ANÁLISIS DE LA EFICACIA DE LOS MODELOS DE ÁRBOLES DE DECISIÓN PARA LA GESTIÓN DE POLÍTICAS PROMOCIONALES CENTRADAS EN LA SEGMENTACIÓN DE CLIENTES

Raquel Flórez López

RESUMEN

La segmentación de clientes resulta necesaria para la adecuada gestión de las políticas promocionales que, como el mailing, tienen como objetivo la captación de clientes ante el lanzamiento de un nuevo producto o servicio. Ahora bien, en el entorno competitivo actual, las cambiantes características sociodemográficas de los consumidores y la especialización de la oferta y demanda, obligan a la dirección de la empresa a gestionar de forma dinámica a sus clientes, al objeto de mejorar su cifra de resultados y conseguir una cuota de mercado superior a la competencia.

En este sentido, el análisis de bases de datos sobre la eficacia de mailings pasados suele proporcionar una información estratégica para el diseño de futuras políticas, siendo habitual el empleo de técnicas predictivas a posteriori basadas en modelos estadísticos.

No obstante, la creciente acumulación de información y la presencia de múltiples variables caracterizadoras de cada individuo, disminuye la eficacia real de estos modelos, que suelen ser incapaces de explicar de forma razonable los factores determinantes de las decisiones que proponen.

Ante esta situación, el empleo de la técnica de Aprendizaje-Máquina conocida como "árboles de decisión", permite conseguir un óptimo desempeño junto con una elevada interpretabilidad del sistema, facilitando la toma de decisiones de la gestión, tal como se contrasta en el presente trabajo respecto a una base de datos de clientes de una entidad aseguradora.

PALABRAS CLAVES: Marketing-mix, gestión de clientes, árboles de decisión, políticas promocionales

ABSTRACT

Clustering clients is necessary for the appropriate management of promotional decisions in the firm, such as *mailing*, which objective is the reception of clients facing the launching of a new product or service. In this moment, the current competitive environment, the socio-demographic characteristics of consumers, together to the specialization of sellers and consumers, force the company to get a dynamic management of clients, in order to improve its profit and to get a higher market share than competitors.

This way, the analysis of databases related to the effectiveness of last mailings, uses to provide strategic information for designing future politics, through the development of predictive "a posteriori" techniques based on statistical models.

Nevertheless, the each time higher accumulation of information, together to the existence of multiple attributes for each individual, reduce the real effectiveness of these models, which are usually unable to explain reasonably the decisive factors of the proposed decisions.

Facing that, the employment of the technique of Machine-Learning called "decision trees", allows to get a good performance together to a high interpretability of the system, making easier the decisions of managers, just as we analyse in this paper, regarding a database of clients from an insurance company.

KEYWORDS: Marketing-mix, customer targeting, decision trees, promotional decisions.

1. LA POLÍTICA PROMOCIONAL Y SU GESTIÓN EN LA EMPRESA

En los últimos años, la creciente globalización de los mercados, así como el incremento de la competencia empresarial y la aparición de nueva formas de aproximación a los clientes no basadas en el trato directo (Internet, venta por catálogo, venta por teléfono etc.), ha generado un cambio significativo en el comportamiento del consumidor, resultando éste mucho más volátil y difícil de predecir que en el pasado.

Por otro lado, los patrones sociodemográficos se modifican cada vez más rápido, lo que fuerza a las organizaciones a la gestión dinámica de sus clientes, a fin de reconocer subpoblaciones de consumidores con patrones de comportamiento similares y que, de esta forma, respondan en la misma dirección ante la presentación de patrones promocionales específicos (Kim et al., 2000).

Ante la reducida área de atracción del negocio físico y la especialización de oferta y demanda, el gestor de empresa debe centrar su atención en incrementar la eficacia de la comunicación con los clientes, al objeto de ofrecerle productos y servicios adecuados específicamente a sus necesidades.

Ahora bien, el análisis de la literatura en la materia permite observar la inexistencia de un consenso generalizado acerca de la segmentación óptima de clientes (Brijs, 2002, p. 195). Una de las razones principales de esta diversidad de opiniones radica en las distintas definiciones realizadas para el término "segmentación de mercados". Así, desde el trabajo pionero de Smith (1956), se han propuesto diversos enunciados alternativos, centrados en el análisis de la segmentación de mercados bien como *estrategia*, o bien como *metodología*; mientras que la primera se centra en el emparejamiento de los productos ofertados respecto a las necesidades de los individuos, el segundo presta más atención a las técnicas y métodos utilizados.

Con carácter general, el término puede resumirse como el proceso que tiene por objetivo la partición del mercado potencial en subgrupos de clientes heterogéneos entre sí y homogéneos respecto a sus componentes, dando lugar a la identificación de clusters de consumidores que responden de forma diferente ante las estrategias del marketing-mix de la empresa.

Así, entre los beneficios derivados de la implementación de políticas de segmentación de la clientela cabe destacar la localización eficiente de los recursos de marketing, facilitando la oferta de productos y/o servicios especializados y adaptados a las necesidades de cada conjunto de individuos.

Centrando el análisis en la segmentación de mercados como *metodología*, la gestión intensiva de bases de datos de clientes se ha convertido en una de las herramientas más poderosas de las organizaciones, que utilizan la experiencia pasada para identificar el público objetivo deseado, especialmente respecto a la gestión de instrumentos promocionales como el *mailing* o la orientación de procesos de venta telefónica.
Por lo que respecta a las variables utilizadas para la formación de clusters, éstas suelen ser de carácter demográfico, comportamental, económico y psicográfico, si bien su elección concreta depende del negocio

analizado. A su vez, tales atributos pueden ordenarse atendiendo a dos bases de clasificación (Wedel y Kamakura, 2000):

- Variables genéricas vs. específicas de producto, según sean o no independientes de los productos, servicios y otras circunstancias relacionadas con la venta.

