6 Modelos predictivos de la redención de puntos para una tarjeta de fidelización

6.1 Objetivos e hipótesis

6.1.1. Introducción: la tarjeta Travel Club

Travel Club es el programa líder de fidelización en España, gracias a los más de 5 millones de hogares socios y que cada día pasan su Tarjeta Travel Club por los más de 11.000 establecimientos asociados al programa que hay repartidos por toda España.

Entre los patrocinadores, o establecimientos asociados, están: Grupo Eroski, Grupo Repsol, BBVA, Telefónica, Iberia, AVIS, Multiópticas, Viajes Ecuador, Sun Planet, Euromaster, Direct Seguros, Fagor, Fnac, Tryp Hoteles, hasta 30 patrocinadores líderes a su vez en sus respectivos sectores.

Travel Club es marca comercial de *Air Miles España*, S.A., una empresa que está presente en países como Reino Unido, Holanda, Emiratos Árabes Unidos o Canadá. En España nació en noviembre de 1996 y desde entonces ha entregado más de 2.500.000 de viajes y regalos a sus socios. Su catálogo completo de premios y una curiosa "calculadora de puntos" están fácilmente accesibles en línea en su web http://www.travelclub.es.

Por otro lado, y como parte de las actividades del Máster de Sistemas de Investigación de Mercados de ESIC, en su curso 2002-2003, se realizó un caso práctico con una base de datos proporcionada por Travel Club. Uno de los problemas esenciales mencionados por los representantes de esta empresa fue la baja tasa de redención de puntos que estaban consiguiendo, para cuyo estudio prepararon un "*data mart*" o subconjunto de su base de datos, con casos con una cierta variedad en "inactividad", esto es, con varios meses de no acumular puntos.

Independientemente de los objetivos de aquel estudio, uno de los hechos relevantes era esa baja tasa de redención de puntos, un 7%, como veremos, y el hecho de que potencialmente había muchos clientes con puntos suficientes para su redención (intercambio por un regalo) que no lo estaban haciendo. Ya hemos visto en el capítulo 2 que la redención de puntos es esencial en un programa de este tipo, su razón de ser. ¿Por qué se estaba produciendo esto? Quizá nos resulte imposible saberlo. Pero sí podemos realizar los análisis necesarios para identificar cuáles de entre esas personas "inactivas" están en riesgo de abandonar el interés por el programa.

6.1.2. Objetivos

Nuestro objetivo de investigación será establecer umbrales en la acumulación de puntos a partir de los cuales debamos ponernos en contacto con el cliente que no haya redimido sus puntos por el riesgo que existe de que abandone el programa. Para ello elaboraremos varios modelos predictivos de la conducta de redención de puntos, y los evaluaremos mediante análisis de curvas ROC. Posteriormente, y una vez decidido el modelo, realizaremos un análisis coste-beneficio que nos permita establecer los umbrales o puntos de corte en puntos a partir de los cuales procederemos a una campaña de contactos con los clientes para realizar tareas de fidelización, o para conocer las razones por las que no están intercambiando sus puntos.

6.1.3. Hipótesis

Hipótesis 1: Si las curvas ROC son el mejor método para evaluar la capacidad predictiva de indicadores individuales, entonces

H1a: Mediante el análisis de curvas ROC sobre las variables de nuestra base de datos, individuales o agregadas de una manera simple, seremos capaces de encontrar un indicador con capacidad predictiva estadísticamente significativa.

H1b: Obtendremos conocimiento sobre la capacidad predictiva de las variables de interés en nuestra base de datos, mediante el cálculo de la curva ROC empírica.

H1c: Una vez calculadas éstas, podremos realizar contrastes estadísticos de significación de la capacidad predictiva de forma no paramétrica.

H1d: Y si obtenemos un indicador cuya distribución sea suficientemente normal, podremos optimizar la estimación de los indicadores de la curva ROC mediante la aplicación del modelo binormal.

H1e: Y a partir de la elección de un modelo de curva ROC de los pasos anteriores, seremos capaces de encontrar puntos de corte óptimos después de un análisis coste-beneficio.

Hipótesis 2: Si es posible estimar un modelo de regresión logística o un modelo de árbol de decisión sobre nuestros datos, entonces

H2a: Podremos decidir entre uno u otro para proponerlo en competición con el indicador individual que hemos encontrado antes.

H2b: A partir de su aplicación en la base de datos, podremos realizar análisis de curvas ROC que nos permitan comparar la capacidad predictiva de las dos aproximaciones (modelo estadístico vs. indicador único).

6.2. Método

6.2.1. Participantes

Esta investigación se basa en datos secundarios. La base de datos contiene 7411 registros en lo que es un "*data mart*" para la evaluación de la conducta de poseedores de la tarjeta Travel Club. Se trata de una base de datos de clientes que a la fecha de su confección (enero de 2003) llevaban al menos 1 mes inactivos.

Sabemos que un 32% son varones y un 28% mujeres, pero para el 40% restante desconocemos el sexo. Hay que tener en cuenta que los procedimientos de registro en Travel Club son muy variados y algunos de ellos no obligan al suscriptor a especificar sexo o edad.

En cuanto a ésta última variable, también hemos observado datos raros que hemos tenido que eliminar, como 96 personas con más de 100 años. Después de hacer las correcciones oportunas, la edad media de los participantes en la base de datos fue de 45 años. No hay diferencias significativas entre hombres y mujeres en cuanto a la edad.

Esta base de datos contiene asimismo participantes de 18 comunidades autónomas, aunque para 1497 de ellos este dato no existe.

Por otro lado hemos comprobado que en la base de datos original, y mediante una variable de segmentación denominada PMP (perfil multipatrocinador), los codificados como 0 no presentaban ninguna actividad durante al menos los últimos 6 meses, y una suma de puntos mínima, por lo que se consideraron totalmente inactivos. Estos registros fueron eliminados de la base de datos, por lo que la base final para el análisis contaba con 6032 filas.

El resto de variables de clasificación se describen en detalle en el apartado 6.3 de resultados. Hemos de señalar que en ningún momento tuvimos acceso a ningún dato personal de los participantes en la base de datos.

6.2.2. Aparatos, materiales y diseño experimental

Por la naturaleza de esta tesis, no hubo ninguna captura de datos que no fueran los que estaban disponibles en la base de datos.

Dado el carácter fuertemente metodológico de esta tesis, a continuación reseñamos el software utilizado y la plataforma de análisis:

- La descripción de datos se ha realizado principalmente mediante el SPSS para Windows, v. 10, y mediante el paquete estadístico NCSS 2004.
- El análisis de curvas ROC, cuando era posible realizarlo, se ha realizado mediante el módulo correspondiente del paquete estadístico NCSS 2004.
- Los modelos de regresión logística y de árbol de decisión se han realizado con el módulo Enterprise Miner v.4.1, sobre SAS v. 8.02 para Windows.
- Los gráficos se han realizado o con MS-Excel 97, o con gnuplot para Windows, v. 3.5

Todos los análisis se han realizado en un PC con Windows 98.

6.3. Resultados

6.3.1. Exploración de los datos

6.3.1.1. Variables sociodemográficas

La descripción de las variables sociodemográficas aparecen descritas en la tabla 6.1. (sexo y edad) y en la tabla 6.2. (variables geográficas).

Edad

Observamos que un 39.1% de los registros de nuestra base de datos no dispone de edad (este campo está vacío). Además hay algunos datos raros: además de un caso con 8 años, hay 96 personas con entre 102 y 103 años, que representa claramente un error. Todos ellos se han pasado a "missing", pero su uso significaría la pérdida de una cantidad importante de casos. Por lo que estudiaremos primero su capacidad predictiva para ver si es necesario realizar una imputación de estos datos que faltan. Si la capacidad predictiva de esta variable, por sí sola, no es suficientemente grande, la excluiremos de los análisis. La media de edad está en torno a los 45 años y no encontramos diferencias significativas en la edad entre varones y mujeres.

Sexo

La base contiene datos sobre esta variable para un 60% del total de suscriptores, que se distribuyen entre 2378 varones (código 1) y 2086 mujeres (código 2). Los restantes son datos *missing*.

Variables de situación geográfica

La base de datos contiene variables para recoger el código postal (aunque con muchas faltas) que sirven a su vez para codificar si el participante vive en hábitat rural o urbano, el código de provincia o la comunidad autónoma de pertenencia del cliente. La variable código postal contiene muchos errores y valores perdidos, por lo que será muy difícil de usar en nuestros análisis. En general, la única variable relevante con un buen conjunto de datos es el código de comunidad autónoma.

Características adicionales sobre el titular y la tarjeta Travel Club

Hay dos características que son muy importantes desde el punto de vista del negocio de Travel Club, que son si se dispone de tarjetas adicionales (hay varios miembros de la unidad familiar con una misma tarjeta) y el registro via *web*. Éste último dato es muy importante porque es un canal muy potente para mantener la relación con el cliente.

Tabla 6.1.Variables sociodemográficas del titular de tarjeta Travel Club

Variable	N	%	Media	Sx	Mínimo	Máximo
CDTITULA	6032	100%	n.a.	n.a.	1	7411
EDAD						
Datos válidos	4408	73.1%	43.7	14.36	19	91
SEXO	6032		n.a.	n.a.	-	-
Varón	2378	39.42%				
Mujer	2086	34.58%				
Valores perdidos (0)	1568	25.99				

Tabla 6.2.Variables geográficas del titular de tarjeta Travel Club

Variable	N	%	Media	Sx	Mínimo	Máximo
MUNI	4537	75.21%	n.a.	n.a.	3	52001
CDPROVIN	4537	75.21%	n.a.	n.a.	1	60
CODCCAA	6032	100%	n.a.	n.a.	1	18
99= valor perdido	1497	24.81%				
SITU	4469	74.1%	n.a.	n.a.	-	-
1= Urbano	3151	52.23%				
2= Rural	1318	29.49%				

Tabla 6.3.Variables sobre características adicionales de la tarjeta Travel Club

Variable	N	%	Media	Sx	Mínimo	Máximo
ONLINE	6032	100%	n.a.	n.a.	-	-
0= no alta en web	5677	94.11%				
1= alta en web	355	5.88%				
TARJ_ADI	6032	100%	n.a.	n.a.	-	-
0= no tarjeta adic.	5282	87.56%				
1= sí tarjetas adic.	750	12.44%				

Las figuras 6.1 a 6.3 muestran la distribución de las variables de tipo sociodemográfico más relevantes para nuestro análisis, que son la edad, el sexo y la comunidad autónoma, mientras que las figura 6.4. y 6.5. representan las proporciones de posesión de tarjetas adicionales y el registro en la web.

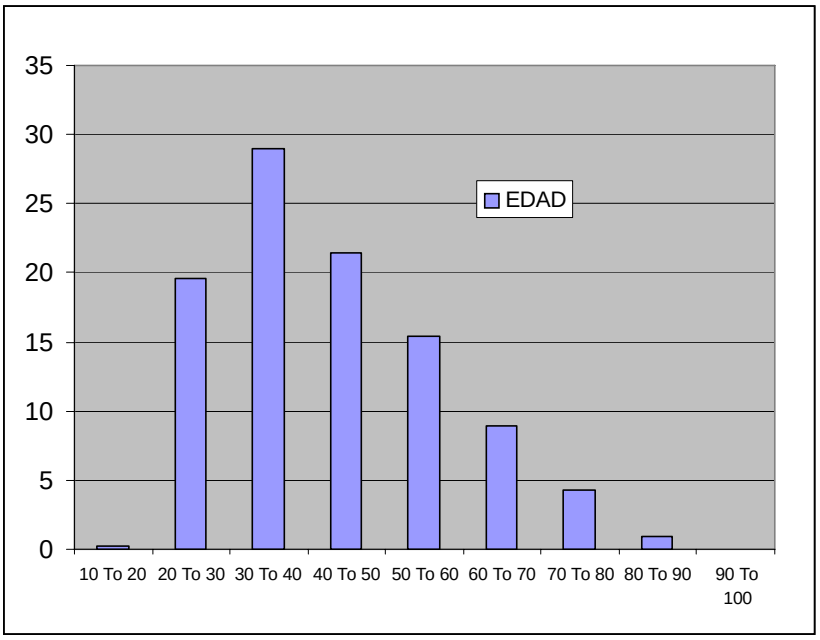


Figura 6.1. Distribución de edades en toda la base de datos

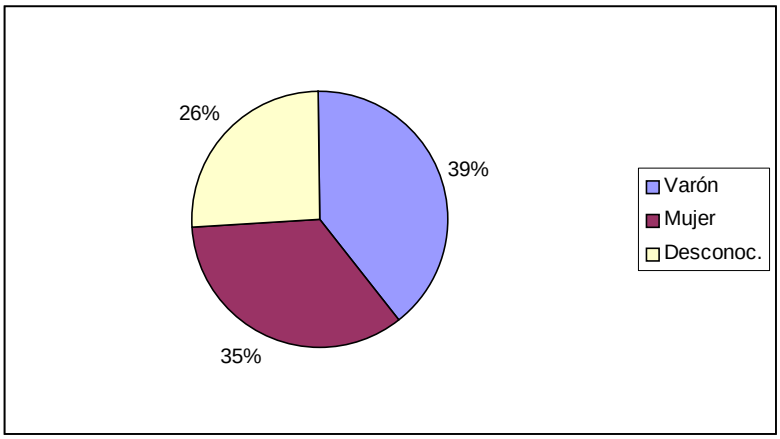


Figura 6.2. Distribución de sexo

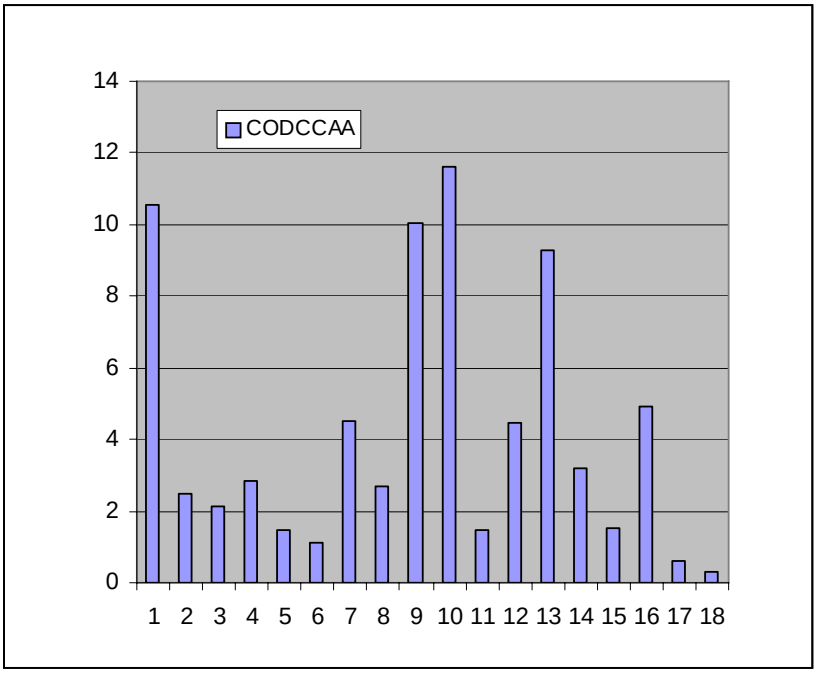


Figura 6.3. Distribución por Comunidad Autónoma (eliminando valores perdidos)

