de datos y la creación de modelos.

Supongamos una compañía de seguros que desea ofrecer una póliza de seguro de casas rodantes a sus clientes. A fin de reducir los costos de envío de folletos, se desea enviar los mismos a aquellos clientes con mayor probabilidad de que compren la póliza. Para tal fin, se selecciona una muestra al azar de clientes con casas rodantes y se les envía la oferta. Luego de unas semanas se cuenta con la lista de clientes a los cuales se les envió el folleto y cuáles de ellos compró la póliza.

Considerando que se conocen muchos datos acerca de los clientes a los que se les envió la oferta, se supone que quizás haya muchas características demográficas y de su comportamiento pasado que podrían haber sido útiles para prever con cierto grado de exactitud quiénes comprarían y quiénes no.

Si consideramos esta situación como un sistema de comunicación, podríamos usar la Teoría de la Información para su análisis. Tratemos entonces de identificar las diferentes partes de la situación a fin de relacionarlas con las partes del sistema de comunicación.

Cada cliente tiene una serie de variables asociadas que toman distintos valores según cada caso.

	V1	V2	V3	V4	V5
Cliente 1					
Cliente 2					
Cliente 3					
..					
..					
..					
..					
Cliente n					

Algunos clientes tendrán valores iguales



Otro tendrán valores levemente diferentes



Y otros completamente distintos



En cualquier caso, podemos pensar que las diferentes combinaciones de valores de las variables representan una señal que envía la situación que estamos analizando. O sea, en el mundo real los clientes tienen distintos comportamientos. La combinación de estos comportamientos se codifican como una señal que se transmitirá a través de una canal. Así como en el primer ejemplo que dimos de un sistema de comunicación decíamos que el canal era el diario, en este caso el canal serán los datos.

Por otro lado, tenemos otra serie de datos con las respuestas a la oferta de la póliza de cada cliente y que representan otro conjunto de señales. Queremos averiguar cuánta información comparten ambos grupos de señales. O sea, queremos analizar la entropía relativa entre estos dos conjuntos de señales.

A las señales formadas por la combinación de variables les llamaremos señales de Entrada o simplemente X, porque ingresan al canal de comunicación. A las señales recibidas como respuesta de los clientes, las llamaremos señales de Salida o Y.

¿Cuánta información acerca de Y transmite X? Lo podemos averiguar de dos maneras

$$\begin{aligned} \text{Información transmitida} &= \text{Información de Entrada} - \text{Equivocación} \\ \text{Información transmitida} &= \text{Información de Salida} - \text{Ruido} \end{aligned}$$

O lo que es lo mismo

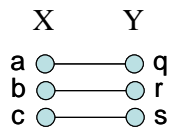
$$\text{Información transmitida} = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Ruido y Equivocación en los datos

Si bien ya hablamos del ruido y la equivocación, ahora analizaremos más detalladamente lo que significan en el contexto de los datos.

Dijimos que los datos de entrada transmiten información acerca de los datos de la salida. Los datos de entrada son los que en estadística se conocen como variables independientes y a los datos de salida como variable dependiente. La frase “las variables independientes tienen un cierto grado de correlación con la variable dependiente” es similar a decir que X e Y comparten información. Es importante notar que no estamos asumiendo ningún tipo especial de relación entre X e Y. O sea, cuando hablamos de información compartida entre dos variables, no nos referimos a correlaciones lineales o no lineales. Simplemente decimos que ambas variables llevan información en común. Esa información común es la que llamamos información transmitida entre ambas variables.

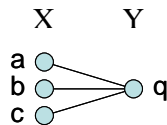
Existen cuatro casos en los que las variables podrían estar relacionadas (o compartiendo información). El primero es el más simple. Cada vez que se envía una misma señal de la entrada X, se recibe otra misma señal en la salida.



		Y		
		q	r	s
X	a	1.0		
	b		1.0	
	c			1.0

Por ejemplo, cada vez que el mensaje de X es a , siempre en Y es q , lo mismo con b/r y c/s . Este es un caso ideal en donde no hay dudas de cuál será la salida para esa entrada.

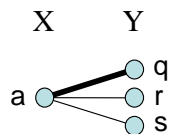
Pero las situaciones reales son algo más complejas. Los datos rara vez encierran relaciones tan claras. Así que también podemos tener otro caso al que llamamos Equivocación



		Y		
		q	r	s
X	a	1.0		
	b	1.0		
	c	1.0		

Varios mensajes distintos de entrada resultan en el mismo mensaje de salida. Varias voces distintas dicen la misma cosa.

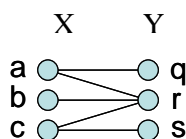
Otro caso es cuando un mismo mensaje de entrada algunas veces resulta en un mensaje de salida y otras veces en otro distinto.