- Variables observables y no observables, según puedan ser medidas directamente, o deban ser inferidas (respectivamente).

Tabla 1. Distintas bases de clasificación de atributos

	GENERAL	**ESPECÍFICA**
OBSERVABLE	Variables culturales, demográficas, geográficas y socioculturales	Atributos relacionados con la frecuencia de uso, lealtad a la marca, patronazgo, situaciones de uso y momento de compra
NO OBSERVABLE	Variables psicográficas, valores personales, personalidad y estilo de vida	Variables psicográficas, económicas, percepciones, elasticidades, preferencias, intenciones

Fuente: Wedel y Kamakura (2000).

Por lo que respecta al análisis de esta serie de datos, los métodos de segmentación se han centrado tradicionalmente en técnicas estadísticas, si bien en los últimos años se están incorporando nuevas herramientas basadas en métodos bayesianos y en paradigma de Inteligencia Artificial (árboles de decisión, redes neuronales artificiales, algoritmos genéticos, etc.)

Ahora bien, a pesar de la variedad de técnicas existentes, resulta posible su clasificación entorno a dos dimensiones fundamentales: (1) técnicas a priori vs. a posteriori; y (2) técnicas predictivas vs. descriptivas (Tabla 2).

Tabla 2. Clasificación de las técnicas de segmentación

	A PRIORI	**A POSTERIORI**
DESCRIPTIVA	Tablas de contingencia, modelos log-lineales	Cluster jerárquico, Cluster óptimo, modelos de clusterización basados en clases latentes
PREDICTIVA	Tabulación cruzada, regresión, análisis discriminante	Técnicas de Aprendizaje-Máquina, modelos de regresión basados en clases latentes, análisis discriminante

Fuente: Adaptado de Wedel y Kamakura (2000).

Así, mientras que los métodos descriptivos analizan la segmentación entre un conjunto de variables segmentación, sin distinguir aquellas que operan como dependientes o independientes, los métodos predictivos estudian la relación entre una variable explicada y un conjunto de variables explicativas.

Por otra parte, mientras en la segmentación a priori las variables utilizadas son definidas por el gestor de empresa, estableciéndose los segmentos según la heurística aplicada o a partir de la experiencia acumulada, en los métodos a posteriori se emplean técnicas estadísticas para obtener los subgrupos de clientes con características similares, a partir de un conjunto de variables de segmentación.

Así, los métodos predictivos a posteriori permiten establecer las categorías de clientes a partir de la información implícita en una base de datos, siendo el propio algoritmo el que establece las relaciones entre las variables analizadas. De esta forma, no resulta necesario considerar reglas de negocios o percepciones previas de los gestores, facilitando la gestión dinámica de la clientela y la adaptación al entorno de la toma de decisiones.

Si bien tradicionalmente las técnicas estadísticas han dominado esta categoría de métodos, la creciente complejidad del entorno empresarial y el abaratamiento de los sistemas de almacenamiento de datos ha llevado a la acumulación de una gran cantidad de información compleja, de forma que cada cliente representa un registro que puede contener decenas o incluso centenas de variables caracterizadoras. Este exceso de información se traduce en dos problemas fundamentales, a considerar por el gestor:

- ✓ La dificultad para determinar las variables relevantes para el análisis, dificultando la interpretación de las decisiones tomadas por el modelo construido.

- ✓ La presencia de una elevada multicolinealidad entre las variables explicativas, con el consiguiente incumplimiento de las hipótesis de partida de los principales métodos estadísticos.

Ante esta situación, se plantea la necesidad de desarrollar un método de análisis que, evitando las restrictivas hipótesis de partida de los modelos estadísticos, permita caracterizar de forma óptima a las distintas categorías de consumidores, desarrollando reglas de decisión fácilmente comprensibles para el experto humano.

En este punto, los árboles de decisión representan una alternativa muy atractiva para la clusterización de consumidores, alcanzando resultados superiores a los de las técnicas tradicionales con un reducido grado de complejidad.

De esta forma, el presente artículo se centra en el estudio de un problema de segmentación de clientes a partir de variables socioeconómicas y psicográficas, mediante el empleo de técnicas predictivas a posteriori. Para ello, se considerará la ejecución de modelos estadísticos tradicionales y de diversos árboles de decisión, comparando el desempeño obtenido y la complejidad de la toma de decisiones subyacente.

2. ANÁLISIS METODOLÓGICO: LOS ÁRBOLES DE DECISIÓN

2.1. CONCEPTOS BÁSICOS DE ÁRBOLES DE DECISIÓN

Los árboles de decisión constituyen estructuras de clasificación jerárquica que llevan a cabo de forma recursiva distintas particiones respecto a un mismo conjunto de datos. En su forma más simple, cada árbol presenta tres componentes diferenciados: nodos de decisión, ramas y nodos terminales u hojas.