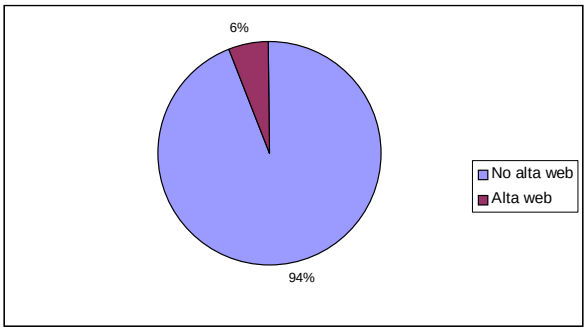


Figura 6.4. Distribución en función de si está suscrito en la web o no

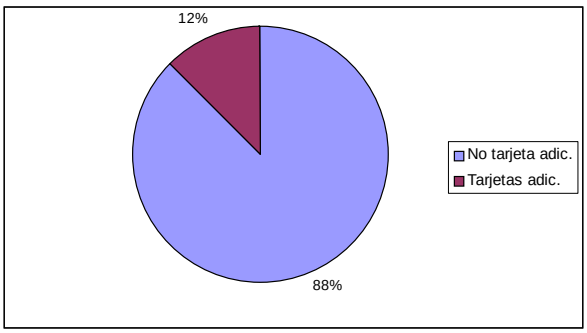


Figura 6.5. Distribución en función de si dispone de tarjetas adicionales o no

6.3.1.2. Variables de segmentación propias del negocio

Una variable que se muestra básica dentro de este grupo es la denominada PMP_ANO, que resume entre otras cosas el perfil multipatrocinador anual. Decimos entre otras cosas porque hemos comprobado que no se relaciona biunívocamente con el perfil multipatrocinador, por tanto debe incluir otras variables propias del negocio de Travel Club, que desconocemos. Ya se ha comentado que hemos descartado analizar un grupo definido por un código de esta variable porque carece casi por completo de puntos y hace al menos seis meses que ni suma ni redime puntos. El resto de grupos dentro de esta variable aparecen descritos en la tabla 6.4. Dado el gran desequilibrio existente entre los grupos (muy concentrados en 1 y muy dispersos a partir de 3) y previa exploración de su relación con la redención, decidimos recodificarla en una nueva variable elaborada por nosotros, denominada SEGPMP, de la que hablaremos en el punto 6.3.1.3.

Tabla 6.4.Variables de segmentación propias del negocio Travel Club

Variable	**N**	**%**	**Media**	**Sx**	**Mínimo**	**Máximo**
PMP_ANO	6032		n.a.	n.a.	1	6
0 (eliminados)	-					
1	4273	70.84%				
2	1421	23.56%				
3	303	5.02%				
4	29	0.48%				
5	5	0.08%				
6	1					
HMLNUM	6032		n.a.	n.a.	1	3
1 - Perfil compra bajo	3995	66.23%				
2 - Perfil compra medio.	1656	27.45%				
3 - Pefil compra alto	381	6.32%				
CDICE	6032		n.a.	n.a.	1	5
1	692	11.47%				
2	859	14.24%				
3	2811	46.60%				
4	1205	19.98%				
5	465	7.71%				

HMLNUM es una variable que registra el perfil de compra, con tres perfiles preestablecidos, no sabemos a partir de qué criterio. CDICE se trata de una codificación de la capacidad adquisitiva, y en la práctica es una variable ordinal, desde 1 (más bajo) hasta 5 (más alto) con una distribución aproximadamente simétrica (véase figura 6.8).

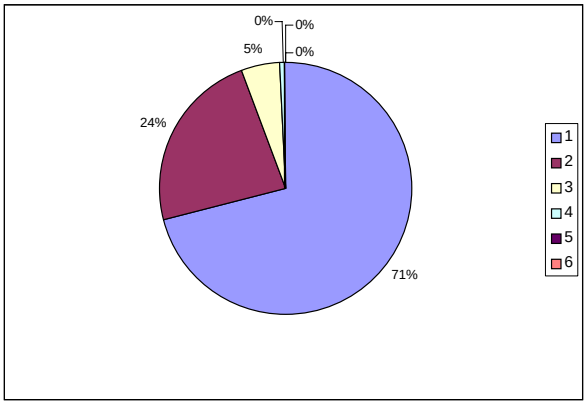


Figura 6.6. Distribución de PMP_ANO (perfil multipatrocinador)

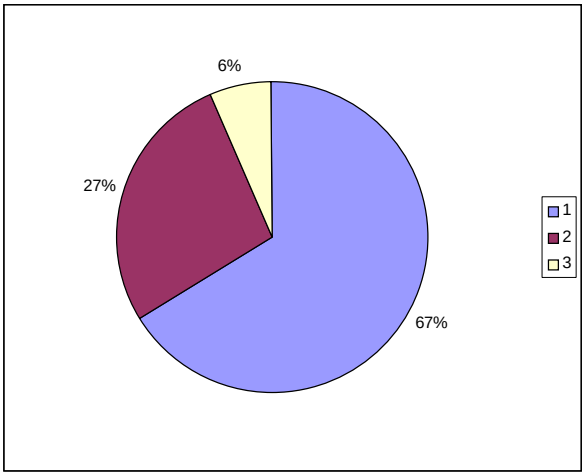


Figura 6.7. Distribución de HMLNUM (perfil de compras)

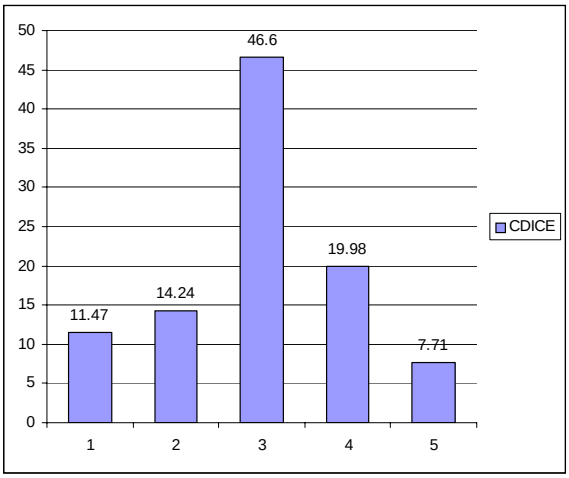


Figura 6.8. Distribución de CDICE, capacidad adquisitiva

6.3.1.3. Variables relacionadas con el tiempo de inactividad en el programa de puntos

Un tema fundamental de estudio es el periodo de inactividad. Es necesario recordar que la base de datos cubre un año (2002), y existen en la misma varias variables para indicar el mes de última actividad, tanto en su escala natural (1= enero, 12= diciembre), como transformada al número de meses inactivo. Dado que nuestro interés no es estudiar estos datos desde una perspectiva histórica, utilizaremos exclusivamente la variable NOMESINA, número de meses inactivo, que como su nombre indica, varía entre 1 (sólo un mes inactivo) hasta 12 (más meses), a mayor número, mayor duración de la inactividad. Sus resultados descriptivos se encuentran en la tabla 6.5. y la figura 6.9 muestra la distribución de esta variable en la población.

Tabla 6.5.Variable sobre inactividad en el programa de puntos

Variable	N	%	Media	Sx	Mínimo	Máximo
NOMESINA	6032	100%	5.21	3.41	1	12

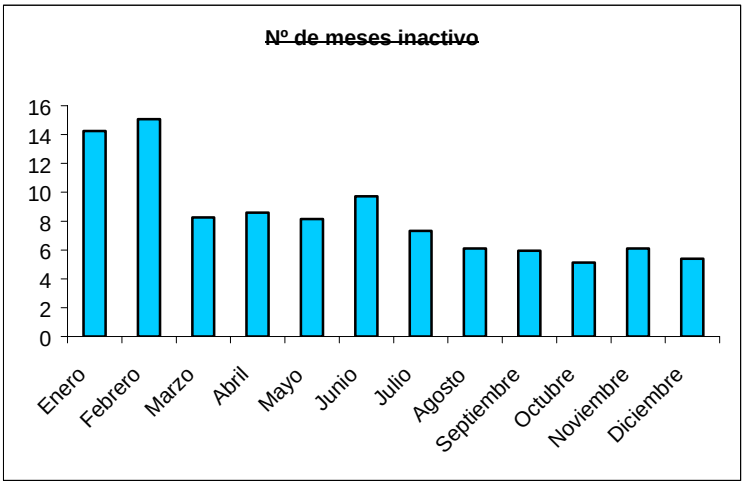


Figura 6.9. Distribución de NOMESINA, número de meses inactivo en el programa de puntos

6.3.1.4. Variables relacionadas con el número de patrocinadores a los que consumen los titulares

Uno de los objetivos básicos de un programa multipatrocinador como es Travel Club es que sus titulares sean clientes de cuantos más patrocinadores mejor. Por tanto, una variable esencial para nuestros análisis será la clasificación de los clientes en función de en cuántos y en qué patrocinadores compran. Se realizó un análisis exhaustivo de todas las variables de adquisición de puntos presentes en la base de datos para obtener la variable SEGPMP.

Tabla 6.6. Clasificación cruzada patrocinadores en los que se consume y la variable PMP_ANO

PATROCINADORES	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**	**Porc.**
mono eroski	854	2301	10					3165	42.71
mono rps	340	808	13					1161	15.67
mono bbv	1	684	3					688	9.28
mono tel		419	7					426	5.75
Sólo pequeños patrocinadores	35	58	8					101	1.36
2 patrocinadores (eroski + respsol)	125		783	18				926	12.49
2 patrocinadores (sin eroski ni repsol)	18	3	597	23	1	0	0	642	8.66
3 patrocinadores siempre con eroski	6	0	0	230	6	0	0	242	3.27
3 patrocinadores SIN eroski	0	0	0	32	1	0	0	33	0.45
más de 3 patrocinadores	0	0	0	0	21	5	1	27	0.36
Total	1379	4273	1421	303	29	5	1	7411	
Porc.	18.61	57.66	19.17	4.09	0.39	0.07	0.01	100.00	
Porc. acum.	18.61	76.27	95.44	99.53	99.92	99.99	100		

Si hacemos la tabulación cruzada entre los patrocinadores en los que se consume y esta variable de segmentación obtenemos la tabla 6.6. Esta clasificación se realizó antes de eliminar los codificados 0 en PMP_ANO, y por ese motivo aparecen todos en la tabla 6.6.

Podemos observar cómo esta distribución es muy desequilibrada. Pero hay algunos hecho significativos:

- Eroski es quien más clientes añade al programa con mucha diferencia. Pero una parte importante de estos clientes están en la categoría 0 de PMP, que no han redimido nunca y tienen un perfil de consumo bajísimo.
- De los perfiles multipatrocinador, Eroski sigue siendo quien más clientes añade al programa.
- Con los perfiles PMP 1 y 2 cubrimos las 3/4 partes de la base de datos.
- Según se va haciendo más complejo el perfil multipatrocinador, la dispersión de la base de datos se va acentuando, por lo que no compensará ni utilizar la variable original (PMP_ANO) ni la variable directa de número de patrocinadores en los que se consume

Por todas estas razones, se simplificó la clasificación cruzada anterior, quedando una nueva variable denominada SEGPMP con la distribución que se muestra en la tabla 6.7

Tabla 6.7.Distribución de la variable SEGPMP que aglutina y redistribuye el perfil multipatrocinador y la variable PMP_ANO

Variable	N	%	Media	Sx	Mínimo	Máximo
SEGPMP	6032		n.a.	n.a.	1	8
1 - mono-patroc. Eroski	2311	38.31%				
2 - mono-patroc. Repsol	821	13.61%				
3 - mono-patroc. BBV	687	11.39%				
4 - mono-patroc. Telefónica	426	7.06%				
5 - Sólo pequeños patroc.	66	1.09%				
6 - 2 patroc. (sin Eroski)	1105	18.32%				
7 - 2 patroc (con Eroski)	320	5.31%				
8 - más de dos patroc.	296	4.91%				

La figura 6.10. muestra la distribución de esta variable.

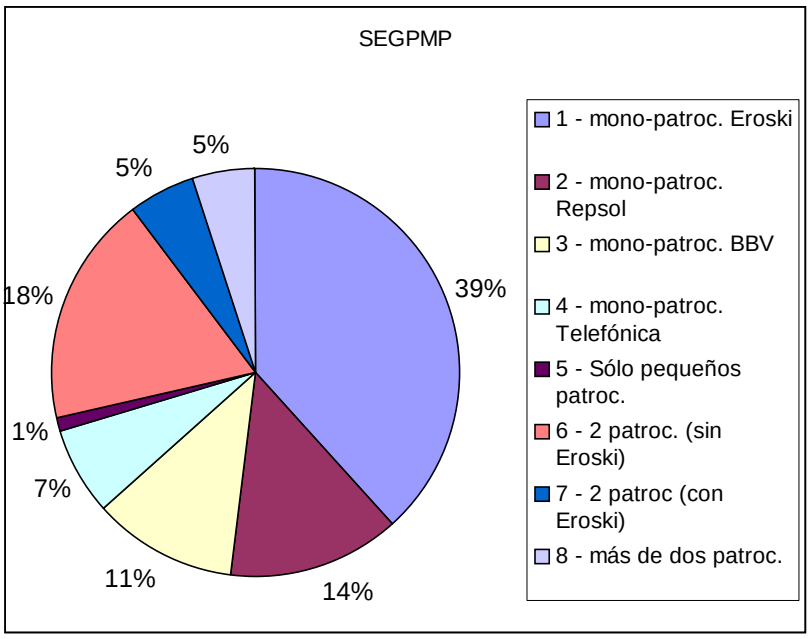


Figura 6.10. Distribución de la variable de clasificación SEGPMP.

6.3.1.5. Variables relacionadas con la obtención de puntos

Trataremos en este estudio únicamente las variables agregadas (totales anuales), esto es, totales de puntos obtenidos, puesto que el uso de modelos dinámicos está fuera del alcance de esta tesis.

Hay dos variables esenciales de suma de puntos, que son los puntos acumulados hasta el año 2002, y luego la suma de puntos obtenidos durante todo el año 2002, incluyendo por supuesto todos los patrocinadores en los que haya hecho visitas.