		Y		
		q	r	s
X	a	0.8	0.15	0.05
	b			
	c			

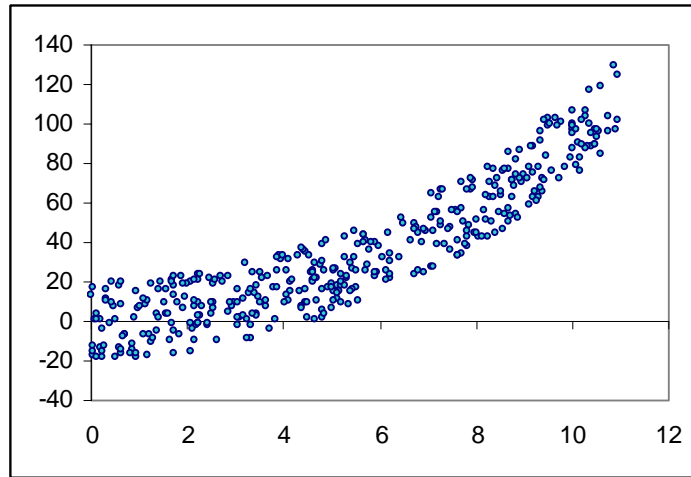
Esta es una situación desagradable porque frente a un mensaje de entrada no tenemos claro cuál podría ser el mensaje de Salida. A esto se le llama Ruido. En el caso representado arriba se supone que cada vez que se envía el mensaje a , se recibe q el 80% de las veces, r el 15% y s el 5%.

Con datos reales lo normal es que exista una mezcla de información, ruido y equivocación. Si bien la equivocación no nos preocupa, el ruido es lo que impide que un modelo tenga una buena exactitud.



		Y		
		q	r	s
X	a	0.5	0.5	
	b		1.0	
	c		0.5	0.5

Otra manera de ver el ruido es mediante el siguiente ejemplo



Esta relación contiene ruido porque para un mismo valor de X existen varios valores distintos de Y.

Midiendo la entropía de los datos.

Supongamos que queremos conocer la entropía de una variable categórica, por ejemplo, estado civil. Comenzamos obteniendo una tabla de distribuciones de las categorías y luego calculamos las probabilidades y los logaritmos, tal como lo hicimos para el caso de la moneda. Por ejemplo

Estado Civil				
Categoría	%	p	$-\log_2(p)$	$-p \cdot \log_2(p)$
Casado	46%	0.46	1.120	0.515
Soltero	28%	0.28	1.837	0.514
Separado	15%	0.15	2.737	0.411
Divorciado	11%	0.11	3.184	0.350

H	1.790
---	-------

Así que la variable Estado Civil lleva 1.79 bits de información. Esta información es acerca de sí misma, no acerca de la variable de salida. Para calcular la entropía de una variable numérica, primero hay que transformarla en categórica. A este proceso se le llama Binning y existen varias maneras de hacerlo. El más simple, aunque no el más recomendado, es separar la variable en rangos iguales. Por ejemplo, si una variable tiene como valor máximo 10 y mínimo 0, entonces se puede transformar en categórica asignando las siguientes categorías según el valor que toma la variable

Rango	Bin
≤ 2	B1
>2 y ≤ 4	B2
>4 y ≤ 6	B3
>6 y ≤ 8	B4
>8	B5

En este ejemplo se transformó la variable numérica en otra categórica con 5 categorías o bins.

Otro método más recomendable que el anterior es el que asigna la misma cantidad de casos (filas) a cada bin.

Los dos métodos anteriores son no supervisados. Existe un mejor método aún que trata de perder la menor cantidad de información posible. Cada vez que una variable numérica se transforma en categórica agrupando valores similares en una sola categoría, se pierde información, no sólo acerca de la variable en si misma por el hecho de reducir la cantidad de mensajes disponibles, sino que se pierde información acerca de la variable de salida. Para una misma cantidad de bins se puede perder mayor o menor información, dependiendo de cómo se elijan los puntos de corte.

Veamos un ejemplo, supongamos que la variable es la edad y la usaremos para transmitir información sobre si la persona está jubilada o no. Supongamos para no complicar la explicación, que deseamos usar sólo dos bins. Es probable que el mejor punto de corte sea alrededor de los 60 años. O sea, si transformamos la variable de la siguiente manera

Rango	Bin
≤ 60	B1
> 60	B2

perderemos menos información que si el punto de corte está en 40 años.

Pero si la variable a predecir es el estado civil, entonces el mejor punto de corte quizás es 30 años.

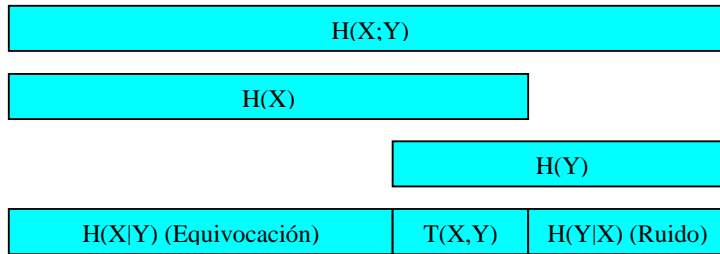
El método “Menor Pérdida de Información” o LIL (Least Information Loss) busca los mejores puntos de corte para maximizar la información acerca de la variable de salida. Este es un método supervisado porque depende de la variable a predecir.

Una vez que una variable numérica se transformó en categórica, se puede medir la información de la misma manera que con variables categóricas.

Para medir la información de un grupo de variables (categóricas y/o numéricas transformadas en categóricas), sólo hay que combinar los diferentes bins que toma cada variable del grupo para formar un nuevo Bin o mensaje, y luego se trabaja con esta nueva variable.