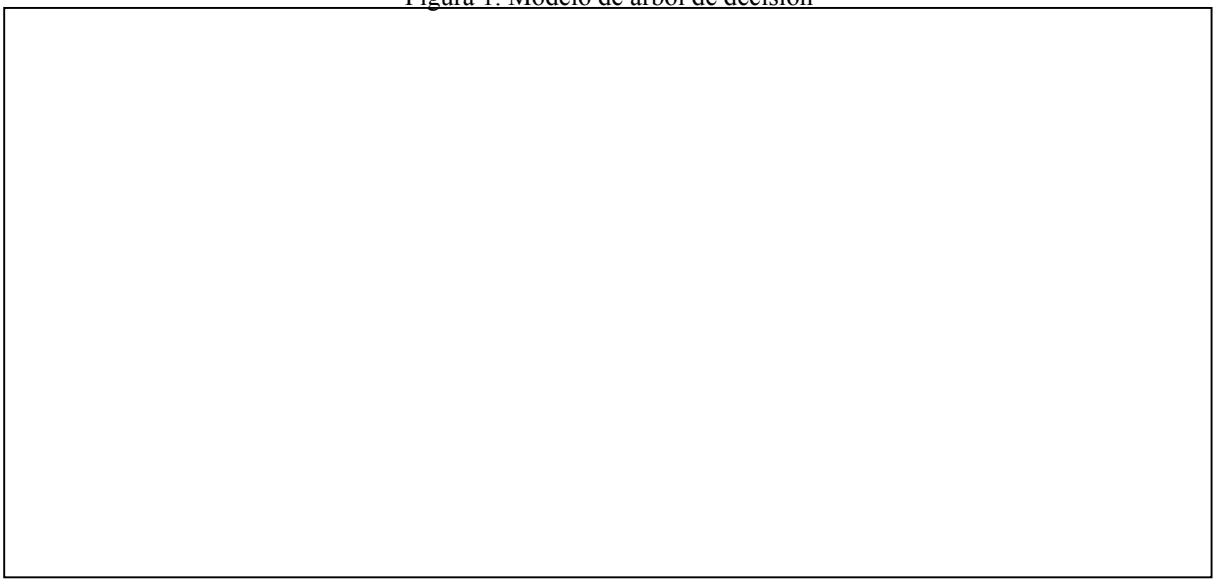
- ✓ Cada *nodo de decisión* desarrolla un test lógico relativo a los atributos presentes en los datos, al objeto de dividir éstos de forma no ambigua en dos o más categorías. Así, una vez que se lleva a cabo el test, se procede a la redistribución de los individuos desde el nodo considerado (“padre”) hacia dos o más nodos secundarios (“hijos”).

- ✓ Cada *rama* lleva a cabo la conexión entre un nodo "padre" y uno de sus nodos "hijo", correspondiente a una subdivisión diferente procedente del test previo.

- ✓ Cada nodo terminal u *hoja* corresponde a un nodo "hijo" asociado a una determinada clase y que no da lugar a nuevas particiones (es decir, es secundario respecto al nodo inicial o "raíz" del árbol). Para ser viable, un árbol de decisión debe contener cero o más nodos de decisión y uno o más nodos terminales.

La Figura 1 resume un esquema de árbol de clasificación.

Figura 1. Modelo de árbol de decisión



El proceso conocido como "inducción de árboles de decisión" (IDT) corresponde a la construcción de estos modelos a partir de un determinado grupo de datos. Para ello, la mayoría de las técnicas desarrolladas hasta la actualidad emplean la estrategia "divide y vencerás" de forma que, a partir de un conjunto inicial de ejemplos de entrenamiento, (□) llevan a cabo particiones recursivas de los casos hasta que todos los individuos de cada nodo pertenecen a la misma clase o bien no se pueden conseguir particiones más eficaces.

Los algoritmos para la construcción de árboles de decisión han experimentado una rápida evolución desde la aparición de los Sistemas de Aprendizaje Conceptual o "*Concept Learning Systems*" (Hunt, 1962, 1966), considerados pioneros en este campo.

La sucesiva evolución de la Teoría de Árboles de Decisión ha contribuido a que, en el momento actual, este paradigma sea uno de los más utilizados en el ámbito de la Minería de Datos ("*Data Mining*"), existiendo diversas variantes tanto respecto a la forma de dividir los datos presentados (particiones univariantes versus oblicuas) como a la forma de evitar el sobreajuste de los datos ("*overfitting*"), que lleva a la estimación de errores inferiores a los reales. Así, respecto a este último problema, suele ser habitual el empleo de técnicas de

reducción de tamaño o *técnicas de poda* que, a partir de un árbol inicial, llevan a cabo la eliminación de ramas no vitales.

Si bien existen distintos algoritmos desarrollados para la construcción de árboles de decisión (ID3, ChAID, ASSISTANT, C4.5, See5, etc), cabe destacar particularmente el modelo CART *("Classification and Regression Trees*") desarrollado por Breiman et al. (1984), que constituye una de las propuestas más eficaces y potentes en este ámbito.

2.2. CART ("CLASSIFICATION AND REGRESIÓN TREES")

El algoritmo CART, propuesto inicialmente por Breiman et al. (1984), constituye una de los métodos de construcción de árboles de decisión más aplicados en la práctica, que permite la formación de grafos útiles para la caracterización de múltiples problemas en entornos diversos: medicina, economía, finanzas, etc.

Este algoritmo se basa en un proceso de construcción de árboles de decisión en dos fases; así, mientras en la primera fase el proceso constructivo se lleva a cabo a través de un proceso de "arriba-abajo" ("top-down"), a través de la partición recursiva de los nodos del árbol mediante reglas de "split", en una segunda fase se procede a realizar un análisis "abajo-arriba" ("down-top"), al objeto de simplificar al máximo el modelo inicial, mejorando tanto su capacidad de generalización como su comprensibilidad. Este proceso, conocido como "poda de el árbol", fue introducido inicialmente por estos autores, si bien posteriormente se ha extendido a otros algoritmos de inducción (C4.5, etc.)

Debido a este entrenamiento en dos fases y, particularmente, a la fase de poda, el algoritmo CART emplea dos subconjuntos diferenciados de datos: el subconjunto de aprendizaje ("*training set*"), empleado en el proceso constructivos inicial, y el subconjunto de test ("*test set*"), usado para la simplificación del grafo, al objeto de estimar con mayor precisión los errores cometidos y, de esta forma, evitar procesos de sobreaprendizaje.