Además, los puntos redimidos (que habían sido lógicamente obtenidos), aparecían en variables separadas, por lo que decidimos sumarlas a la variable suma de todos los puntos (antes de 2002 y 2002) para formar la variable denominada TOTTOT3. La tabla 6.8 recoge los datos descriptivos de esta variable. No se representa su distribución en puntuaciones directas (puntos obtenidos) puesto que es típica de una distribución exponencial y por tanto de difícil representación.

Para llevar esta distribución a una más próxima a una normal se realizó la transformación logaritmo natural (Ln), con lo que tenemos una variable más, que denominamos LNTOT3. Sus datos descriptivos se recogen en la tabla 6.8.

Tabla 6.8. Descripción de la variable TOTTOT3 y su transformada LNTOT3

Variable	N	%	Media	Sx	Mínimo	Máximo
TOTTOT3	5966	98.90%	no tiene sentido, mediana =184 puntos	no tiene sentido	4	31615
LNTOT3	5966	98.90%	5.20 (=181 puntos)	1.54	1.38	10.36

Después de un estudio exhaustivo de la forma de la distribución de esta variable, hemos podido comprobar cómo, obtenemos una distribución global (para todos los casos) muy próxima a la normal, como se muestra en la figura 6.11.

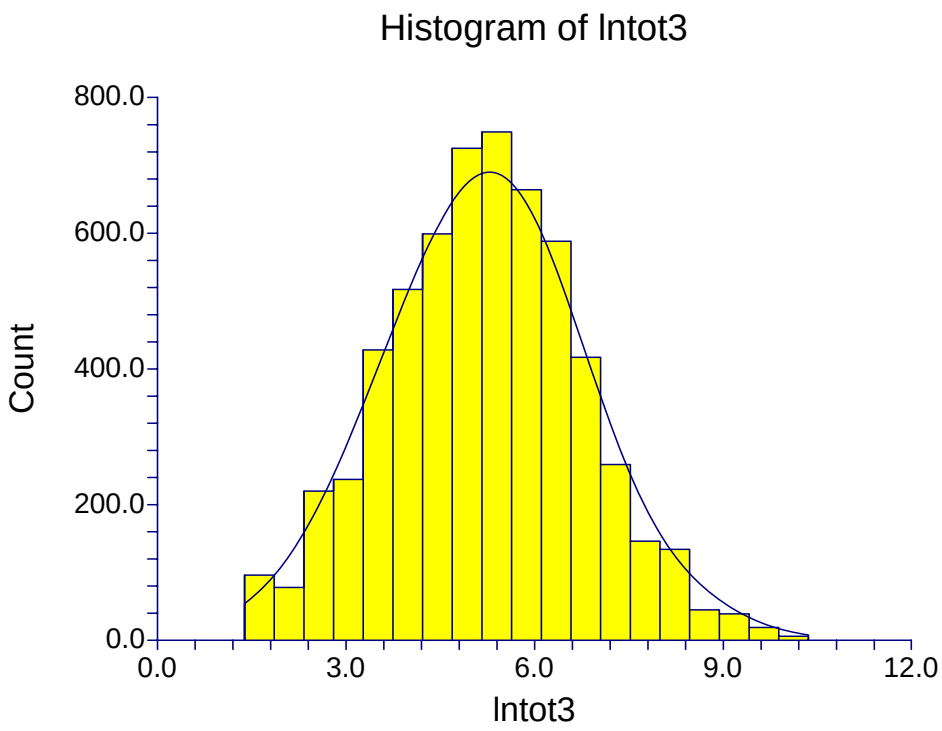


Figura 6.11. Distribución de LNTOT3

Los contrastes estadísticos de normalidad, realizados con el programa NCSS 2004, se presentan en la tabla 6.9. Podemos comprobar que según la prueba de Kolmogorov-Smirnov (0.019, p>0.05) esta distribución es normal.

Tabla 6.9. Pruebas de normalidad que realiza NCSS 2004 sobre la variable LNTOT3

Prueba	Valor del E.C.	Nivel de prob.	Valor crítico 10%	Valor crítico 5%	Decisión -5%
Shapiro-Wilk W	0.9973074	7.74E-03			Se rechaza normalidad
Anderson-Darling	1.625.653	3.57E+02			Se rechaza normalidad
Martinez-Iglewicz	0.9854234		0.9942221	0.9939333	NO se rechaza normalidad
Kolmogorov-Smirnov	1.19E-02		0.014	0.015	NO se rechaza normalidad
D'Agostino Skewness	1.381.571	0.1671034	1.645	1.960	NO se rechaza normalidad
D'Agostino Kurtosis	0.1138	0.909403	1.645	1.960	NO se rechaza normalidad
D'Agostino Omnibus	19.217	0.38257	4.605	5.991	NO se rechaza normalidad

6.3.1.6. Variables relacionadas con la redención de puntos

Ya se ha comentado que en la base de datos figuran variables que registran los puntos redimidos en periodos concretos, pero nuestra variable "objetivo" o "*target*" en el argot de Marketing es si ha redimido en alguna ocasión o no. Esta es la variable REDIME, cuya descripción aparece en la tabla 6.10 y en el gráfico 6.12.

Tabla 6.10. Descripción de la variable TOTTOT3 y su transformada LNTOT3

Variable	N	%	Media	Sx	Mínimo	Máximo
REDIME	6032	100%	n.a.	n.a.	-	-
0= Nunca ha redimido	5471	90.70%				
1= Ha redimido en alguna ocasión	561	9.30%				

Nuestro objetivo de análisis irá dirigido ahora a comprobar qué variables tienen relación con esta variable criterio.

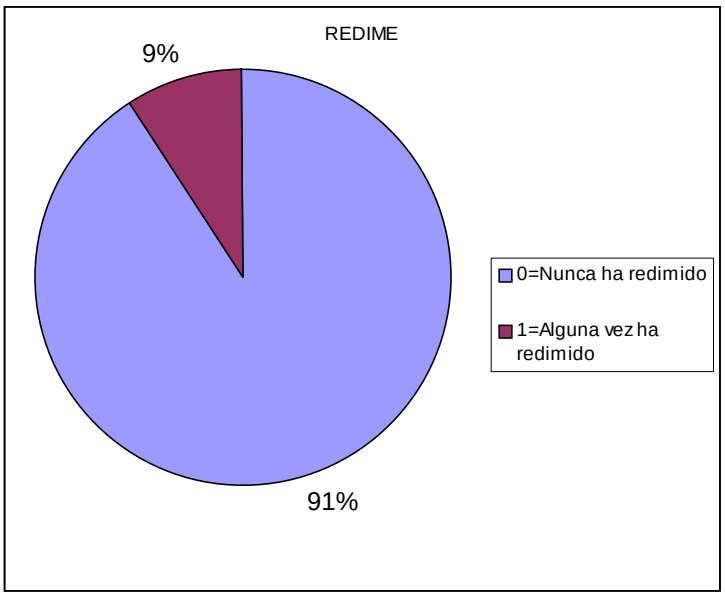


Figura 6.12. Distribución de la redención de puntos (REDIME) en el grupo total

6.3.2. Indicadores relacionados con la conducta de redención

6.3.2.1. Número de meses inactivo

La base de datos incorpora una variable que indica el número de meses inactivo, esto es, el número de meses desde que acumuló puntos por última vez. Podemos observar su relación con el comportamiento de redención en la figura 6.13., a mayor número de meses inactivo menor porcentaje de redención.

Pero, ¿hasta qué punto tiene capacidad predictiva? Como hemos visto, la curva ROC proporciona un indicador único de la capacidad predictiva de un indicador. En este caso esta curva ROC, empírica, es la mostrada en la figura 6.14. Observamos a simple vista que la capacidad predictiva es relativamente baja, aunque pueda ser significativa. La tabla 6.11 muestra que el área bajo la curva ROC es significativamente superior a 0.5 con $p<0.001$. Podemos entonces afirmar que el número de meses inactivo, aislado, predice algo nuestra variable criterio.

Tabla 6.11. Resultados del contraste de significación para la ROC empírica con NOMESINA

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of REDIME	Count
NOMESINA	0.56572	0.01206	5.45	<0.001	<0.001	0.0757	7411

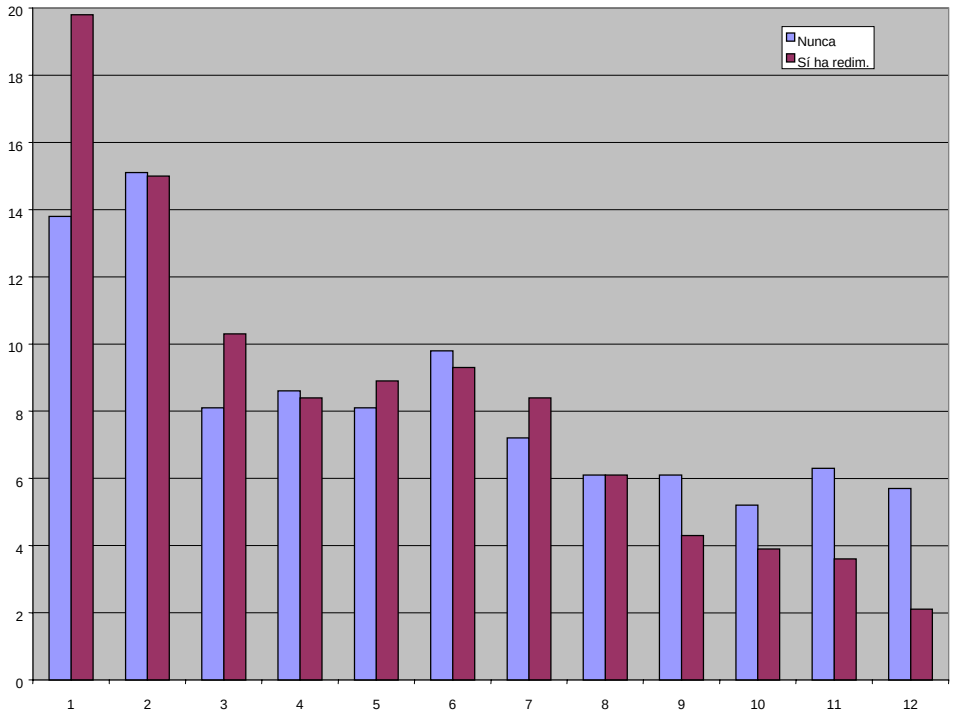


Figura 6.13. Distribución de número de meses inactivo por redención

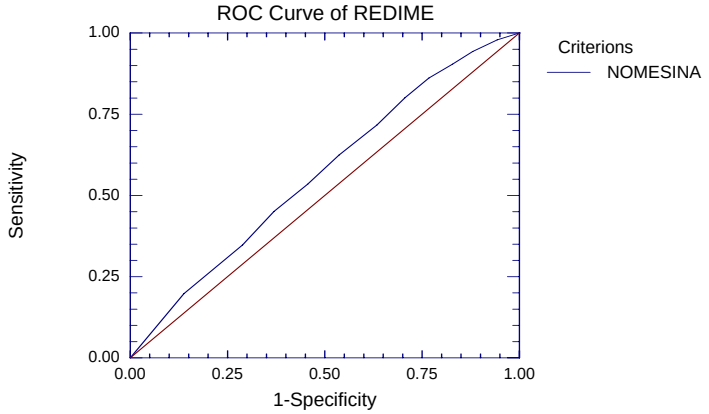


Figura 6.14. Curva ROC empírica de la relación de NOMESINA con REDIME

6.3.2.2. Variables de segmentación

Existe una variable en la base de datos llamada HMLNUM, donde HML significa *High, Medium, Low*, y que es una vez más una variable propia de segmentación de Travel Club. Esta variable también muestra una relación importante con la conducta de redención. Podemos observar su relación con la conducta de redención en la figura 6.15.

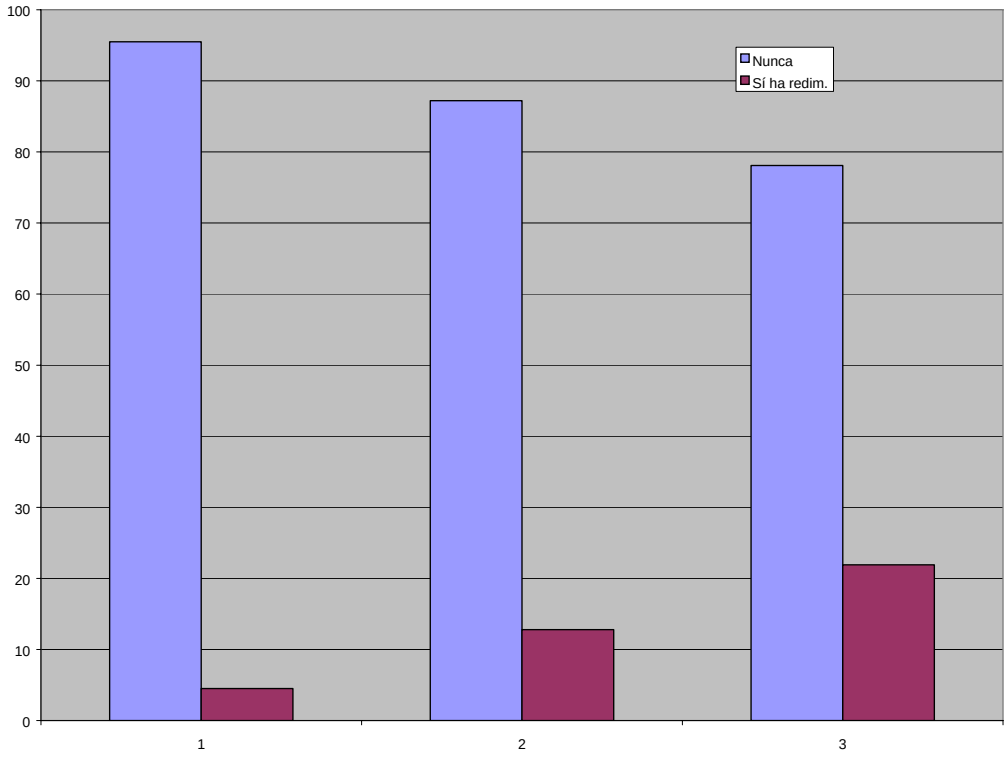


Figura 6.15. Relación entre HMLNUM y REDIME

Con esta variable también podremos dibujar una ROC empírica, aunque NCSS 2004 nos avise que con tan pocas categorías el área bajo la curva puede estar subestimada. Pero es preferible a dibujar conjuntamente esta variable junto con las otras de segmentación, en particular CDICE. No podremos dibujar además SEGPMP puesto que es una variable esencialmente nominal.

Se puede ver que mientras HMLNUM predice significativamente la conducta de redención ($z=12.84$, $p<0.0001$), CDICE no predice significativamente ($z=0.34$, $p=0.36$).