Edad	Sexo	Señal Combinada
B1	F	B1F
B3	M	B3M
B1	M	B1M
B5	M	B5M
B1	F	B1F
B2	F	B2F
B3	M	B3M

El siguiente gráfico nos guiará en el cálculo del ruido, la equivocación y la información transmitida.



$H(X;Y)$ representa la entropía de la Entrada y Salida combinadas y para calcularlas se procede como si fuera un conjunto de variables tal como vimos anteriormente. Del gráfico se puede averiguar cómo calcular el ruido o la equivocación

$$H(Y|X) = H(X;Y) - H(X)$$

$$H(X|Y) = H(X;Y) - H(Y)$$

Conociendo estos valores de los datos es posible anticipar qué tan bueno será el modelo que podamos construir. Por ejemplo, los siguientes ejemplos son para una misma relación a la que se le va agregando ruido.

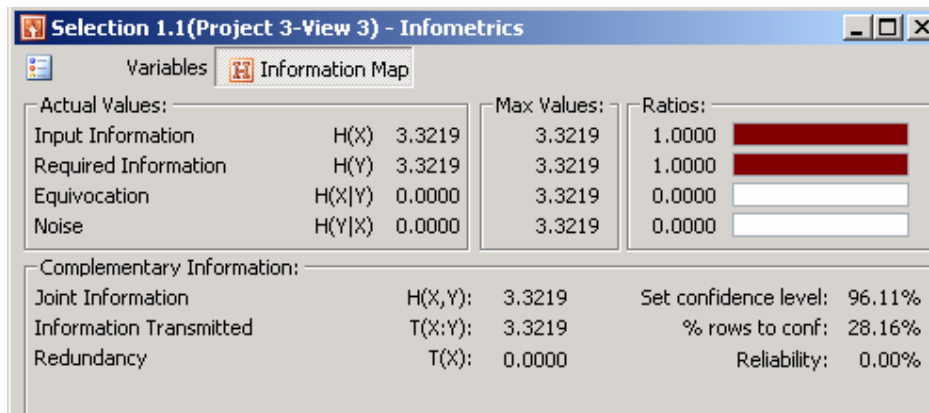
En el primer caso los datos no contienen ruido, solo la relación entre X e Y que es una parábola. La entropía de X y de Y es 3.3219 bits. O sea, se necesitan 3.3219 bits de información para predecir Y con total exactitud.

La equivocación es 0 bits, como también lo es el ruido, por lo que los datos parecen ser ideales para realizar un modelo.

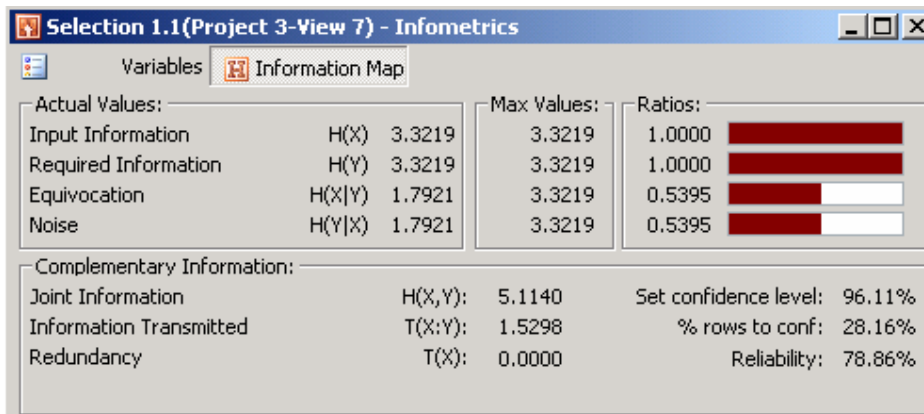
La información transmitida la podemos calcular como

$$T = H(Y) - H(Y|X) = 3.3219 \text{ bits} - 0 \text{ bits} = 3.3219 \text{ bits.}$$