Por lo que respecta al criterio de partición o "split", CART considera dos posibilidades diferenciadas:

- El índice de Gini ("gini index"), basado en la imprecisión de cada nodo (Breiman et. al, 1984).

- El criterio Twoing ("twoing criterion"), basado en la separación binaria del conjunto de categorías analizadas $C=\{1, 2,... J\}$.

$$\phi(s,t) = \frac{p_L p_R}{4}\left[\sum_j \left|p(j|t_L) - p(j|t_R)\right|\right]^2,$$

siendo p la probabilidad asociada con cada clase en cada nodo, que da lugar a la partición de los datos en dos subconjuntos (izquierda "L" y derecha "R"), tal que y , de forma que $p_L+p_R=1$ para cada nodo.

Por último, resulta conveniente apuntar que apenas suelen darse diferencias entre los árboles construidos a partir del índice de Gini y los grafos derivados del criterio Twoing. Con carácter general, la comparación de ambos

métodos respecto a un conjunto de datos significativo ha permitido observar que el índice de Gini genera particiones ligeramente más efectivas que el criterio Twoing, por lo que Breiman el a. (1984) recomiendan el empleo de la primera alternativa.

Una vez analizadas las características más destacadas de los árboles de decisión y, particularmente, del modelo CART para la construcción de grafos de inducción, a continuación se plantea la aplicabilidad real de este tipo de técnicas para la segmentación de clientes, comparando su desempeño con técnicas estadísticas clásicas y considerando diferentes costes de error.

3. APLICACIÓN EMPÍRICA

La gestión promocional relacionada con el *mailing* a clientes ha sido tradicionalmente una de las políticas más empleadas para dar a conocer al mercado un nuevo producto o servicio. No obstante, habitualmente muchos de los receptores no se encuentran interesados en la oferta recibida, por lo que no prestan atención a la información que reciben. Esto se traduce en un incremento de los costes promocionales de la empresa, que utiliza muchos recursos para obtener información acerca de consumidores que apenas sí generan rendimiento.

Ante esta situación, la empresa se plantea la necesidad de conseguir un conocimiento más completo de sus potenciales clientes, identificando a los individuos interesados respecto a aquellos que no lo están, y facilitando así el diseño de estrategias comerciales que reduzcan costes y eviten desperdicios de tiempo y esfuerzo.

En este sentido, la base de datos analizada recoge información relativa a 5.822 potenciales de una empresa aseguradora, que se plantea como objetivo determinar las características más significativas de aquellos individuos que adquirieron una póliza de vehículo-autocaravana tras un proceso de mailing[6]. Esta información serviría de base a la entidad para el desarrollo de un nuevo mailing en otra zona geográfica diferente, destinado a 4.000 nuevos clientes, al objeto de incrementar la tasa de respuesta afirmativa (que en el primer caso fue tan sólo de 348 individuos, esto es del 5,97%), y minimizar los costes promocionales correspondientes.

Respecto a la base analizada, cada cliente se encuentra caracterizado por 85 variables, de carácter económico, sociodemográfico y psicográfico (ver Anexo), de las que 5 tienen naturaleza categórica y el resto se encuentran medidas en escala ordinal. Por otro lado, para el aprendizaje de los modelos se ha utilizado el subconjunto de entrenamiento formado por los 5.822 individuos iniciales, utilizando la submuestra de test (4.000 nuevos individuos) para verificar el desempeño relativo de cada técnica.

Por otro lado, debe considerarse que la capacidad de cada modelo depende particularmente de dos variables fundamentales:

- El coste de cada contacto promocional (c_i).
- El ingreso obtenido por cada contacto afirmativo o venta (p_i).

De esta forma, resulta posible definir una función de coste-beneficio para el modelo analizado, como sigue:

[6] La base de datos es una de las propuestas en el CoIL Challenge (2000). Para más información puede consultarse http://www.dcs.napier.ac.uk/coil/challenge

$$N = N_{i1} + N_{i2}$$

siendo C_i el coste total asociado con el escenario *i*-ésimo, N el total de individuos analizados, N_{i1} el total de individuos a los que se ha mandado el mailing, N_{i2} el total de individuos a los que se ha decidido no enviar el mailing, A_{i1} el total de individuos que han contestado afirmativamente al mailing recibido y A_{i2} el número de individuos que hubieran contestado afirmativamente en caso de haber recibido tal mailing.

Como puede observarse, sea cual sea el escenario analizado, la función de coste-beneficio anterior presenta siempre dos límites bien definidos:

- Si el mailing es absolutamente correcto ($A_{i1}=N_{i1}, A_{i2}=0$) :
- Si el mailing es absolutamente erróneo ($A_{i1}=0, A_{i2}= N_{i2}$): $C_{erróneo} = -N_{i1}c_i - N_{i2}p_i = -(N_{i1}c_i + N_{i2}p_i)$

A objeto de análisis, resulta posible relacionar el ingreso por acierto (p_i) con el coste del mailing (c_i), obteniéndose la expresión:

$$t_i = \frac{p_i}{c_i},$$

que puede interpretarse como la relación que existe el coste y el beneficio de la política; así t_i=10 indicaría que cada venta realizada genera un ingreso que permite cubrir 10 envíos de mailing, t_i=20 informa de que cada respuesta afirmativa permite financiar 20 envíos, etc.