Tabla 6.12. Resultados del contraste de significación para la ROC empírica con HMLNUM y CDICE

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of REDIME	Count
HMLNUM	0.64698	0.01144	12.84	0.0000	0.0000	0.09300	6032
CDICE	0.50456	0.01324	0.34	0.3652	0.7304	0.09300	6032

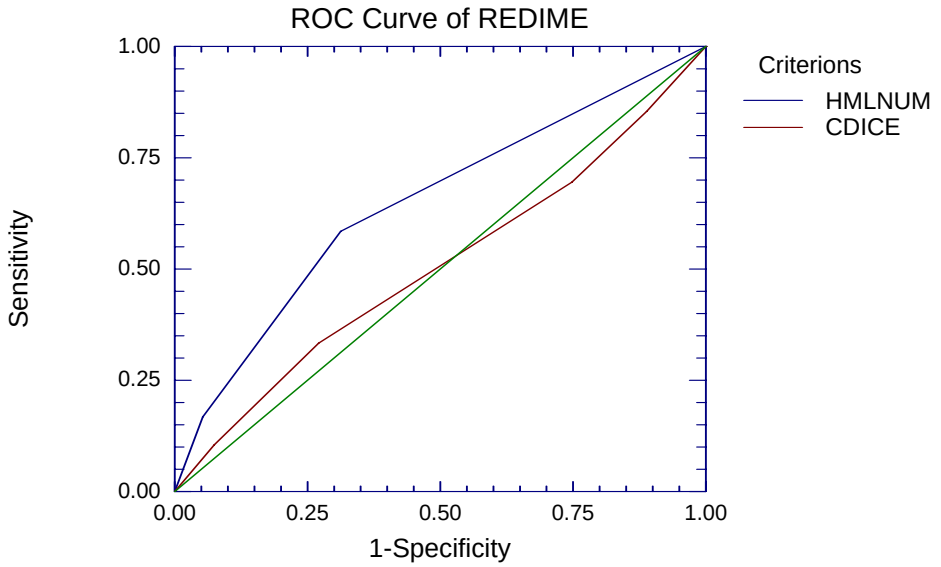


Figura 6.16. Curvas ROC para HMLNUM (superior) y CDICE (inferior)

En cualquier caso, no predice suficiente o muestra relación más baja por sí solo con conducta de redención que otros indicadores en la base de datos.

6.3.2.4. Edad y otras variables del cliente

Edad es una variable importante en cualquier análisis. Veremos a continuación su relación con la conducta de redención a partir del cálculo de la curva ROC empírica. La tabla 6.13 muestra los resultados de las pruebas de significación, y la figura 6.17 la curva ROC. Se puede comprobar que por sí sola tiene una capacidad predictiva significativa.

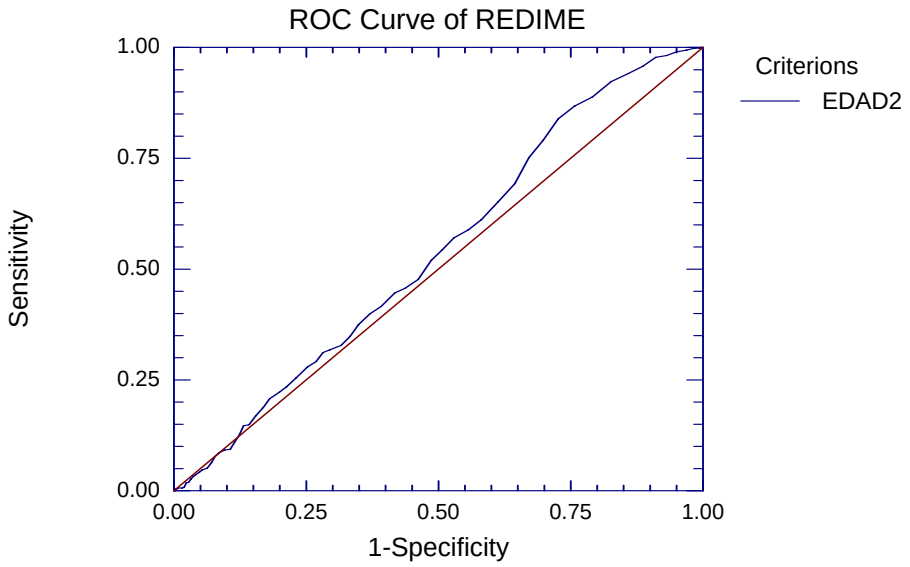


Figura 6.17. Curvas ROC para EDAD (eliminados casos extraños)

Tabla 6.13. Resultados del contraste de significación para la ROC empírica con EDAD2

Criterion	Empirical Estimate of AUC	AUC's Standard Error	Z-Value to Test AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalence of REDIME	Count
EDAD2	0.54012	0.01275	3.15	0.0008	0.0017	0.11139	4408

Otra variable que es importante en este contexto es si el cliente está registrado en la web de Travel Club de tal manera que puede interactuar via web con el programa de puntos. Los datos de suscripción en la base de datos, en relación con su conducta de redención se representan en la figura 6.18. Podemos ver cómo es una minoría la que está registrada en la web, menos de un 5% del total de la base de datos. Pero a simple vista podemos observar una relación importante con la conducta de redención.

Por otro lado está la disponibilidad de tarjetas adicionales, que también muestra una cierta relación con la conducta de redención. Esta relación se representa en la figura 6.19.

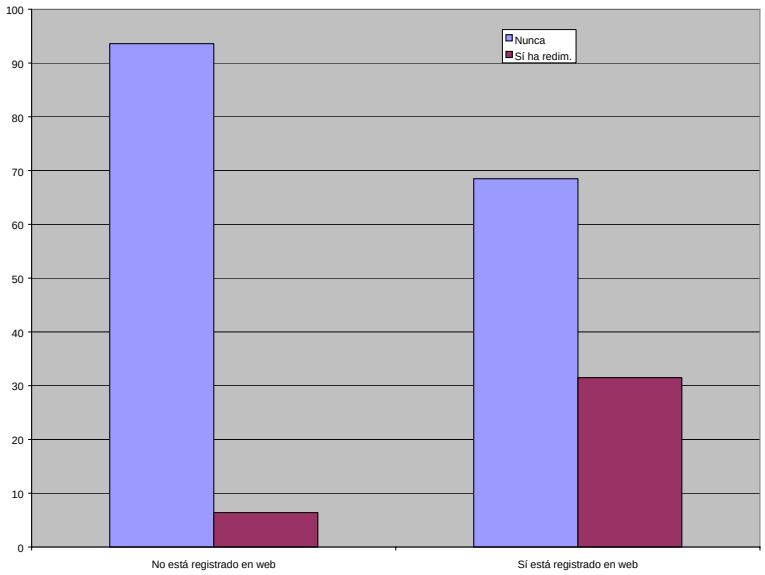


Figura 6.18. Relación entre registro online y REDIME

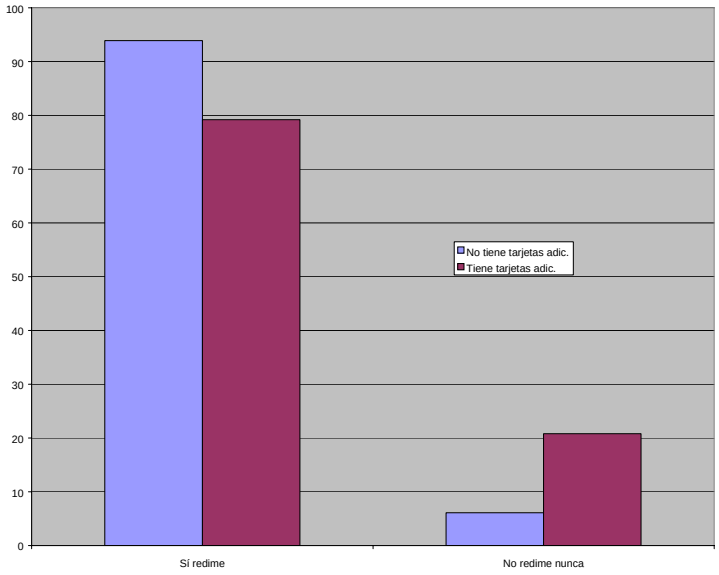


Figura 6.19. Relación entre disponibilidad de tarjetas adicionales y REDIME

6.3.2.3. Total de puntos acumulados

En primer lugar, calculamos un indicador basado en los puntos obtenidos en el año inmediatamente anterior a la obtención de la base de datos (2002). Esta variable la denominamos TOTTOT. Comprobamos mediante un análisis de curvas ROC empíricas

que su capacidad predictiva de la conducta de redención era significativa (AUC=0.73, z=20.75, p<0.001).

Pero por otro lado, en la base de datos hay variables que indican también el total acumulado de puntos históricamente. Con ello podemos elaborar la variable TOTTOT2, y a su vez calculando la curva ROC empírica tendremos una indicación de su valor predictivo, que después del contraste proporciona un AUC de 0.72, z=19.39, p<0.001.

Una de las grandes ventajas del análisis de curvas ROC empírico (o no paramétrico), es la posibilidad de comparar mediante un contraste estadístico el área bajo la curva para varios grupos independientes. Tomando como variable de agrupación HMLNUM, tenemos las siguientes curvas ROC empíricas (figura 6.20).

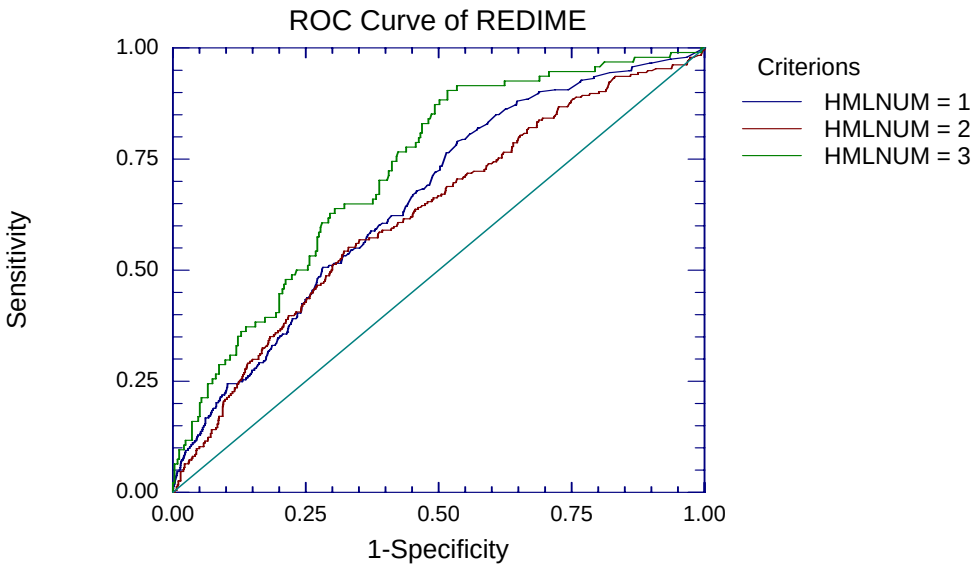


Figura 6.20. Curvas ROC empíricas para el indicador TOTTOT2 por HMLNUM

Y los contrastes estadísticos de diferencias entre las áreas de la curva se representan en la tabla 6.14. Podemos observar cómo según hacemos mayor el valor de HMLNUM (clasificación por gasto), aumentamos la capacidad predictiva de TOTTOT2. Pero ¿podemos mejorar la capacidad predictiva de esta suma total de puntos obtenidos?

Tabla 6.14. Resultados del contraste de significación para la ROC empírica con HMLNUM

HMLNUM	AUC1	AUC2	Difference Value	Difference Std Error	Z-Value	Prob Level
1, 2	0.65749	0.62722	0.03027	0.02618	1.16	0.2476
1, 3	0.65749	0.7248	-0.06732	0.03295	-2.04	0.0411*
2, 3	0.62722	0.7248	-0.09758	0.03428	-2.85	0.0044**

6.3.3. Análisis de curvas ROC sobre el indicador de puntos totales acumulados

En primer lugar comparamos mediante la curva ROC empírica las curvas ROC para comprarar esta variable con las sumas anteriores. Los resultados (obtenidos con el SPSS v. 10) se presentan en la figura 6.21.

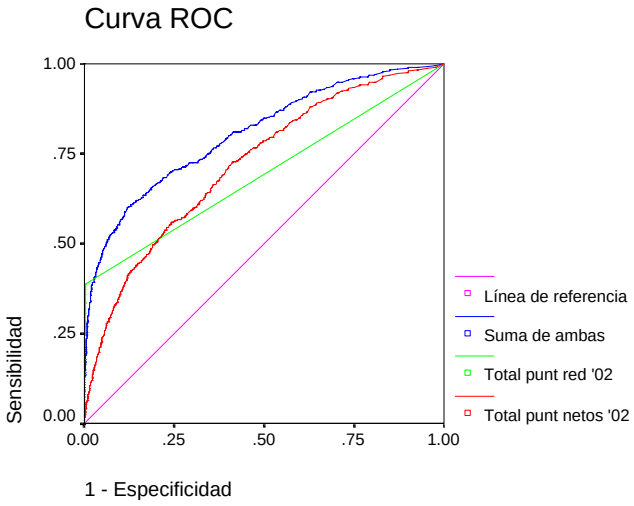


Figura 6.21. Curvas ROC empíricas para los tres indicadores de sumas de puntos

Se ha demostrado que la suma total 3 (TOTTOT3), después de la transformación Ln, presenta una distribución normal. Lo que sigue es un análisis detallado de su relación con la conducta de redención. El hecho de mantener una mayor capacidad predictiva, además de presentar esta forma de distribución, permite aprovechar toda la potencia del

modelo de curva ROC binormal, que mejora la estimación de los parámetros propios de la curva ROC, a lo que se dedica el apartado 6.3.3. a continuación

6.3.3.1. Curva ROC empírica y binormal

En primer lugar puede ser interesante observar la forma de la distribución para los dos grupos, los que nunca han redimido y los que lo han hecho en alguna ocasión. Esto se realiza mediante dos histogramas superpuestos (figura 6.22).