O sea, los datos transmiten toda la información necesaria para predecir Y

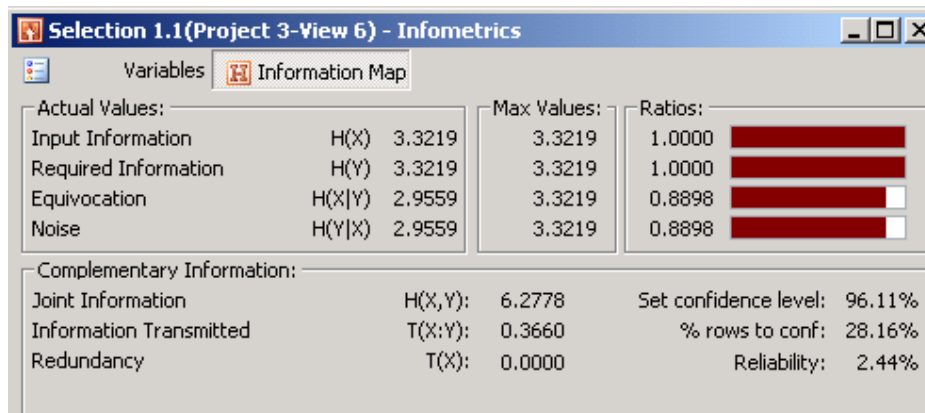


Ahora, si le agregamos algo de ruido, el análisis entrópico nos lo mostrará



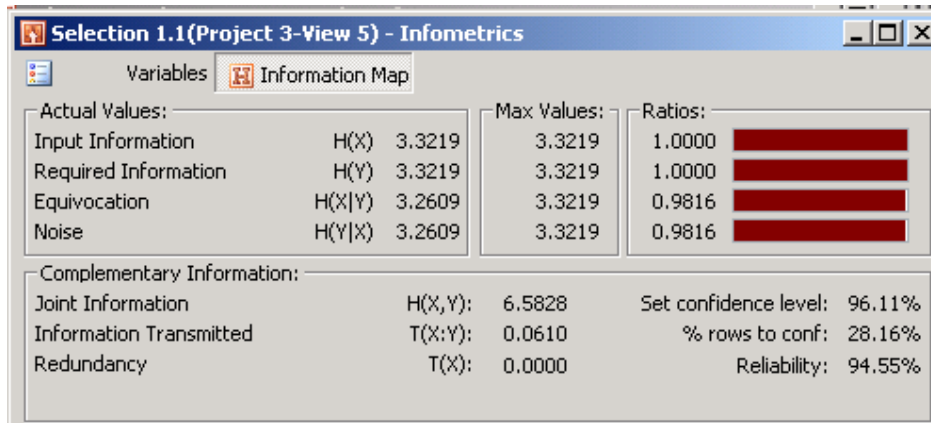
La cantidad de información requerida sigue siendo la misma, 3.3219 bits, pero ahora aparecieron el ruido y la equivocación. Tenemos 1.7921 bits de ruido lo que hace que la información transmitida sea 1.5298 bits, alrededor de un 46% de la información necesaria para predecir Y.

Ahora agregamos más ruido aún



El ruido agregado es de 2.9559 bits y de esta manera sólo se transmiten 0.3660 bits de información útil

Finalmente agregamos tanto ruido que la variable Y queda en su mayor parte enmascarada con este ruido

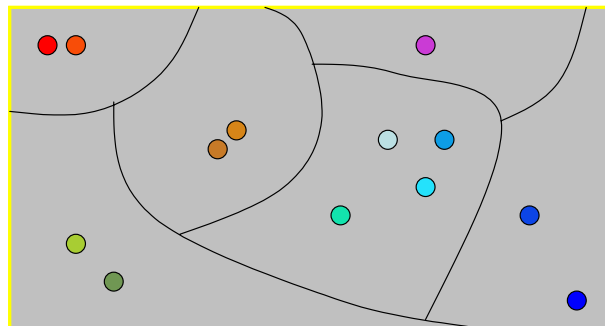


La información útil es casi nula, 0.0610 bits. No existirá ninguna herramienta que sea capaz de modelar estos datos ya que no contienen información útil acerca de la variable a predecir.

Modelos de predicción

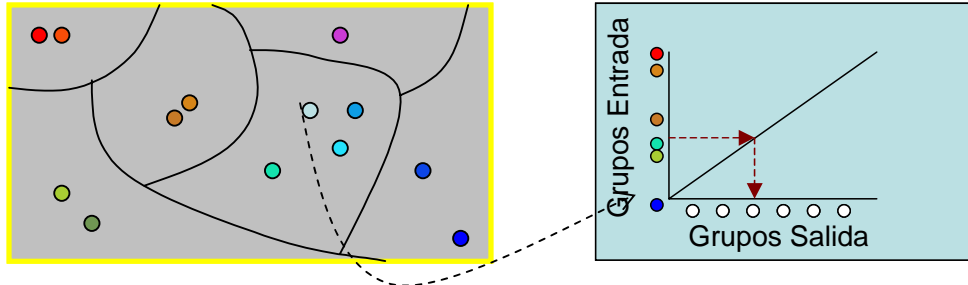
Si bien conocer la información útil contenida en los datos es muy importante, podemos avanzar un poco más y construir un modelo de predicción.

El primer paso será construir un mapa de información que contenga las señales de entrada. Cada señal estará posicionada en un espacio n-dimensional de acuerdo a la similitud que tenga con el resto de señales. Señales similares serán vecinas y las que no se les parezcan estarán más lejos.



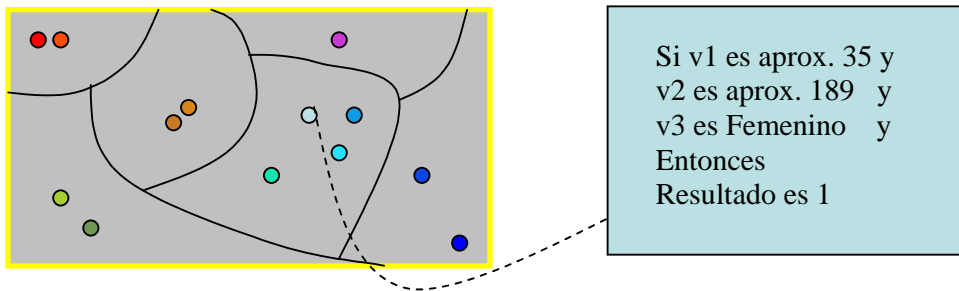
El gráfico muestra un espacio de sólo dos dimensiones pero en realidad el espacio tiene tantas dimensiones como variables de entrada existen.