De esta forma, resulta posible redefinir los límites anteriores, como sigue:

- Si el mailing es absolutamente correcto ($A_{i1}=N_{i1}, A_{i2}=0$) : $\frac{C_{correcto}}{c_i} = N_{i1}(t_i - 1)$
- Si el mailing es absolutamente erróneo ($A_{i1}=0, A_{i2}= N_{i2}$): $\frac{C_{erróneo}}{c_i} = -N_{i1}1 - N_{i2}t_i = -(N_{i1}1 + N_{i2}t_i)$

Por su parte, el coste relativo de un escenario determinado tomaría la expresión:

$$\frac{C_i}{c_i} = A_{i1}t_i - N_{i1}1_i - A_{i2}t_i$$

Así, a partir de un determinado modelo, y considerando cada i-ésimo escenario, puede estimarse su desempeño relativo como sigue:

$$D_i = \frac{(C_i - C_{erróneo})}{C_{correcto} - C_{erróneo}},$$

o bien, de forma redefinida:

$$\frac{D_i}{c_i} = \frac{(C_i - C_{erróneo}) / c_i}{(C_{correcto} - C_{erróneo}) / c_i} = \frac{(C_i / c_i - C_{erróneo} / c_i)}{(C_{correcto} / c_i - C_{erróneo} / c_i)}$$

Asumiendo que el coste del mailing es fijo, y de esta forma que los diversos escenarios de la empresa se refieren específicamente a la relación particular entre los ingresos por ventas y el coste de cada esfuerzo promocional, resulta posible normalizar ambos límites, de forma que $\frac{C_{correcto}}{c_i}$=1 y $\frac{C_{erróneo}}{c_i}$=0. Así, el desempeño normalizado del modelo se encontrará siempre entre los límites:

$$0 \leq D_{iN} \leq 1,$$

siendo más eficaz cuanto más próximo se encuentre a 1, y menos eficaz cuanto más cercano esté de 0.

La comparación de los distintos desempeños normalizados (submuestra de test) y escenarios respecto a diversos métodos de segmentación permitirá obtener una visión general de su capacidad de caracterización relativa. En concreto, se han considerado dos métodos de segmentación diferentes, uno de ellos de carácter estadístico clásico y el otro correspondiente a un modelo CART:

- Análisis discriminante lineal.

- CART, con partición "Gini", y poda basada en una submuestra de validación integrada por el 33% de los individuos de la submuestra inicial de entrenamiento.

Por su parte, para cada metodología se han considerado cinco escenarios diferenciados, relativos a las diversas relaciones entre p_i y c_i, como sigue: t_i=5; t_i=10; t_i=20; t_i=50; t_i=100.

La Tabla 3 recoge los resultados obtenidos para cada metodología y escenario analizado, medidos a través del desempeño normalizado D_{iN}, respecto a las submuestras de aprendizaje y test. Su representación gráfica puede consultarse en la Figura 2.

Asimismo, la Tabla 4 resume los individuos clasificados correcta e incorrectamente por cada método y escenario (submuestra de test).

Tabla 3. Resultados obtenidos para los distintos métodos y escenarios (función de coste-beneficio normalizada)

		t_i=5	**t_i=10**	**t_i=15**	**t_i=20**	**t_i=50**
Análisis discriminante lineal	*Aprendizaje*	0,9001 (30 variables)	0,9454 (36 variables)	0,9524 (42 variables)	0,9737 (44 variables)	0,9618 (56 variables)
	Test	0,82825	0,9251	0,9278	0,9316	0,9244
CART (Gini + poda)	*Aprendizaje*	0,9159 (2 variables)	0,9348 (2 variables)	0,9476 (3 variables)	0,9483 (1 variable)	0,9649 (3 variables)
	Test	0,9013	0,9234	0,9384	0,9437	0,9782

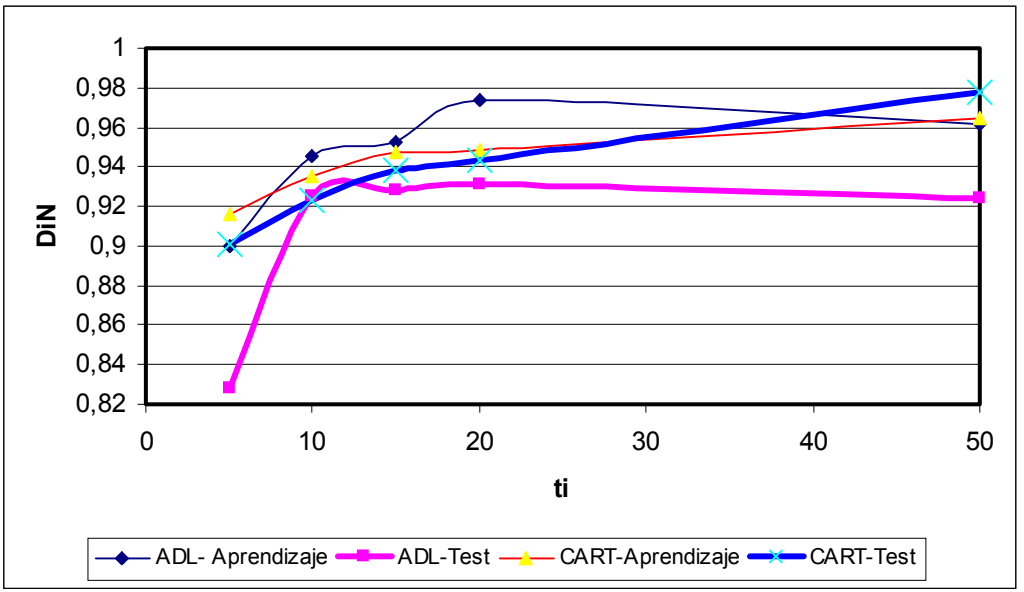
Tabla 4. Errores cometidos para los distintos métodos y escenarios

	t_i=5		**t_i=10**		**t_i=15**		**t_i=20**		**t_i=50**	
	P=1	**P=2**	**P=1**	**P=2**	**P=1**	**P=2**	**P=1**	**P=2**	**P=1**	**P=2**