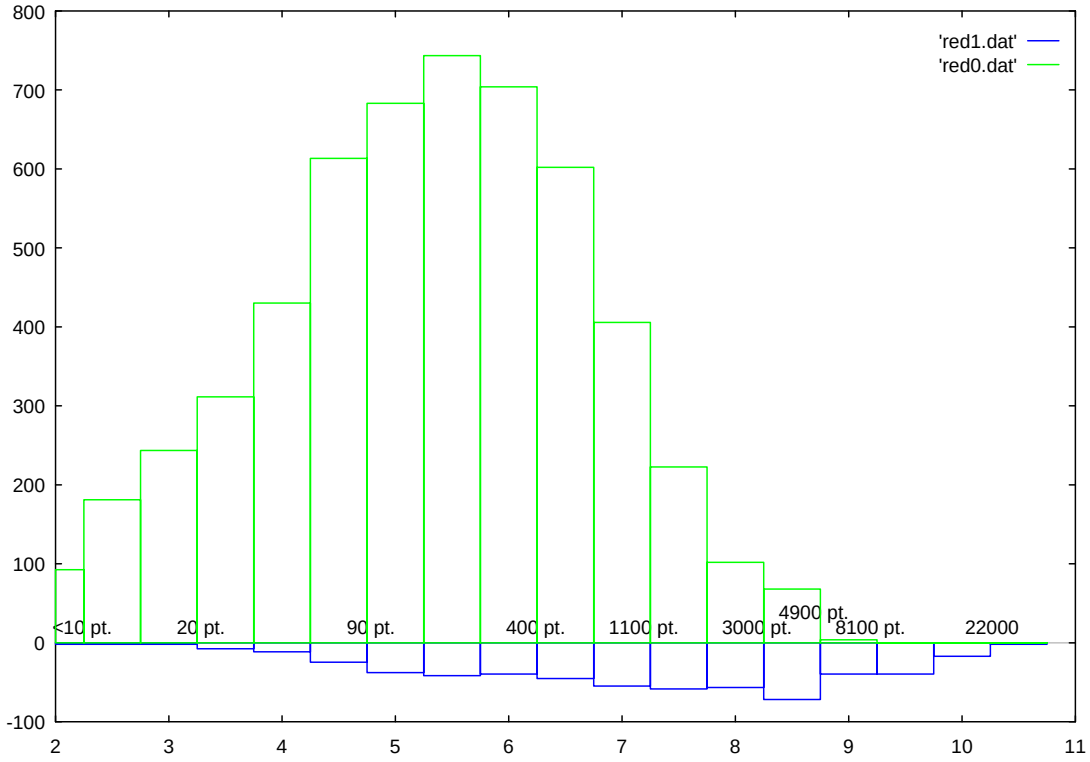


Figura 6.22. Distribución del Ln(total de puntos inc. históricos) en nuestra base de datos (N=5966). Encima del eje de abscisa correspondencia en puntos directos de los valores transformados

La figura 6.22 presenta las distribuciones de frecuencias considerando los dos grupos, los que sí redimen y los que no. Observamos que mientras la distribución del grupo que no redime sigue presentando una distribución prácticamente normal, no es así en el caso del grupo que sí redime, aunque sí que se muestra un perfil de distribución simétrica. Una de las dificultades que señalaremos de este enfoque es el garantizar que las dos distribuciones sean normales, puesto que la solución de realizar una transformación en

una de ellas, si la otra distribución es normal, produce otra vez discrepancias con el supuesto de binormalidad.

En cualquier caso, la forma de la distribución nos garantiza una aproximación suficiente a la curva normal, y por tanto lo usaremos como ejemplo para el cálculo de la curva ROC binormal. Hemos de señalar la ventaja de este enfoque que únicamente requiere de la transformación de la puntuación, y no elimina casos extremos, que para muchos puede suponer una distorsión en los datos.

La estimación de las curvas ROC tanto empíricas como binormales se llevó a cabo mediante el programa NCSS 2004, a la fecha de escribir esta tesis, el único programa completo de estadística que lo realiza.

La curva ROC binormal se presenta en la figura 6.23. superpuesta a la curva ROC empírica. Podemos observar las sutiles diferencias entre ambas, en especial el hecho de que la curva binormal suele proporcionar mayor área bajo la curva. En este caso no está tan claro, pero los resultados en la literatura sugieren que la curva empírica subestima el área bajo la curva.

Como ya hemos visto en el capítulo 4, otra forma de representar la curva ROC es en el espacio definido por las puntuaciones típicas de la distribución normal. Esta representación produce, en el caso de curvas ROC binormales, una recta, y en el caso de las curvas ROC empíricas, una colección de puntos que, en función de su aproximación a distribuciones normales subyacentes, se acercará más o menos a la recta de la ROC binormal. La figura 6.24 a continuación presenta esta representación para el mismo indicador que tratamos antes.

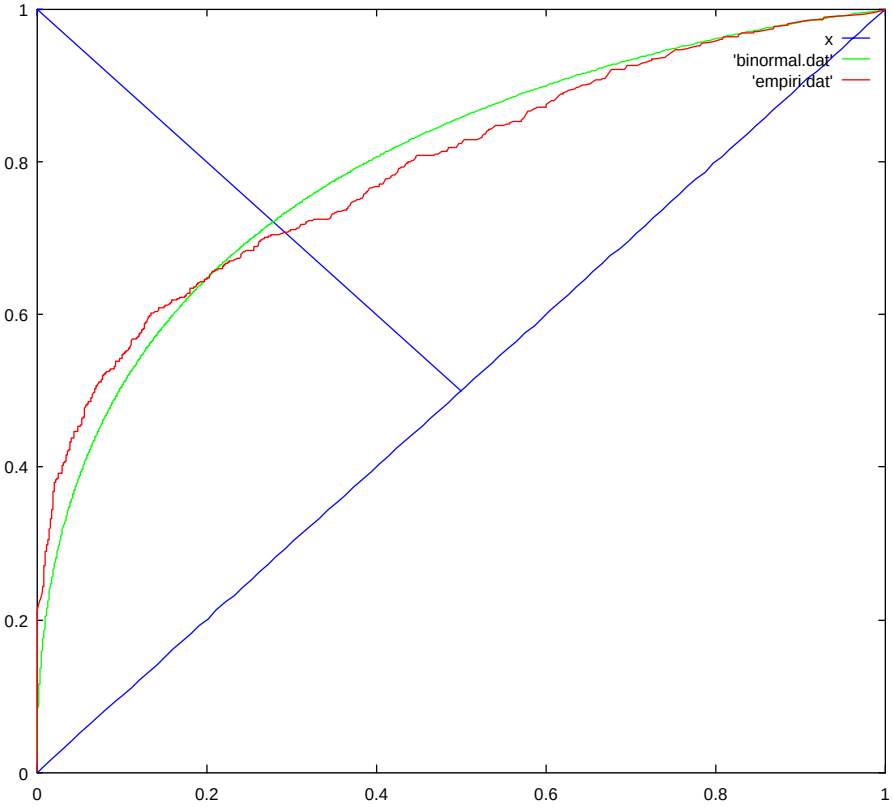


Figura 6.23. Curvas ROC empírica y binormal para el indicador de redención de puntos Ln(total de puntos conseguidos)

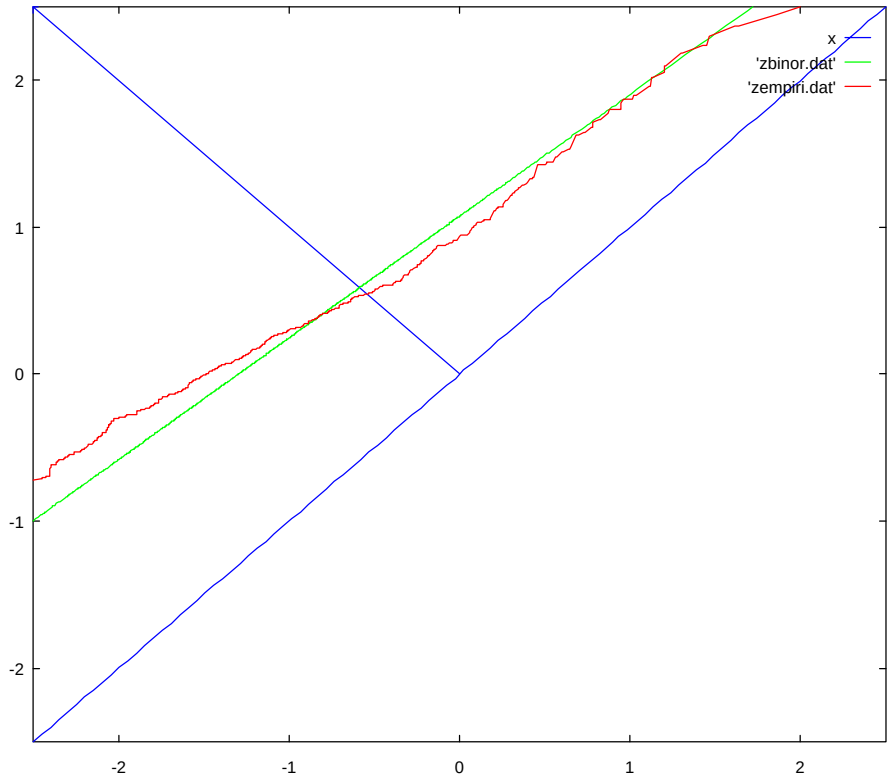


Figura 6.24. "Rectas" ROC empírica y binormal para el indicador de redención de puntos Ln(total de puntos conseguidos) en el espacio definido por las puntuaciones típicas correspondientes a los valores de sens. y (1-espec.)

Las tablas 6.15 y 6.16 muestran los resultados de los contrastes estadísticos para evaluar si este indicador tiene capacidad predictiva significativa. Se puede comprobar que bajo en el enfoque empírico es así (AUC=0.79, z=25.51, p<0.0001)

Tabla 6.15. Resultados del contraste de significación para la ROC empírica con LNTOT3

Criterio	Estimador empírico de AUC	Error típico de AUC	Valor Z para probar AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalencia de REDIME	N
lntot3	0.79150	0.01143	25.51	<0.0001	<0.0001	0.09303	5966

Tabla 6.16. Resultados del contraste de significación para la ROC binormal con LNTOT3

Criterio	Estimador empírico de AUC	Error típico de AUC	Valor Z para probar AUC > 0.5	1-Sided Prob Level	2-Sided Prob Level	Prevalencia de REDIME	N
lntot3	0.79608	0.01049	18.80	<0.0001	<0.0001	0.09303	5966

En el caso de la curva binormal, la tabla de resultados se presenta en la tabla 6.16, y podemos observar que mejora ligeramente el área bajo la curva (0.796 frente a 0.791), y sigue teniendo capacidad predictiva significativamente diferente de 0.5 (z=18.80, p<0.0001).

Las tablas 6.17 y 6.18 presentan los intervalos de confianza para las curvas ROC empírica y binormal, respectivamente, ajustadas para nuestro indicador LNTOT3.

Tabla 6.17. Intervalo de confianza para la ROC empírica con LNTOT3

Criterio	Estimador empírico de AUC	Error típico de AUC	Límite inferior (n.c. 95%)	Límite superior (n.c. 95%)	Porcentaje observado de casos positivos	N total
lntot3	0.79608	0.01049	18.80	<0.0001	<0.0001	0.09303	5966

Tabla 6.18. Intervalos de confianza curva ROC binormal con LNTOT3

Criterio	Estimador empírico de AUC	Error típico de AUC	Límite inferior (n.c. 95%)	Límite superior (n.c. 95%)	Porcentaje observado de casos positivos	N total
lntot3 ROC empírica	0.79150	0.01143	0.76802	0.81286	0.09303	5966

6.3.3.3. Optimización del punto de corte para el modelo de curva ROC ajustado

Siguiendo el siguiente razonamiento: los riesgos, los costes a largo plazo vendrán de personas de gran valor pero que pueden abandonar nuestro programa. Se trata por tanto de detectar aquéllos que NO han redimido sus puntos, aunque por el número de puntos, y en comparación con la población de clientes sí lo deberían haber hecho. Si trasladamos esto a nuestra tabla de doble entrada tendremos lo siguiente:

Tabla 6.19. Razonamiento base para el análisis coste beneficio en la predicción de redención de puntos en una tarjeta de fidelización

		Estado real	
		Sí redime	No redime
Predicción por nuestro indicador o modelo predictivo	Sí redime	Verdaderos Positivos (VP) Persona que responde a nuestro modelo, el coste vendrá de los regalos que haya redimido. Los beneficios podrían ser hasta 1.5 el coste	Falsos Positivos (FP) (Falsas alarmas) Persona que no responde a nuestro modelo. Aunque estemos ahorrando dinero de los regalos, potencialmente estamos perdiendo 1.5 su coste
	No redime	Falsos Negativos (FN) (omisiones) Personas que según nuestro modelo no deberían redimir pero sí lo hacen. No son interesantes desde el punto de vista del programa (habituales o por inercia). Sólo consideraremos el coste directo del regalo.	Verdaderos Negativos (VN) Personas que predecimos que no redimirán y así sucede. Estas personas no contribuyen en nada al beneficio, y por tanto su coste-beneficio es 0 (para los propósitos de nuestro análisis)

Recordemos la fórmula de optimización del punto de corte:

$$\text{sensibilidad} - m\,(1 - \text{especificidad}) \text{ es máximo} \qquad (6.1)$$

Donde a su vez

$$m = \frac{P(condición = \ falso)}{P(condición = verdadero)}\left(\frac{C_{FP} - C_{VN}}{C_{FN} - C_{VP}}\right) \qquad (6.2)$$

A partir también de lo que sabemos de la revisión de la literatura podemos razonablemente esperar que tengamos un 20% de buenos clientes. Estableceremos nuestra "prevalencia" en 0.20. Una vez hecho esto, se trata de resolver la razón de costes a la derecha de 6.2. Para ello podemos suponer:

Tabla 6.20. Estimación de partida para la razón coste-beneficio

		Estado real	
		Sí redime	No redime
Predicción	Sí redime	Cvp = 1.5 - coste del reaglo	Cfp= 1.5 o más
	No redime	Cfn = coste del regalo	Cvn =0

Aplicando la fórmula tendríamos:

$$r_{CB} = \left(\frac{C_{FP} - C_{VN}}{C_{FN} - C_{VP}}\right) = \frac{(C_{FP}) - 0}{\text{coste regalo} - (\text{C}_{\text{VP}} - \text{coste regalo})}$$

$donde\, C_{FP} > 1.5$

$C_{VP} < 1.5$

Con lo que r_{cb} será siempre estrictamente mayor que 1, y jugaremos con los siguientes valores: 1.1, 1.3, 1.5, 1.7, suponiendo que perder un cliente beneficioso nos supone un 10%, un 30%, un 50% o un 70% más que conservarlo.

Pues bien, ¿qué punto de corte en puntos estableceremos para contactar con estos clientes? A partir de la salida del programa NCSS 2004, que, recordemos, permite introducir rangos o valores de la razón coste beneficio según (6.1) y (6.2), se han obtenido funciones de coste que se presentan en las figuras 6.25 y 6.26.

Observamos cómo según la razón es menor, más rápidamente se alcanza un máximo de beneficio. Tengamos en cuenta que hemos fijado la prevalencia en el 20%, esto es, esperamos que en la población un 20% de nuestra base de clientes haya redimido sus puntos más pronto o más tarde.