Una vez agrupadas las señales muy similares, sólo restará relacionar estos grupos con las señales de salida. Esto lo podemos lograr de dos diferentes formas, una es encontrar una función de transferencia lineal que relacione ambas señales.



Mapeo de la información usando una Función de Transferencia Lineal

La otra manera es hacerlo mediante un grupo de reglas.



Mapeo de la información usando un conjunto de Reglas

Resumen de todo el proceso

Comenzamos con una tabla de datos que contiene variables que servirán para hacer predicciones respecto de otra variable. En estadística, las primeras son llamadas variables independientes y a la variable a predecir se la llama variable dependiente. En Teoría de la Información, las primeras son llamadas variables de entrada (Inputs) y la variable a predecir se llama variable de salida (Output).

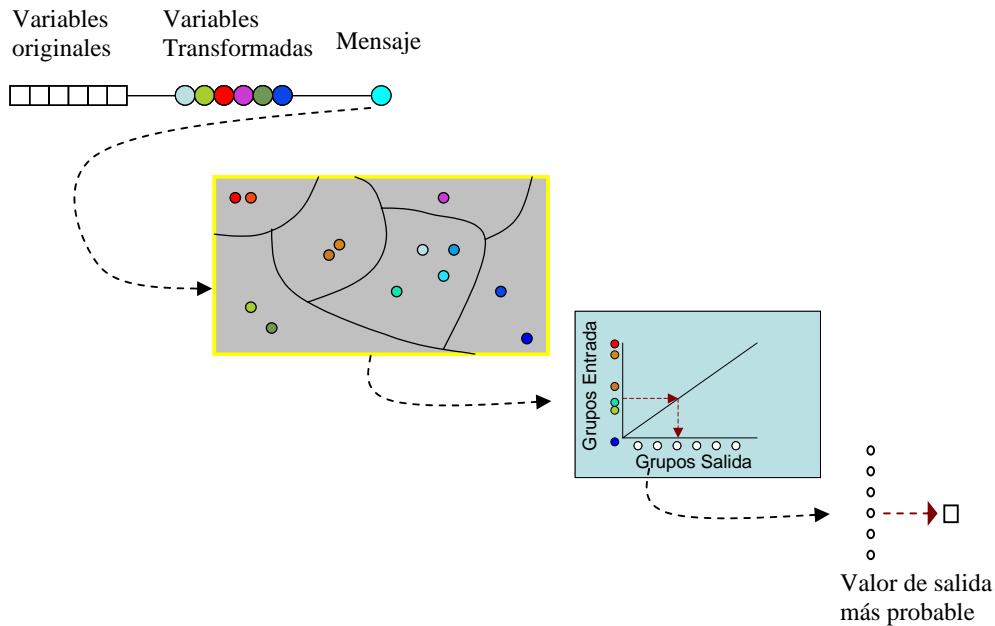
Las variables de entrada deben ser categóricas, por lo que si existen variables numéricas, primero deben ser transformadas usando algún método de Binning. La combinación de estas variables en cada caso (fila) de la tabla representa una señal o mensaje del sistema.

Cada uno de estos mensajes lleva información respecto de la variable de salida. Esta información se mide por medio de la entropía. Cuando el mismo mensaje es usado por distintos casos para relacionar distintos estados de la salida, decimos que existe ruido. El ruido es indeseable porque enmascara la señal y no nos permite obtener el mensaje correcto.

La información que se transmite desde la entrada a la salida se calcula como la información recibida menos el ruido. Esta información transmitida se puede usar para predecir los valores de la variable de salida. Una de las grandes ventajas de medir información es que podemos conocer de antemano si los datos con que contamos serán suficientes para predecir una variable de interés o no. Dicho de otro modo, si la información que se transmite de la entrada a la salida es muy baja, ninguna herramienta de modelado será capaz de hacer un buen trabajo. Si no existe

información, no hay manera de realizar un modelo, no importa si lo intentamos con una red neuronal, un árbol de decisión o un algoritmo asombroso y recién descubierto.

Es posible armar un mapa que vincule los mensajes o señal de entrada con los mensajes de la salida usando una función de transferencia lineal o un conjunto de reglas.



Cabe aclarar que si bien la función de transferencia es lineal, el modelo como un todo (Binning, mapa de información y función de transferencia) es capaz de tratar con relaciones lineales y no lineales.

Selección de variables

La ventaja de poder medir la información que lleva una o más variables sobre la variable a predecir es que se puede diseñar un método simple y directo para seleccionar un grupo de variables que lleve la mayor cantidad posible de información.

Los pasos son muy simples

1. Seleccionar la variable que mayor información contenga acerca de la variable a predecir
2. Seleccionar la siguiente variable con mayor información *adicional* acerca de la variable a predecir
3. Continuar con el paso 2 hasta que la cantidad de información que aporte la variable no justifique la pérdida de representatividad

El paso 3 tiene en cuenta que cuantas más variables son seleccionadas mayor deberá ser la cantidad de casos (filas) necesarios para que los datos sean representativos de la población.

Este algoritmo evita automáticamente la selección de variables correlacionadas linealmente o no, ya que si una variable estuviera correlacionada con alguna previamente elegida, no estaría aportando demasiada información adicional y no sería elegida.