Análisis discriminante lineal	**R=1**	35	203	117	121	110	128	113	125	90	148
	R=2	1608	2154	698	3064	622	3140	595	3167	471	3291
CART (Gini + poda)	**R=1**	52	186	122	116	151	87	160	78	224	14
	R=2	300	3462	859	2903	1174	2588	1457	2305	2967	795

Nota: R=real; P=predicho; 1=Respuesta afirmativa al mailing; 2=Respuesta negativa al mailing

Figura 2. Desempeño relativo de los distintos métodos y escenarios



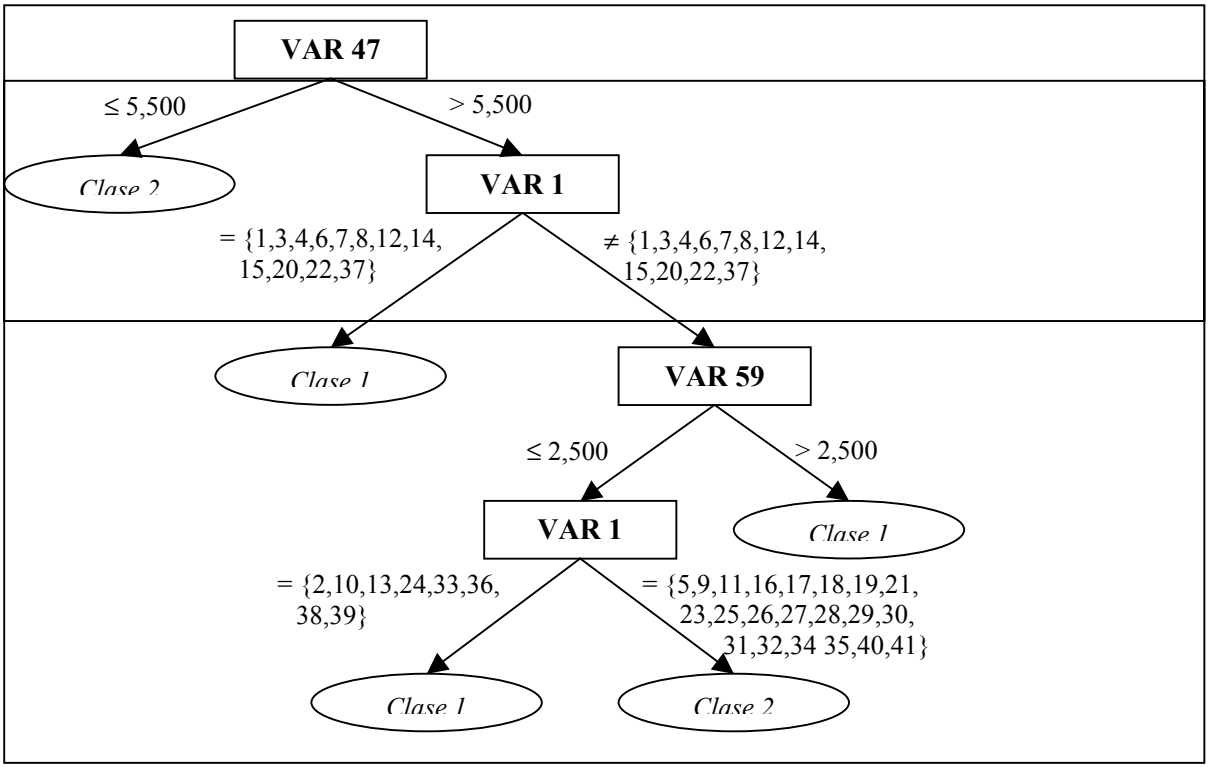
Como puede observarse, el modelo CART mejora los resultados obtenidos para el conjunto de patrones de test respecto a todos los escenarios analizados, con la única excepción del t_i=10, donde los resultados son similares. En este sentido, los árboles de decisión mejoran de forma continua el beneficio obtenido con el mailing, mientras que el modelo discriminante alcanza un máximo en t_i=10, para comenzar a continuación un descenso acusado.

No obstante, la característica más importante de los árboles de decisión no se refiere a su capacidad predictiva, sino a su facilidad de interpretación, al generar un número restringido de reglas fácilmente interpretables por el decisor. En este sentido, los modelos obtenidos han generado entre 2 y 5 reglas de decisión, integradas únicamente por entre 1 y 3 variables explicativas, que han demostrado un óptimo comportamiento para el conjunto de 5.822 patrones de entrenamiento y 4.000 patrones de test. El análisis discriminante múltiple, por el contrario, precisa entre 30 y 56 variables explicativas, incrementando la dimensionalidad del problema y dificultando la interpretación de los resultados obtenidos.

La Figura 3 resume el árbol de decisión relacionado con el escenario t_i=0.15, así como la función discriminante asociada al mismo.

Por otro lado, debe considerarse que la obtención de información acerca de la base de clientes, así como su posterior almacenamiento, no resulta gratuita a la empresa sino que, por el contrario, requiere un coste significativo de tiempo y recursos económicos. De esta forma, el modelo CART, al precisar un número de variables muy reducido, disminuye también este tipo de costes, por lo que la función absoluta de coste-beneficio resulta comparativamente aún más positiva.