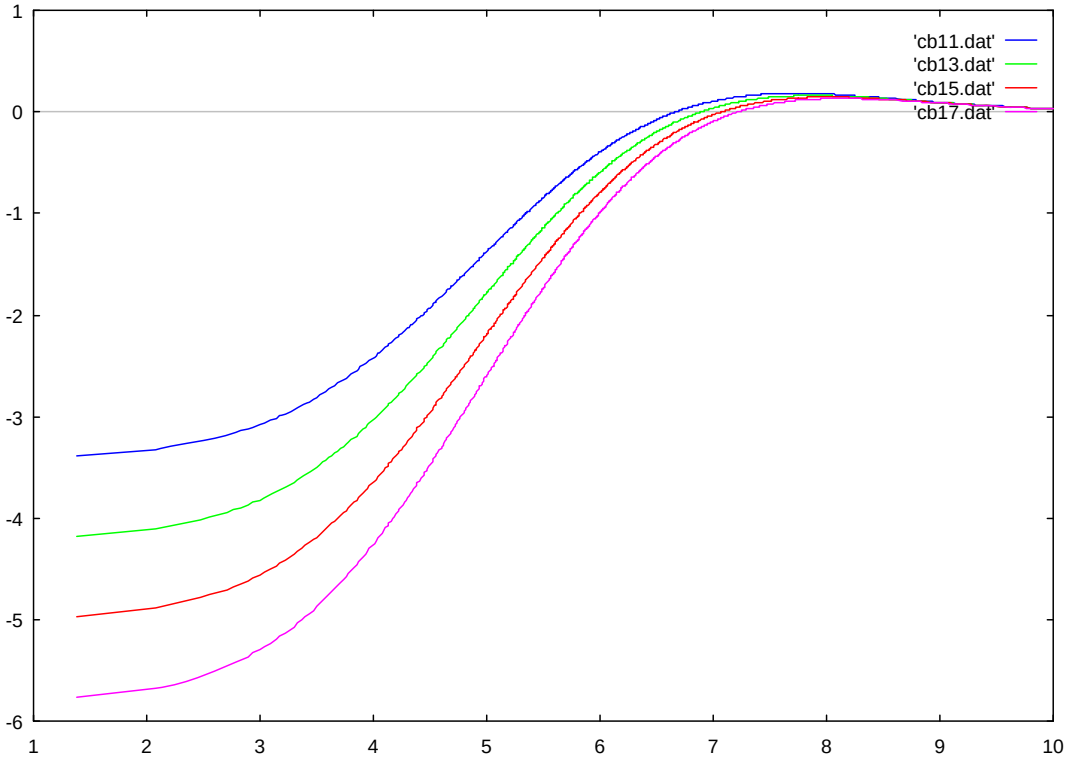


Figura 6.25. Funciones de coste suponiendo cuatro razones diferentes de coste-beneficio (1.1, 1.3, 1.5, 1.7). De superior a inferior en la figura significa de menor (razón) a mayor. En el eje de abscisa, Ln(total de puntos acumulados). En el eje de ordenadas, coste promedio

La figura anterior muestra la relación en función del logaritmo natural del total acumulado de puntos. Podríamos hacer la representación en términos directos de los puntos. Lógicamente no veríamos nada en la escala completa de puntos, pero si utilizamos el rango entre 500 y 5000 puntos obtenemos la representación que se presenta en la figura 6.26. Esta representación es muy útil para comprobar la relación entre el punto de corte y el beneficio promedio total, en función de varias razones de costes. Se ve fácilmente como la función consigue un coste 0 o beneficio más fácilmente cuanto menor sea la razón coste-beneficio.

La tabla 6.21 presenta los puntos máximos para cada una de las razones con las que se han calculado las funciones.

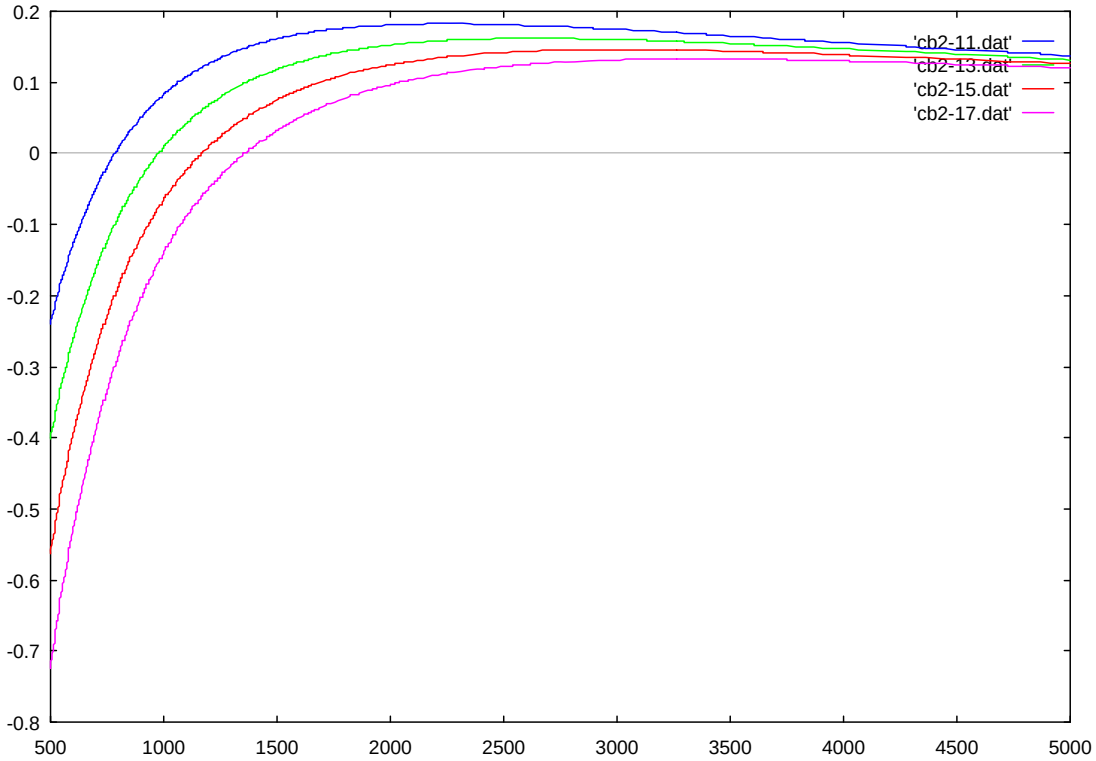


Figura 6.26. Funciones de coste suponiendo cuatro razones diferentes de coste-beneficio (1.1, 1.3, 1.5, 1.7). De superior a inferior en la figura significa de menor (razón) a mayor. En el eje de abscisa, total de puntos acumulados directos. En el eje de ordenadas, coste promedio

Tabla 6.21. Máximos y puntos a partir de los cuales función de costes es positiva

Razón CB	Punto a partir del cual retorno es positivo	Retorno máximo de la función	Puntos en los que consigue máximo
1.1	788	0.18	2298
1.3	982	0.16	2980
1.5	1164	0.146	2892
1.7	1352	0.1331	3294

De la tabla 6.21 se pueden concluir varias cosas:

- El umbral a partir del cual la función es positivo crece rápidamente según aumentamos la razón CB.
- Sin embargo, los máximos de las distintas funciones convergen en unas puntuaciones umbral de en torno a 3000 puntos.

A partir de estos resultados se puede concluir que un punto de corte de 3000 puntos maximizaría el beneficio global de nuestra campaña, supuestas razones CB entre 1.1 y 1.7, y supuesta también una prevalencia de 0.20 en nuestra población.

En cualquier caso, queda mostrada la gran potencia del análisis de curvas ROC con su esquema de coste-beneficio para poder realizar estos cálculos. Todos los autores expertos en curvas ROC señalan que no hay ninguna regla que permita establecer este punto de corte, sino que la decisión se tiene que tomar teniendo en cuenta todos los posibles costes y beneficios. Algo que no hemos podido hacer en este estudio por no disponer de más datos pero que mejoraría mucho este enfoque.

6.3.3.4. Capacidad predictiva del modelo de curva ROC

Por otro lado, debemos considerar la capacidad predictiva. Para ello calculamos las funciones de capacidad predictiva positiva (figura 6.27) que es la que nos interesa, pues queremos detectar positivos.

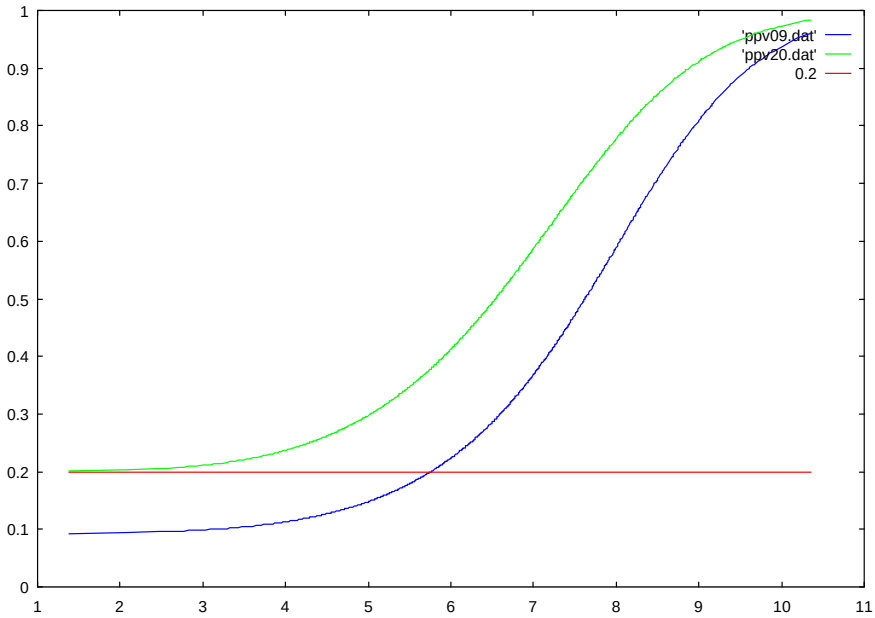


Figura 6.27. Funciones de potencia predictiva positiva. En el eje de abscisa Ln del total de puntos acumulados. En el eje de ordenadas, poder predictivo. La curva superior muestra la función para una prevalencia de 0.20. La función inferior para una prevalencia de 0.09. Como función constante aparece la prevalencia de 0.20.

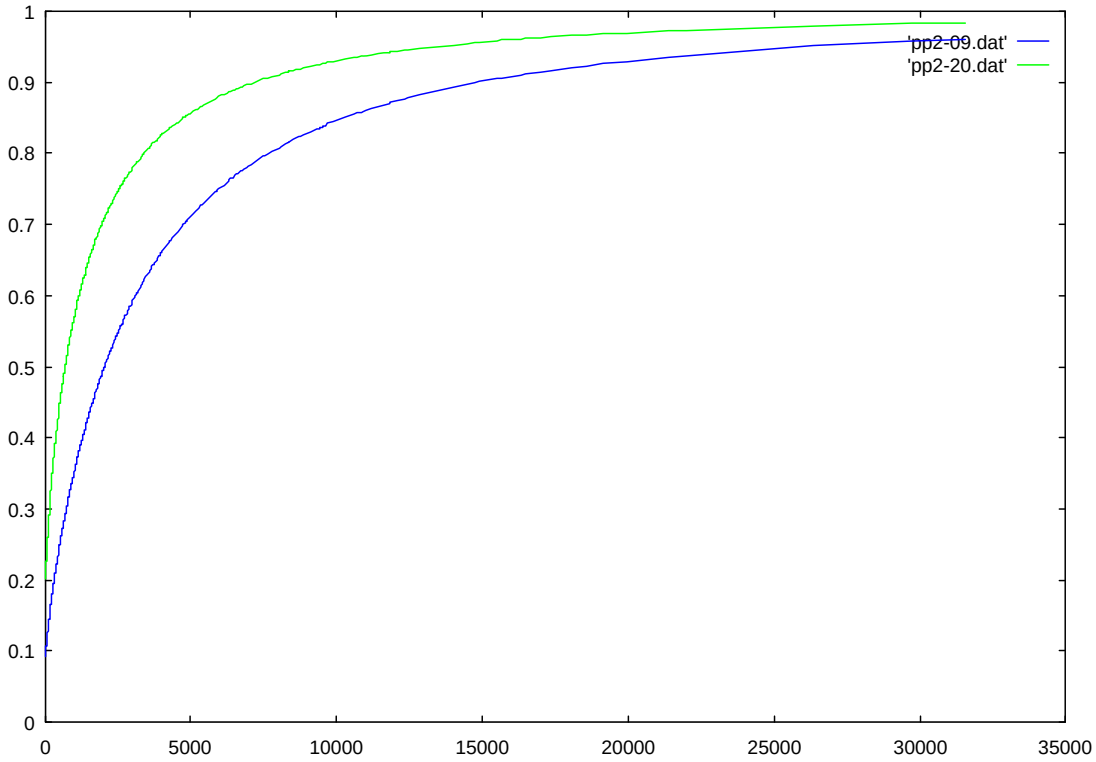


Figura 6.28. Funciones de potencia predictiva positiva. En el eje de abscisa puntos acumulados. En el eje de ordenadas, poder predictivo. La curva superior muestra la función para una prevalencia de 0.20. La función inferior para una prevalencia de 0.09.

Concentrándonos en un intervalo razonable de puntos, entre 500 y 5000, tendremos las funciones que se representan en la figura 6.29. En torno a 3400 puntos tendremos un valor predictivo positivo de 0.8, suponiendo una prevalencia de 0.20. En el caso de tener en realidad una prevalencia de 0.09 (la que obtenemos en nuestra base de datos, suponiendo que fuera representativa de la población total y que fuera un fenómeno más o menos estable), para obtener un valor predictivo positivo equivalente, tendríamos que aumentar a 7600 puntos. Este gran aumento por efecto del cambio en la prevalencia y cómo afecta a la decisión sobre el punto de corte se puede observar en la figura 6.30.