Otra ventaja es que si una variable mantiene alguna interacción con algunas de las variables previamente elegidas, su inclusión aportará más información de la que aportaría por si sola, con lo que aumenta las probabilidades de ser elegida y deja la evidencia al analizar las información que se va ganado con cada variable seleccionada.

Ejemplos

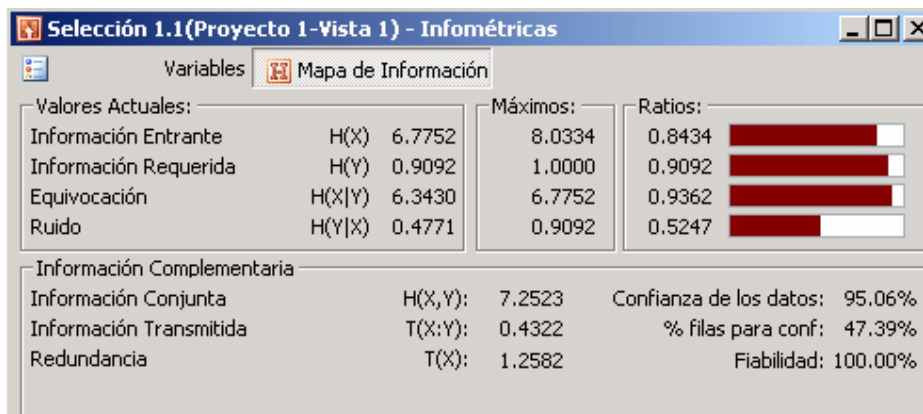
Los siguientes ejemplos fueron realizados usando Powerhouse, una herramienta de Data Mining basada en la Teoría de la Información.

Facultad

El primer ejemplo es muy simple. Se trata de una encuesta realizada a chicos que se reciben del colegio secundario y se les pregunta si su intención es asistir a la facultad el año siguiente. Se cuenta con 4 variables de entrada, el sexo, el ingreso de los padres, el coeficiente intelectual y si fueron o no motivados por sus padres.

El mapa de información muestra que las variables de entrada llevan un total de 6.7752 bits de información, pero 6.3430 bits es equivocación, o sea información duplicada que no agrega nada, así que la información útil es la diferencia entre estos valores, 0.4322 bits.

La entropía de la variable de salida muestra que se necesitan 0.9092 bits para predecirla con total exactitud, pero sólo tenemos disponibles 0.4322 bits, así que ya sabemos que el modelo no será perfecto.



Otra manera de ver la información disponible es considerando el ruido. Hay 0.4771 bits de ruido. Si se lo restamos a la información necesaria obtendremos nuevamente los 0.4322 bits disponibles para predecir si el alumno tiene pensado asistir a la facultad o no.

Realizamos un modelo con una función de transferencia lineal llamado Scorecard

SCORECARD Modelo 1(Proyecto 1-Vista 1-Selección 1.1) - Scorecard

Variable de Salida: Asistirá a la Fac... R2: 0,445
 Selección: Selección 1.1 Lift: 1,421
 Cant. Campos: 4 Eficiencia: 62,33%
 Rango de Score: 0-100

Card View

Variable	Val/Score	Val/Score	Val/Score	Val/Score	Val/Score	Si es nulo
Motivado por Pad...	No Motiv...	Motivado				
	0	28				15
Ingresos Padres	< 33105.0	< 49920.0	< 67940.0	< 74965.0	>=74965.0	
	0	10	30	39	51	13
IQ	< 88.5	< 100.5	< 110.5	< 111.5	>=111.5	
	0	5	11	16	19	9
Sexo	F	M				
	0	2				1

Este es un modelo que genera un score de 0 a 100 y se interpreta del siguiente modo:

Cada variable contribuye con parte del score total. La variable “Motivado por Padres” es la más importante y suma 28 puntos si los padres lo motivaron y 0 puntos si no lo hicieron.

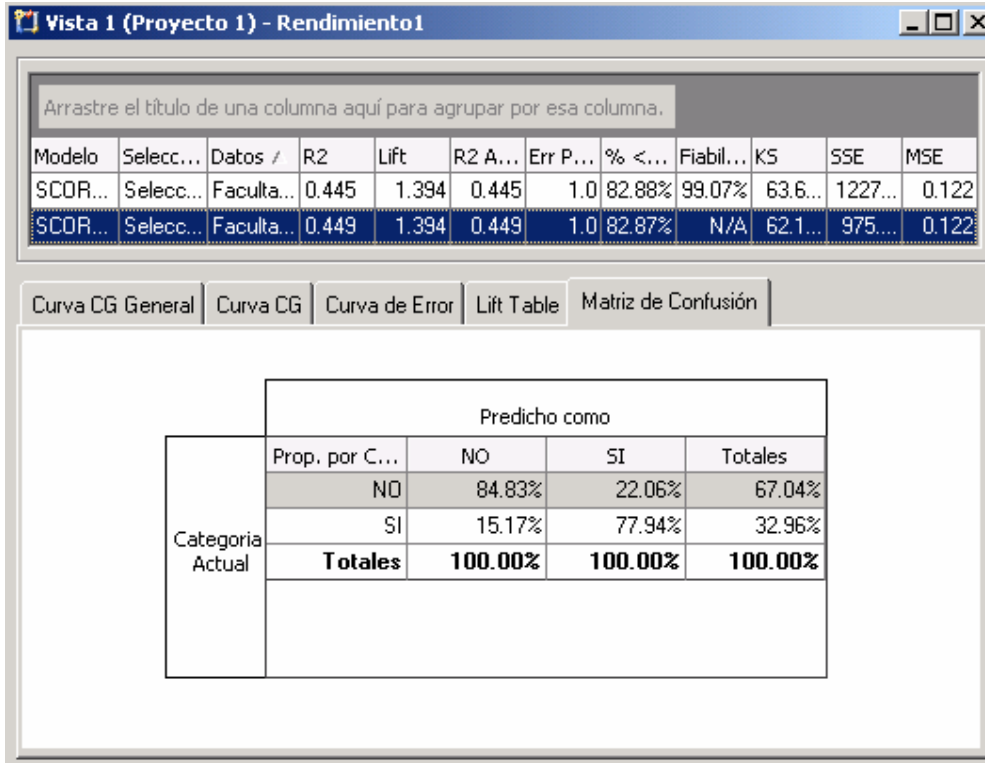
La variable “Ingreso Padres” suma diferentes puntajes según el rango de ingreso de los padres. Se puede notar que cuanto mayor sea el ingreso, mayor será el score.