Figura 3. Árbol de decisión CART asociado con el escenario t_i=20



Función discriminante lineal (coeficientes canónicos no estandarizados):

Y = 0,061VAR1-0,258VAR2-0,238VAR4-0,258VAR5-0,051VAR6+0,025VAR7+0,069VAR8+0,095VAR10-0,044VAR14+0,094VAR16-0,119VAR18+0,074VAR19-0,094VAR21+0,091VAR22+0,048VAR24+0,147VAR28-0,037VAR30+0,064VAR32-0,039VAR36+0,043VAR38+0,054VAR40-0,240VAR41+0,058VAR42+0,063VAR43+0,436VAR440,388VAR46+0,210VAR47-0,056VAR49-0,362VAR50+0,190VAR51-0,409VAR53+0,174VAR57+0,965VAR58+0,286VAR59+0,685VAR62-0,617VAR65-0,565VAR75-0,697VAR77-4,034VAR79-0,642VAR80+1,704VAR82+0,418VAR85-3,693

Como resulta evidente, el árbol de decisión proporciona información significativamente más comprensible que la función discriminante que, debido al elevado número de variables explicativas que considera, resulta prácticamente ininteligible por parte del gestor de empresa.

Por el contrario, el árbol generado mediante CART genera únicamente cinco reglas de aprendizaje, que puede expresarse a través de reglas lógicas del tipo "si… entonces…", como sigue:

"Si VAR47<=5,5 entonces NO ENVIAR EL MAILING"
"Si VAR47>5,5 Y VAR1□{1,3,4,6,7,8,12,14,15,20,22,37} entonces SÍ ENVIAR EL MAILING"
"Si VAR47>5,5 Y VAR59 <=2,5 Y VAR1□{2,10,13,24,33,36,38,39} entonces SÍ ENVIAR EL MAILING"
"Si VAR47>5,5 Y VAR59 <=2,5 Y VAR1□{resto categorías} entonces NO ENVIAR EL MAILING"
"Si VAR47>5,5 Y VAR59>2,5 entonces SÍ ENVIAR EL MAILING"

Teniendo en cuenta que la variable 47 representa el número de pólizas de seguro de automóviles, la variable 59 el total de pólizas de seguro de incendio (variable *proxy* de la propensión del cliente a la cobertura de riesgos) y que la variable 1 se refiere al subtipo de cliente, siendo las primeras clases (1, 2, 3, etc.) las correspondientes a consumidores con alto nivel adquisitivo y estilo de vida liberal, mientras que las últimas categorías (27, 28, 29, etc) se refieren a personas con menor nivel adquisitivo y estilo de vida conservador, puede concluirse que el mailing debe dirigirse principalmente a clientes que, o bien posean más de dos automóviles asegurados, o bien no lleguen a dicha cantidad pero demuestren un estilo de vida liberal y moderno, y/o una elevada propensión a la cobertura de riesgos.

Por el contrario, el estudio del modelo discriminante impide obtener ningún tipo de resultado explicativo acerca de los consumidores, lo que llevaría a los gestores a tomar decisiones que no comprenden y, de este modo, que no pueden controlar.

3. CONCLUSIONES

La promoción de ventas constituye una de las decisiones más importante para el gestor de empresa, especialmente ante el reto del lanzamiento de nuevos productos y/o servicios.

No obstante, la política promocional tiene un coste significativo para la entidad, lo que obliga a la dirección a segmentar sus clientes de forma diferenciada, al objeto de direccionar claramente el esfuerzo de comunicación.

Así, en los últimos años se ha planteado la gestión intensiva de bases de datos a través de modelos predictivos a priori que, a partir de los atributos observados sobre clientes pasados, permitan predecir los consumidores potenciales más interesados en el producto ofertado.

Si bien las técnicas estadísticas tradicionales constituyen una alternativa muy válida en este ámbito, sin embargo, la presencia de un número creciente de variables caracterizadoras, muy relacionadas entre sí, dificulta el cumplimiento de sus hipótesis de partida, al tiempo que genera modelos extremadamente complejos y muy difíciles de explicar.

Ante esta situación, las técnicas basadas en Aprendizaje-Máquina y, en concreto, los árboles de decisión, permiten la construcción de modelos robustos fácilmente interpretables que, mediante la generación de sentencias lógicas del tipo "si…entonces…" facilitan el análisis por parte del gestor de las conclusiones obtenidas por el modelo.

Al objeto de contrastar la capacidad real de estos sistemas, se ha analizado una base de datos de clientes de entidades aseguradoras, definida mediante 85 variables explicativas y 5.822 individuos. La aplicación alternativa de los modelos de análisis discriminante lineal y árboles CART para distintos escenarios (diferentes relaciones

coste-beneficio) ha verificado, no sólo el mejor desempeño de los segundos, sino su capacidad para caracterizar de forma sencilla a los clientes potencialmente más atractivos para la empresa.

De esta forma, puede concluirse que los árboles de decisión constituyen una metodología de análisis muy útil en el ámbito de la gestión promocional, que se comportan de forma robusta con independencia de la función de coste-beneficio utilizada, y que generan reglas de decisión comprensibles para el usuario y ajustadas a la información subyacente en los datos.

BIBLIOGRAFÍA

Bhattacharyya, s. (2000): "Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing". *Proceedings 6th ACM SICKDD Int'l Conference on Knowledge Discovery & Data Mining (KDD-00),* págs. 465-473.

Breiman, L.; Friedman, J.H.; Olshen, R.A. y Stone, C.J. (1984): *Classification and Regression Trees,* Chapman & Hall, New York.

Brijs, T. (2002): *Retail Market Basket Analysis: A Quantitative Modelling Approach.* Tesis Doctoral. Department of Applied Economics, Limburg University Center.

Hunt, E.B. (1962): *Concept Learning: An Information Processing Problem.* Wiley, New York.

Hunt, E.B.; Marin, J. y Stone, P.J. (1966): *Experiments in induction,* Academic Press, New York.