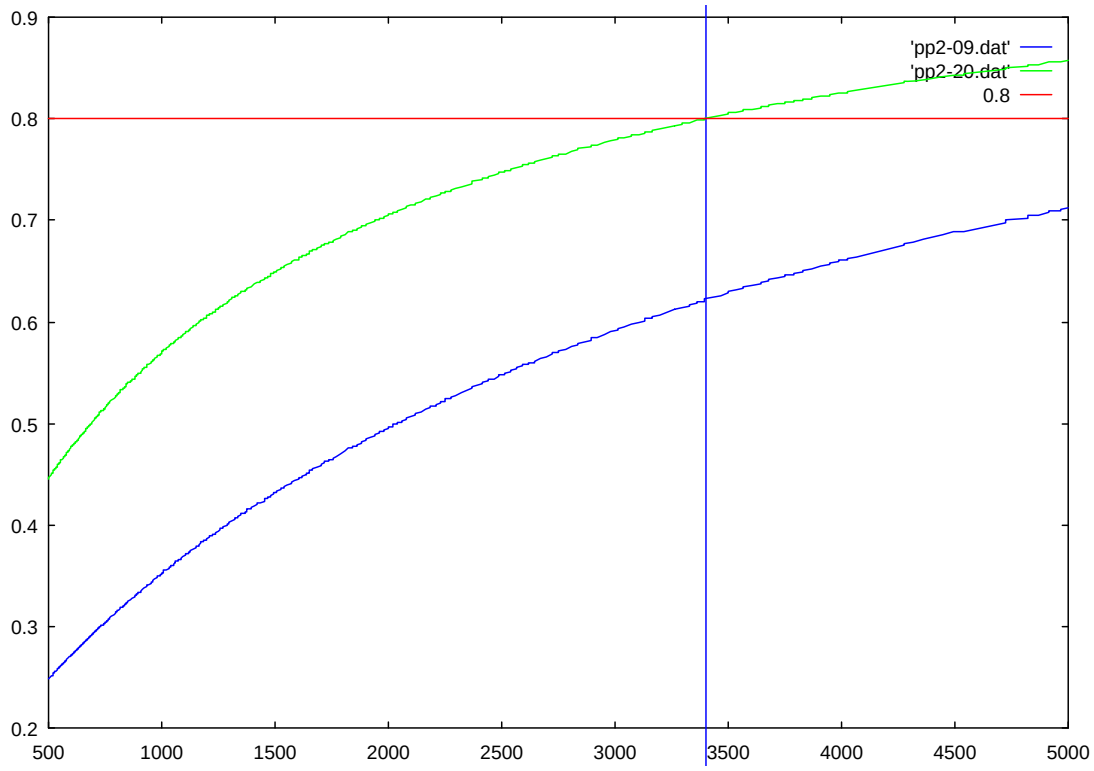


Figura 6.29. Funciones de potencia predictiva positiva para el intervalo entre 500 y 5000 puntos. En el eje de abscisa puntos acumulados. En el eje de ordenadas, poder predictivo. La curva superior muestra la función para una prevalencia de 0.20. La función inferior para una prevalencia de 0.09.

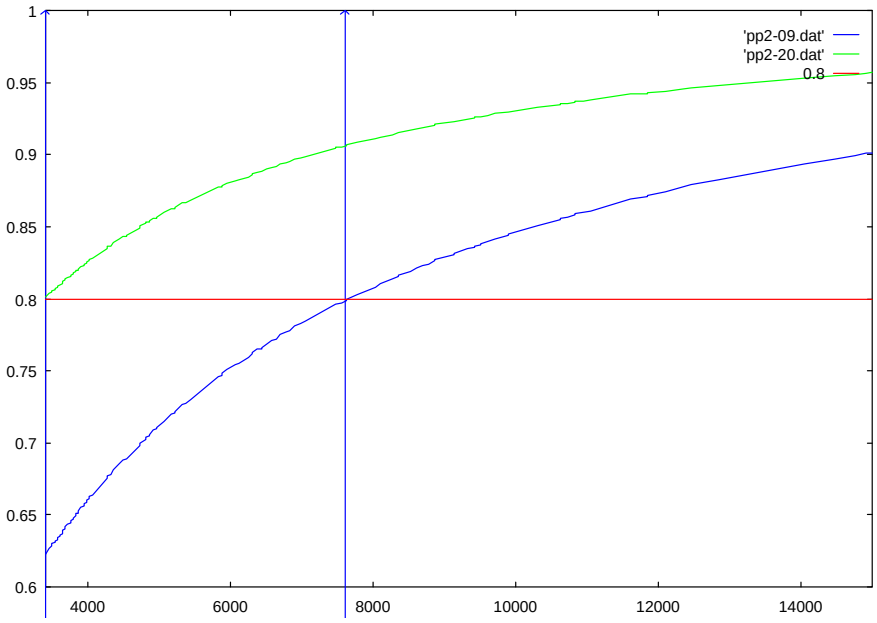


Figura 6.30. Elección del punto de corte a partir de la función de potencia predictiva positiva para una prevalencia de 0.09 y con el objetivo de conseguir poder predictivo de 0.80. En el eje de abscisa puntos acumulados. En el eje de ordenadas, poder predictivo. La curva superior muestra la función para una prevalencia de 0.20. La función inferior para una prevalencia de 0.09.

6.3.3.5. Conclusiones provisionales

En conclusión de todo este largo apartado, podemos establecer que:

- Las curvas ROC binormales, si cumplimos los supuestos de partida para su estimación, son un marco muy potente para la evaluación de la capacidad predictiva a partir de indicadores relativamente simples.
- Las curvas ROC empíricas no tienen esos supuestos tan estrictos, y sirven por tanto para comparar la capacidad predictiva de una gran variedad de indicadores. Además existe un esquema completo de contraste de hipótesis que permite tomar decisiones sobre las diferencias tanto entre indicadores, como entre grupos con un único indicador.

Para el problema que nos ocupa, hemos mostrado que:

- En torno a 3000 puntos podemos establecer un umbral para contactar con los clientes, que maximiza la función de coste beneficio para varias razones r_{cb} >1.1

- Para poder garantizar una capacidad predictiva global de 0.80, necesitamos establecer el umbral en 3400 puntos, pero suponiendo una prevalencia (tasa real de redención en la población) de 0.20. Si esa prevalencia fuera menor, el umbral sería mucho más alto, y en el peor de los casos, en el que la prevalencia fuera la que se ha obtenido en la muestra, podría llegar hasta los 7600 puntos.

6.3.4. Modelos estadísticos predictivos de la conducta de redención

Podemos preguntarnos si no existe un modelo predictivo que nos permita mejorar la capacidad predictiva de las curvas ROC con indicadores sencillos obtenidas en el apartado 6.3.3. Para poder responder a esta pregunta es necesario estimar modelos estadísticos que permitan la inclusión simultánea de varios predictores, puesto que anteriormente hemos estado jugando con un único predictor.

Dentro de los modelos multivariantes para la predicción de resultados dicotómicos, un modelo perfectamente establecido es el de la regresión logística (Hosmer y Lemeshow, 1989).

Por otro lado, mucho más recientemente han surgido con fuerza enfoques computacionales, uno de los cuales presenta muchas ventajas en nuestro contexto: se llaman árboles de decisión, en los que no entraremos en detalle. Una excelente revisión es la de Murthy (1998).

6.3.4.1. Proceso

La estimación de ambos modelos predictivos se ha llevado a cabo con el software Enterprise Miner v.4.1, sobre SAS v. 8.02 para Windows. Se trata de uno de los mejores paquetes software para realizar lo que ha venido en llamarse "minería de datos". Además de procedimientos estadísticos estándar, como modelos lineales de cualquier tipo, incluyendo modelos lineales generalizados, incorpora redes neuronales, razonamiento basado en memoria (aunque señalan que es un módulo experimental), y árboles de decisión.

La interfaz de usuario responde a una metáfora de diagrama de flujo en el que el usuario va poniendo módulos que representan fases del proceso de análisis. La figura 6.31 presenta el diagrama de flujo creado para este análisis en particular, junto con las funcionalidades estándar en este módulo.

En primer lugar es necesario elegir las variables que entrarán en los análisis. Esto se realiza en el primer nodo del diagrama (carga de la base de datos). En el primer paso se pregunta al usuario qué tamaño de la muestra querrá utilizar (esto no es la partición de datos). En este punto, puesto que nuestra base de datos tampoco es tan grande, elegimos utilizar toda la base de datos.

A continuación se incluye el módulo de partición de datos. Se trata de establecer 3 muestras, una de entrenamiento, una de validación, y otra de prueba. Utilizamos la configuración estándar del programa, de 40, 30 y 30% de la muestra, respectivamente. Los algoritmos harán la estimación con la muestra de entrenamiento y la validarán con la correspondiente.

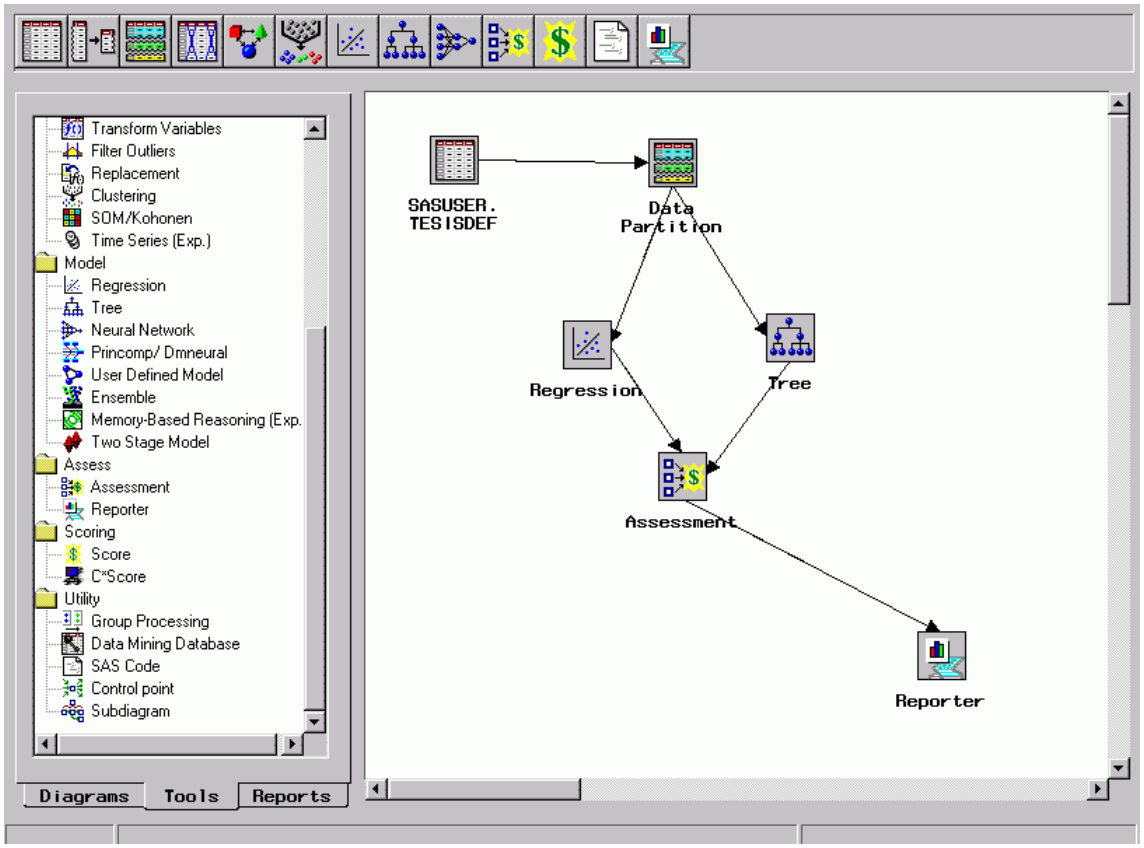


Figura 6.31. Interfaz de usuario del Enterprise Miner de SAS, con el proceso llevado a cabo para este análisis

A continuación se proponen los dos modelos alternativos que vamos a considerar, regresión (logística puesto que nuestra variable objetivo es dicotómica) y árbol de decisión.

Un nodo de evaluación enlazado con la salida de estos dos procedimientos permite evaluar la capacidad predictiva de sus salidas. Por último un nodo de informe permite obtener un conjuntos de ficheros *html* con las salidas de cada uno de los procedimientos.

Las variables que se utilizarán para los modelos aparecen reflejadas en la tabla 6.22. SAS realiza codificación *dummy* automática en el caso de tener variables nominales, o categóricas que no estén codificadas 0, 1. Nuestra variable "*target*" está así codificada, y por tanto los análisis serán específicos para este tipo de variables.

Tabla 6.22. Variables de entrada en el proceso de modelización

Nombre	Rol en el modelo	Tipo de var
ONLINE	input	binary
TARJ_ADI	input	binary
LNTOT3	input	interval
NOMESINA	input	interval
TOTBBV	input	interval
TOTERK	input	interval
TOTPAT	input	interval
TOTRPS	input	interval
TOTTLF	input	interval
VISBBV	input	interval
VISERK	input	interval
VISRPS	input	interval
VISTLF	input	interval
CDICE	input	nominal
CODCCAA	input	nominal
HMLNUM	input	nominal
SEGPMP	input	nominal
SEXO	input	nominal
REDIME	target	binary

6.3.4.2. Regresión logística

El procedimiento para la estimación de una regresión logística en SAS se denomina "*DMREG Procedure*". Utiliza "*Dual Quasi-Newton Optimization*", y "*Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)*".

La optimizacion convergió en 17 iteraciones y se cumplió el criterio de convergencia.

Los resultados de ajuste del modelo se muestran en la tabla 6.23 y los parámetros significativos del modelo en la tabla 6.24

Tabla 6.23. Resultados de ajuste del modelo de regresión logística

Fit Statistic	Training	Validation	Test
Akaike's Information Criterion	1111,3184023	,	,
Average Squared Error	0,0551377013	0,0629335398	0,0600882153
Average Error Function	0,2116283469	0,2439525316	0,2241802151
Degrees of Freedom for Error	2368	,	,
Model Degrees of Freedom	45	,	,
Total Degrees of Freedom	2413	,	,
Divisor for ASE	4826	3620	3618
Error Function	1021,3184023	883,10816452	811,08401832
Final Prediction Error	0,0572333065	,	,
Maximum Absolute Error	0,998764697	0,9996794302	0,9996464644
Mean Square Error	0,0561855039	0,0629335398	0,0600882153
Sum of Frequencies	2413	1810	1809
Number of Estimate Weights	45	,	,
Root Average Sum of Squares	0,2348141846	0,2508655812	0,2451289769
Root Final Prediction Error	0,2392348354	,	,
Root Mean Squared Error	0,2370348158	0,2508655812	0,2451289769
Schwarz's Bayesian Criterion	1371,8065753	,	,
Sum of Squared Errors	266,0945464	227,81941413	217,39916296
Sum of Case Weights Times Freq	4826	3620	3618
Misclassification Rate	0,0683796104	0,0745856354	0,0696517413
Total Profit for REDIME	224	167	170
Average Profit for REDIME	0,0928305015	0,0922651934	0,0939745716

Tabla 6.24. Componentes en el modelo de regresión logística estimado

Parámetro		g.l.	Estimador	Error Típico estimador	Wald Chi-square	Pr > Chi-square		Estimador tipificado	exp(Est)
Intercept		1	-88.528	31.660	7.82	0.0052	.		0
LNTOT3		1	12.102	0.0985	150.87	<.0001		1.043.314	3.354
ONLINE	0	1	-0.8054	0.1222	43.47	<.0001	.		0.447
SEGPMP	3	1	-12.073	0.2921	17.09	<.0001	.		0.299
SEGPMP	4	1	13.673	0.4427	9.54	0.002	.		3.925
HMLNUM	1	1	0.5236	0.1826	8.22	0.0041	.		1.688
NOMESINA		1	-0.0515	0.0267	3.71	0.0541		-0.096121	0.95
VISERK*		1	-0.0123	0.00647	3.6	0.0576		-0.137652	0.988

*Esta variable ya no entra en el modelo

Hay que aclarar que ONLINE es la variable dicotómica de estar registrado en línea, SEGPMP 3 compara el código 3 (los monopatrocinadores BBV) de esta variable con la última categoría, esto es, con la 8 (los de perfil más multipatrocinador), y SEGPMP 4 es igulamente la comparación entre el segmento 4 (monopatrocinadores Telefónica) y el 8.