Lo mismo sucede con el IQ y el Sexo.

La última columna, llamada “Si es nulo” contiene los puntos a sumar al score en caso de que esa variable para un alumno en particular no tenga ningún valor asociado.

Finalmente se suman los scores de cada una de las variables y se obtiene el score final. Cada score tiene asociada la probabilidad de que el alumno tenga pensado asistir a la facultad. A mayor score, mayor probabilidad de asistir.

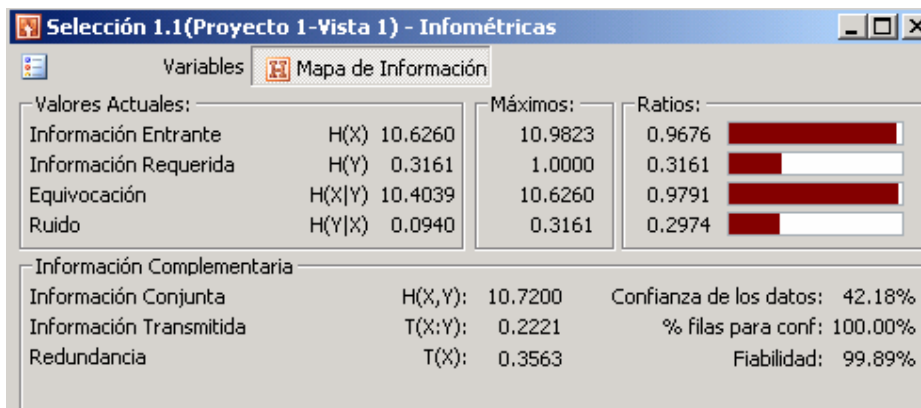
Existen varias medidas para evaluar el rendimiento del modelo tanto con los datos de entrenamiento como con los de prueba, tal como se puede apreciar en la siguiente ventana. El valor Lift está calculado en base a la información que transmite el modelo. Dicho de otro modo, el modelo representa los datos, por lo tanto transmite información. Si el Lift fuera 1, el modelo no estaría aportando nada para predecir Y.



Compañía de seguros

El siguiente ejemplo viene de una competencia de Data Mining en la que hay que armar una lista con 800 clientes que tengan la mayor probabilidad de comprar una póliza de seguros de una casa rodante. Los datos contienen 5800 clientes con 85 variables independientes (entrada) y una variable dependiente (salida). Existe además una tabla con otros 4000 clientes sobre los cuales se debe aplicar el modelo para seleccionar los 800 clientes con mayores posibilidades de comprar la póliza.

Comenzamos cargando los 5800 clientes y dejando que Powerhouse separe una muestra del 60% para obtener el modelo y dejando el 40% restante para la prueba final sobre datos nunca vistos. El paso siguiente es seleccionar las variables más importantes y al mismo tiempo armar el mapa de información



Se necesitan 0.3161 bits para predecir si un cliente comprará la póliza o no y las variables seleccionadas aportan 0.2221 bits, o sea alrededor de un 70% de lo necesario, lo que indica que un modelo realizado con estos datos podría tener un buen rendimiento.

Las variables seleccionadas fueron 5 y son listadas en la siguiente ventana que muestra además otros detalles. Sólo nos concentraremos en los nombres de las variables y la ganancia que aportan.

Nro.	Nombre	Tipo	H(X)	Hmax...	Relac...	H(Y x)	H(Y X)	Gana...	Conf.	Bal
47	PPERSAUT	Numér...	1.4456	2.0000	0.7228	0.9351	0.9351	6.49%	96.58%	0.0627
59	PBRAND	Numér...	2.1801	3.0000	0.7267	0.9406	0.8827	11.73%	95.48%	0.1120
5	MOSHOO...	Numér...	2.9925	3.3219	0.9008	0.9720	0.8009	19.91%	94.31%	0.1877
31	MHKOOP	Numér...	3.2673	3.3219	0.9836	0.9847	0.6094	39.06%	74.76%	0.2920
15	MFWEKIND	Numér...	3.0033	3.3219	0.9041	0.9909	0.2974	70.26%	42.18%	0.2964

La ganancia total es de 70.26% como lo indica el valor final de lo que se va ganando con la inclusión de cada variable. La primera variable, PPERSAUT es la que mayor información aporta en forma individual (6.49%), pero vemos por ejemplo, que la última variable (que debería aportar menos de 6.49%) aporta 31.20% (70.26-39.06). ¿Por qué? La explicación está en las interacciones. Esta variable está interactuando con una o más variables seleccionadas previamente.