Kim, Y.; Street, W.N. y Menczer, f. (2000): *"An evolutionary multi-objective local selection algorithm for custormer targeting"*. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Porter, M.E. (1986): *"Estrategia Competitiva"*. Campus. Rio de Janeiro.

Simmonds, (1981): "Strategic Management Accounting", *Management Accounting,* abril.

Smith, W. (1956): "Product differentiation and market segmentation as alternative marketing strategies". *Journal of Marketing,* vol. 21, págs. 3-8

Wedel, M. y Kamakura, W.A. (2000): *Market Segmentation: Conceptual and Methodological Foundations"*. Kluwer Academic Press, Boston.

ANEXO

Tabla 3. Atributos exógenos incluidos en la base de datos ("caravane insurance segmentation")

Nº	**Nombre**	***Descripción***	***Nº***	**Nombre**	***Descripción***
1	MOSTYPE	Subtipo de cliente (*)	44	PWAPART	Póliza seguro particular a terceros (*****)
2	MAANTHUI	Número de casas 1-10	45	PWABEDR	Póliza seguro a terceros (empresas)
3	MGEMOMV	Tamaño medio de la casa 1-6	46	PWALAND	Póliza seguro a terceros (agricultura)
4	MGEMLEEF	Edad media (**)	47	PPERSAUT	Póliza seguro de automóviles
5	MOSHOOFD	Tipo principal de cliente (***)	48	PBESAUT	Póliz seguro monovolumen
6	MGODRK	Católico (****)	49	PMOTSCO	Póliza seguro motocicletas / scooters
7	MGODP	Protestante	50	PVRAAUT	Póliza seguro camiones
8	MGODOV	Otra religión	51	PAANHANG	Póliza seguro trailers
9	MGODGE	Sin religión	52	PTRACTOR	Póliza seguro tractores
10	MRELGE	Casado	53	PWERKT	Póliza seguro máquinas agrícolas
11	MRELSA	Pareja de hecho	54	PBROM	Póliza ciclomotores
12	MRELOV	Otra relación	55	PLEVEN	Póliza seguro de vida
13	MFALLEN	Soltero	56	PPERSONG	Póliza seguro particular de accidentes
14	MFGEKIND	Hogar sin hijos	57	PGEZONG	Póliza seguro familiar de accidentes
15	MFWEKIND	Hogar con hijos	58	PWAOREG	Póliza seguro de invalidez
16	MOPLHOOG	Educación superior	59	PBRAND	Póliza seguro de incendio
17	MOPLMIDD	Educación de grado medio	60	PZEILPL	Póliza seguro deportes marítimos
18	MOPLLAAG	Educación básica	61	PPLEZIER	Póliza seguro de barco
19	MBERHOOD	Estatus superior	62	PFIETS	Póliza seguro bicicletas
20	MBERZELF	Emprendedor	63	PINBOED	Póliza seguro propiedades
21	MERBOER	Granjero	64	PBYSTAND	Póliza seguro seguridad social
22	MBERMIDD	Mando intermedio	65	AWAPART	Número de terceros (particular)
23	MBERARBG	Trabajadores cualificados	66	AWABERD	Número de terceros (empresa)
24	MBERARBO	Trabajadores no cualificados	67	AWALAND	Número de terceros (agricultura)
25	MSKA	Clase social A	68	APERSAUT	Número de pólizas de automóviles
26	MSKB1	Clase social B1	69	ABESAUT	Número de pólizas de monovolúmenes

27	MSKB2	Clase social B2	70	AMOTSCO	Número de pólizas de motocicletas / scooters
28	MSKC	Clase social C	71	AVRAAUT	Número de pólizas de camiones
29	MSKD	Clase social D	72	AAANHANG	Número de pólizas de trailers
30	MHHUUR	Vivienda arrendada	73	ATRACTOR	Número de pólizas de tractores
31	MHKOOP	Vivienda en propiedad	74	AWERKT	Número de pólizas de maquinas agrícolas
32	MAUT1	1 coche	75	ABROM	Número de pólizas de ciclomotores
33	MAUT2	2 coches	76	ALEVEN	Número de pólizas de vida
34	MAUT0	Sin coche	77	APERSONG	Número de pólizas particulares de accidentes
35	MZFONDS	Servicio nacional de salud	78	AGEZONG	Número de pólizas familiares de accidentes
36	MZPART	Seguro privado de salud	79	AWAOREG	Número de pólizas de incapacidad
37	MINKM30	Renta < 30.000	80	ABRAND	Número de pólizas de incendio
38	MINK3045	Renta 30-45.000	81	AZEILPL	Número de pólizas de deportes marítimos
39	MINK4575	Renta 45-75.000	82	APLEZIER	Número de pólizas de barcos
40	MINK7512	Renta 75-122.000	83	AFIETS	Número de pólizas de bicicletas
41	MINK123M	Renta > 123.000	84	AINBOED	Número de pólizas de propiedades
42	MINKGEM	Ingresos medios	85	ABYSTAND	Número de pólizas de seguridad social
43	MKOOPKLA	Clase con poder de compra básico			

Notas: (*): Se definen 41 categorías; las primeras corresponden a urbanistas, liberales y personas de ingresos elevados; las últimas están relacionadas con clientes que viven en el campo, conservadores, de importantes creencias religiosas y de ingresos medios o bajos.

(**): Se definen 6 categorías, que varían desde 20-30 años hasta 70-80 años (ordenadas).

(***): Se definen 10 categorías, las primeras relacionadas con clases liberales y las últimas con clases conservadoras.

(****): Se definen 9 categorías, que varían del 0% al 100%.

(*****): Se definen 9 categorías, que varían desde 0 hasta superior a 20