De los resultados podemos observar cómo Lntot3 es una de los predictores más significativos. Luego encontramos diferencias muy significativas por ONLINE, pero no por tarjetas adicionales. HMLNUM 1 compara los de la categoría 1 con los de la categoría 3. Número de meses inactivo tiene una contribución significativa al modelo. Pero la siguiente variable en la lista (VISERK, visitas a Eroski) ya no entra en el modelo.

6.3.4.3. Árbol de decisión

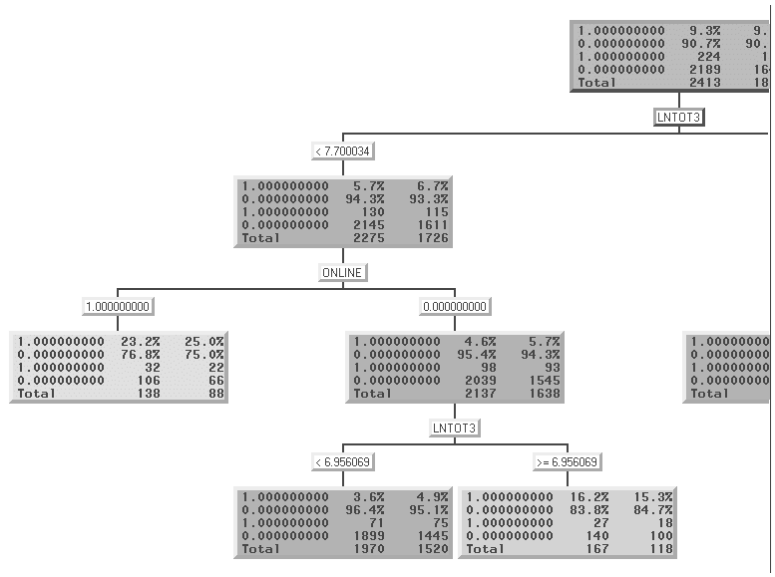
Murthy (1998) recoge en un largo artículo el estado del arte en ese momento de los algoritmos conocidos como "árboles de decisión", desde una perspectiva multidisciplinar.

Los orígenes de los árboles de decisión en estadística comienza por la necesidad de explorar datos de encuestas. A partir de ese momento, aparecen algoritmos como AID,

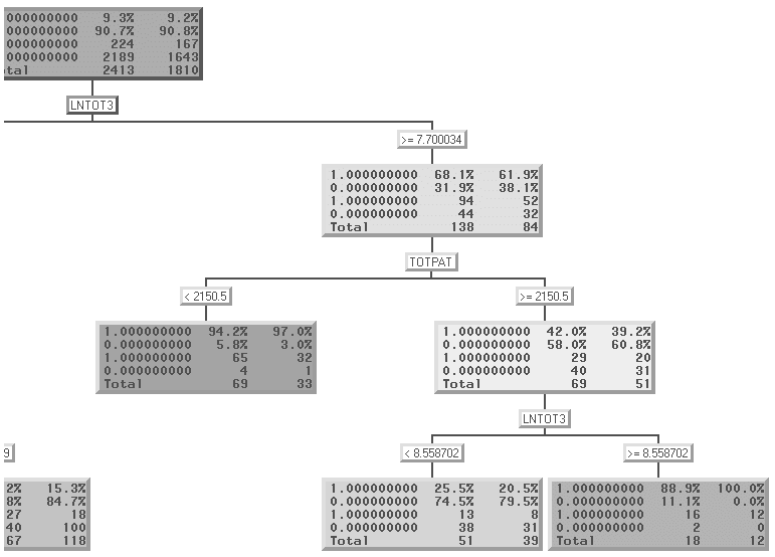
MAID, THAID y CHAID que construyen árboles de segmentación binaria cuyo objetivo es hacer visibles las interacciones entre variables predictoras y dependientes.

Los árboles de decisión son una técnica para representar reglas que están oculta en los datos, mediante estructuras jerárquicas, secuenciales, que hacen particiones en los datos de forma recursiva.

El árbol que ajusta SAS Enterprise Miner puede observarse en la figura 6.32 (a) -lado izquierdo - y 6.32(b)- lado derecho.



(a)



(b)

Figura 6.32. Árbol de decisión ajustado por SAS Enterprise Miner, lados izquierdo (a) y derecho (b)

Se puede observar cómo en el primer paso, el árbol distingue entre los que tienen por debajo de 7.7 en lntot3. Esto supone un punto de corte de 2200 puntos acumulados. A continuación distingue entre los que están suscritos *online*, y los que no, y en éstos divide a su vez los datos entre los que tienen por debajo de 6.95 en LNTOT3. En este punto supone 1000 puntos.

Para los que tienen más de 2200 puntos acumulados, entonces el árbol distingue entre quienes tienen acumulados 2100 puntos o más en diferentes patrocinadores, y a continuación distingue entre quienes tienen más de 5100 puntos.

Como podemos ver, aunque útiles, las reglas que genera un árbol de decisión implican varios puntos de corte, en función de diferentes categorías que son las que van formando los nodos. No parece claro cómo utilizar el análisis coste-beneficio con este procedimiento.

6.3.4.4. Comparación entre los dos modelos predictivos

SAS proporciona un mecanismo de comparación de modelos basado en curvas ROC, pero meramente visuales. No tenemos estimación del área bajo la curva, ni mucho menos contraste estadístico entre los resultados de los modelos. La figura 6.33 muestra las curvas ROC que proporciona Enterprise Miner. Podemos observar que la regresión logística tiene mayor capacidad diagnóstica que el árbol de decisión, pero no podemos decidir con un contraste estadístico.

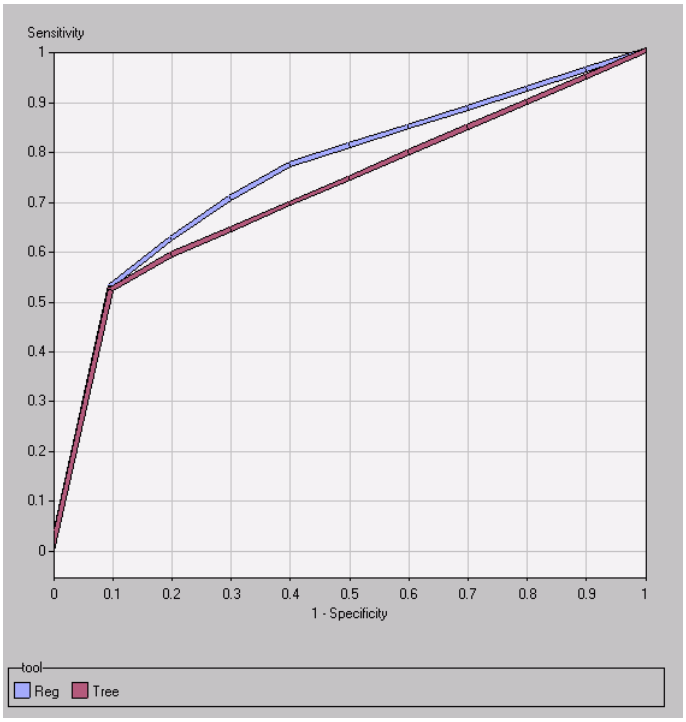


Figura 6.33. Curvas ROC empíricas para evaluar la capacidad discriminativa de la regresión logística y el árbol de decisión.

En relación con este resultado, ya sabíamos que los heurísticos de división ambiciosa "*greedy*" son eficientes y adecuados para muchas aplicaciones, pero son esencialmente subóptimos, aunque existen técnicas para mejorar la inducción. Los árboles individualmente tienen alta varianza en términos de precisión para su generalización, así que muchos autores han sugerido la combinación de resultados de múltiples árboles. Los árboles producen fragmentación en los datos, que reduce la significación probabilística de los nodos más cercanos a los extremos.

A partir de estos resultados continuaremos nuestro análisis para comparar el modelo predictivo de regresión logística con nuestro modelo sencillo basado en un único predictor.

6.3.5. Comparación de la capacidad predictiva del modelo de regresión logística y del modelo de indicador único.

Una vez hemos llegado a este punto nos podemos preguntar, ¿con qué modelo nos quedamos, con uno sencillo basado en datos agregados de puntos, o con uno más

complicado -el de regresión logística- ? ¿Podemos tomar una decisión sobre cuál de los dos predice mejor? ¿Podemos ajustar un modelo binormal a la puntuación de la regresión logística, directamente?

Las curvas ROC nos proporcionan la respuesta. En la figura 6.34 observamos la comparación de las curvas ROC empíricas producidas por el modelo de regresión logística y por nuestro indicador agregado lntot3. Podemos observar cómo mediante la regresión logística mejoramos en capacidad predictiva. La tabla 6.25 presenta los resultados de los contrastes estadísticos que confirman que la regresión logística tiene mayor capacidad predictiva que sólo lntot3 (que también forma parte del modelo de regresión logística).

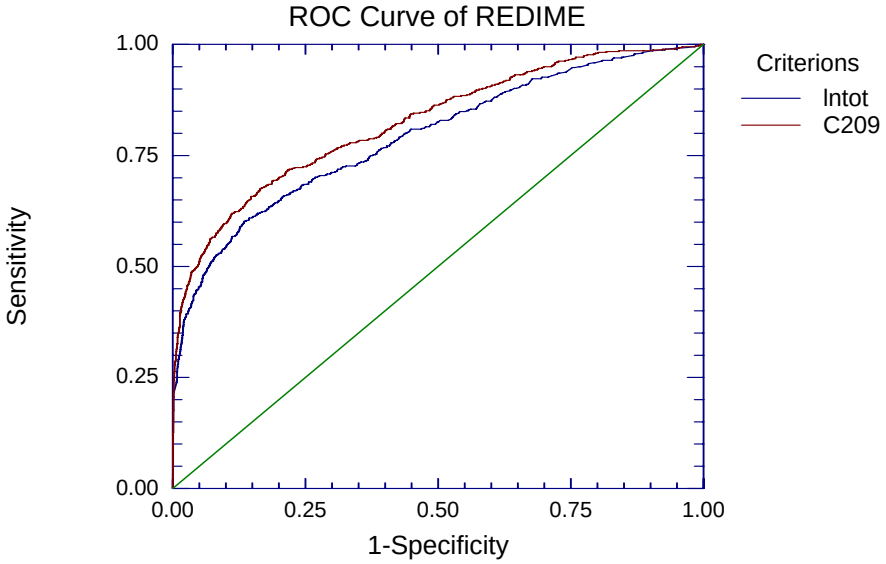


Figura 6.34. Curvas ROC empíricas para evaluar la capacidad discriminativa de la regresión logística y el indicador lntot3.

¿Podemos utilizar el modelo de ROC binormal para esta comparación? No, puesto que la puntuación que obtenemos del modelo de regresión logística no sigue, por definición, una distribución normal. Debiéramos transformarla. Pero es que además hemos perdido la simplicidad de nuestro primer modelo, esto es, partir de una puntuación sencilla que nos permitía establecer un punto de corte independiente de clasificaciones, combinaciones de variables, o cálculos complejos para obtener la puntuación a partir del modelo de regresión logística. Por esto mismo, aunque podemos llevar a cabo el análisis de coste-beneficio que habíamos realizado anteriormente, el resultado será proponer

puntos de corte, no ya sobre la escala de puntos acumulados directos, sino sobre una escala de comprensión mucho más difícil, como es la puntuación que se obtiene a partir de una regresión logística.

Tabla 6.25. Contraste estadístico entre las áreas bajo las curvas ROC de modelo 1 predictor y regresión logística

Criterio	Estimador empírico de AUC1	Estimador empírico de AUC2	Diferencia	E.T. diferencia	Z	Nivel prob
lntot3-score	0.79150	0.82438	-0.03287	0.00533	-6.17	<0.0001

6.4. Conclusiones parciales

Es éste el momento de extraer conclusiones de todos nuestros análisis, en particular ver si las hipótesis de partida que habíamos hecho al principio de este capítulo se sostienen después de haber realizado los análisis.

En cuanto a la hipótesis 1, sobre las curvas ROC y su capacidad de evaluar la capacidad predictiva de indicadores individuales:

- (H1a) Hemos sido capaces de tomar decisiones sobre la capacidad predictiva de distintas variables de nuestra base de datos, de una manera estándar para todas ellas y con una gran potencia y simplicidad. De este modo hemos sido capaces de identificar un valor agregado (una simple suma de puntos) con una capacidad predictiva muy importante.
- (H1b y c) Hemos sido capaces de encontrar un indicador con capacidad predictiva estadísticamente significativa, y hemos podido tomar decisiones entre alternativas mediante el enfoque de curva ROC empírica y contrastes estadísticos basados en la ROC no paramétrica.
- (H1d) Después de realizar las transformaciones necesarias sobre la variable agregada que hemos identificado como más predictiva, hemos realizado curvas ROC según el modelo binormal, aumentando la capacidad predictiva de nuestro modelo. Aun así, reconocemos que este enfoque tiene la gran limitación de cumplir el supuesto de distribuciones binormales que se solapan, que puede ser muy difícil de cumplir en entornos aplicados, y sobre todo en aquellos casos en que el hecho positivo sea muy raro, puesto que nos será más difícil cumplir este supuesto.
- (H1e) Hemos podido establecer un punto de corte que optimice el beneficio esperado. Sin embargo, no hemos sido capaces de estimar suficientemente los costes y beneficios para obtener toda la potencia del modelo, y hemos tenido que realizar simulaciones con varias razones de costes-beneficios.

En cuanto a la hipótesis 2, hemos podido estimar un modelo de regresión logística y un modelo de árbol de decisión, y

- (H2a) Hemos podido compararlos en capacidad predictiva utilizando una forma muy básica de curva ROC, según la proporciona el paquete estadístico sobre el que se han estimado tanto la regresión logística como el árbol de decisión. Pero no hemos sido capaces de hacer contraste estadístico entre el área bajo la curva de cada modelo, puesto que esta funcionalidad no la incorpora el programa estadístico. Por tanto, hemos tomado una decisión a partir de la curva ROC dibujada. No hemos podido extraer las puntuaciones probabilísticas a partir del árbol de decisión de tal manera que pudieran ser comparadas con el rendimiento de la variable identificada en el paso 1. Sí que hemos podido hacer con el modelo de regresión logística, y...
- (H2b) Hemos comparado el modelo de regresión logística con el modelo de un único predictor identificado en el paso 1, llevando a cabo el contraste de hipótesis estadísticas, pero no hemos podido realizar esta comparación con el modelo binormal, puesto que la salida del modelo de regresión logística no cumple el supuesto de distribuirse normalmente.