Si se construye un modelo que usa una función de transferencia lineal llamado OPFIT (es similar al modelo Scorecard), se obtiene el siguiente rendimiento

Modelo	Selección	Datos /	R2	Lift	R2 A...	Err P...	% <...	Fiabil...	KS	SSE	MSE
OPFIT...	Selecc...	carava...	0.053	1.053	0.052	1.0	91.25%	86.39%	40.2...	178...	0.051
OPFIT...	Selecc...	carava...	0.061	1.061	0.059	1.0	91.09%	N/A	42.1...	130.33	0.056

Decil	# Casos	Random Prob...	Model Probab.	Lift
1	232	10%	33.11%	3.31
2	232	20%	50.68%	2.53
3	232	30%	64.86%	2.16
4	232	40%	79.05%	1.98
5	232	50%	88.51%	1.77
6	232	60%	89.86%	1.50
7	232	70%	92.57%	1.32
8	232	80%	97.97%	1.22
9	232	90%	99.32%	1.10
10	235	100%	100.00%	1.00

La tabla de Lift (en este caso el Lift se calcula en forma normal) muestra que seleccionando el 20% de los clientes con mayor score se consigue reunir 50.68% de clientes que compraron la póliza dando un lift de $50.68/20 = 2.53$

Para mostrar que es posible armar un modelo que utilice reglas en vez de una función de transferencia lineal, se construye un modelo MAXIT. El mismo encuentra 21 reglas y la siguiente ventana muestra alguna de ellas sólo como ejemplo

Reglas	Campo	Valor	Score
Regla 1 - Cobertura: 8.23%			
	PPERSAUT	6.5	
	PBRAND	4.5	
	MOSHOOFD	8.5	
	MHKODP	7.5	
	MFWEKIND	4.5	
	CARAVAN	1	0.0799
Regla 2 - Cobertura: 7.57%			
	PPERSAUT	6.5	
	PBRAND	1.5	
	MOSHOOFD	5.5	
	MHKODP	1.5	
	MFWEKIND	4.5	
	CARAVAN	1	0.0717
Regla 3 - Cobertura: 6.49%			
	PPERSAUT	2.5	
	PBRAND	0.5	

La regla 1, que cubre un 8.23% de los casos, dice que si la variable PPERSUAT tiene un valor aproximado de 6.5 y PBRAND un valor aproximado de 4.5 y, etc, etc, entonces el score es 0.0799. el resto de las reglas se interpreta del mismo modo. Este modelo tiene un rendimiento similar al modelo OPFIT según puede verse en la ventana de rendimientos

Modelo	Selecc...	Datos /	R2	Lift	R2 A...	Err P...	% <...	Fiabil...	KS	SSE	MSE
MAXIT...	Selecc...	carava...	0.042	1.06	0.041	1.0	70.71%	63.08%	35.8...	180...	0.052
MAXIT...	Selecc...	carava...	0.066	1.063	0.065	1.0	73.05%	N/A	39.1...	130...	0.056

Decil	# Casos	Random Probab...	Model Probab.	Lift
1	232	10%	36.49%	3.65
2	232	20%	50.68%	2.53
3	232	30%	58.78%	1.96
4	232	40%	75.00%	1.88
5	232	50%	79.05%	1.58
6	232	60%	83.11%	1.39
7	232	70%	88.51%	1.26
8	232	80%	93.24%	1.17
9	232	90%	99.32%	1.10
10	235	100%	100.00%	1.00

Conclusiones

En este documento hemos intentado mostrar algunas ventajas de analizar datos y construir modelos en base a la Teoría de la Información. Con este enfoque en vez de asumir que los datos contienen información, podemos medir la cantidad de información contenida. Este enfoque evita que naveguemos a ciegas en un mar de datos.

Los tres problemas planteados al principio, la preparación de los datos, la selección de variables y la creación de modelos eficientes y simples de entender, pueden resolverse usando algoritmos basados en la Teoría de la Información, tal como lo explicamos y lo demostramos con dos ejemplos.

Para conocer más detalles de la teoría, los métodos y las ventajas se puede consultar la siguiente información:

Pyle, Dorian, *Data Modeling and Data Mining*, Morgan Kaufmann, 2003

Pyle, Dorian, *Data Preparation for Data Mining*, Morgan Kaufmann, 1999

Pierce, John R., *An Introduction to Information Theory*, Dover, 1980

MacKay, David J.C., *Information Theory, Inferences, and Learning Algorithms*, 2001

Shannon, Claude y Weaver, Warren, *The Mathematical Theory of Communication*, University of Illinois Press, 1998

Documentos y artículos del sitio web www.dataxplore.com.ar y del Blog www.powerhousedm.blogspot.com