5 Medir la precisión o efectividad de un sistema diagnóstico: índices de eficacia

En Medicina, el objetivo de un sistema de diagnóstico es muy a menudo discriminar entre dos estados mutuamente exclusivos, tales como la presencia o ausencia de una enfermedad a partir de unas pruebas médicas o decidir si en una placa de rayos X se observa un tumor o no. Esta tarea de detección y decisión es mucho más general y no está en absoluto restringida al campo médico. Por ejemplo, en Psicología Clínica, tendríamos que decidir si se diagnostica a un un jugador como jugador patológico o no a partir de un protocolo o prueba psicológica específica. Y en investigación de mercados, podríamos tener que detectar quiénes de nuestros clientes pueden estar a punto de abandonarnos, o quiénes de entre nuestros clientes responderán a una promoción.

A lo largo de los capítulos anteriores hemos presentado múltiples ejemplos en varios campos, en particular en el capítulo 4 hemos puesto como ejemplo las pruebas de detección temprana o diagnóstico rápido APACHE II y III, y un instrumento de detección del maltrato infantil.

El problema de evaluar la precisión de las herramientas diagnósticas (de cualquier tipo, desde una simple prueba a todo un protocolo de evaluación) requiere primero definir qué entendemos por los conceptos básicos de precisión, eficacia, capacidad predictiva, capacidad diagnóstica.

Nuestro objetivo en este capítulo será primero definir los conceptos de precisión y eficacia diagnóstica, y revisar los índices propuestos para su medida, y su relación con el análisis de curvas ROC.

¿Qué es la eficacia diagnóstica de un sistema?

Tomemos un ejemplo de Medicina (al fin y al cabo no cabe duda de que es el campo en que todos estos conceptos y modelos están más desarrollados): Altman (1991) comenta el caso de la clasificación de diagnósticos a partir de una prueba de mamografía por dos radiólogos. Podemos comparar estas clasificaciones mediante una tabla de doble entrada en la que tendríamos todos los posibles cruces y el número de coincidencias. Se trataría entonces de estimar el acuerdo entre jueces. El estadístico adecuado es *kappa*, que proporciona un valor de 1 para indicar un acuerdo perfecto y 0 para indicar la falta total de acuerdo, no más allá del azar. También puede tomar valores negativos, que indican un acuerdo peor incluso que el propio por azar (discrepancias completas en la clasificación).

Altman (1991) proporciona la siguiente regla simple y básica para clasificar los índices de acuerdo obtenidos (positivos):

Tabla 5.1. Clasificación de la fuerza del acuerdo a partir del estadístico kappa,según recoge Altman (1991)

Valor de kappa	Fuerza del acuerdo
<0.20	"Pobre" o Malo
0.21 - 0.40	Mediano o Regular
0.41 - 0.60	Moderado
0.61 - 0.80	Bueno
0.81 - 1.00	Muy bueno

Al igual que sucede con la correlación de Pearson y otros estadísticos, la reducción de toda la riqueza de los datos a un único número proporciona un mecanismo de comparación entre estudios (supuestos diseños de estudios que permitan esta comparación), pero cuya interpretación puede ser en muchos casos arbitraria.

¿Pero estaríamos midiendo la capacidad predictiva? En el ejemplo anterior hablamos del acuerdo entre clasificadores (ya sean humanos o el resultado o salida de un algoritmo de clasificación). En muchos casos, el diagnóstico implica la detección de un

caso positivo frente a un caso negativo, esto es, una decisión entre sólo dos valores. En ese caso, la evaluación de la eficacia del sistema diagnóstico implica estudiar la capacidad para predecir o detectar el resultado verdadero, esto es de discriminar los verdaderos positivos y los verdaderos negativos. Podemos pensar entonces en los índices básicos de sensibilidad y especificidad tal y como los definíamos en (4.2) y (4.4).

En el caso de la práctica clínica a la que se refiere a menudo Altman (1991) -"El único interés de estudiar un test diagnóstico es usarlo para hacer diagnóstico"-, son los valores predictivo positivo y negativo (4.13) y (4.14) los que resultan significativos, para lo que es impresdincible hablar de la prevalencia.

Altman concluye que, dado que los valores predictivos finales dependen de la prevalencia, y además todo el esquema de evaluación (basado en sensibilidad y especificidad) se sustenta sobre el conocimiento del estado real, "no debiéramos tomar los valores predictivos observados en una muestra como aplicables universalmente". De hecho una de las grandes dificultades de cualquier estudio será conocer la prevalencia en la población, a los que por supuesto habrá que añadir los ya mencionados del coste de cada uno de los casos producto de la decisión, según se recoge en (4.10). Otra enorme dificultad desde el punto de vista práctico algunas veces, pero en otras desde el punto de vista teórico, es determinar la certeza de que los verdaderos positivos lo son. Por ejemplo, ¿podremos saber alguna vez que una persona acusada de un delito efectivamente lo ha cometido? ¿O que una persona, o el estado de un automóvil, o de una carretera, han sido en realidad el o la responsable de un accidente?

Podemos ver cómo el cálculo de nuestros índices, estén basados en la curva ROC o en otro enfoque, es sólo el primer paso de un conjunto de decisiones razonadas, ya sea para estimar un punto de corte o para comparar diferentes instrumentos diagnósticos.

El objetivo de este capítulo será, en primer lugar, definir qué entendemos por eficacia o capacidad predictiva de un sistema diagnóstico. A continuación revisaremos los diferentes enfoques propuestos en varios ámbitos, y los compararemos con el enfoque de la curva ROC que estamos proponiendo. Esta información se basa en el excelente artículo de J. A. Swets (1986) para *Psychological Bulletin*, y se complementa con un

apartado específico sobre las medidas de eficacia de algoritmos de recuperación de información, que tienen mucho que ver con las medidas de las curvas ROC.

Finalizamos este capítulo con un apartado sobre los gráficos de elevación, o *lift-charts*, puesto que son una herramienta muy generalizada en Marketing, y que al menos visualmente se parecen mucho a las curvas ROC, aunque no son iguales.

5.1 Medidas de un sistema diagnóstico

Swets y Picket (1982) en su clásico libro "*Evaluation of Diagnostic Systems*" describen 4 formas de evaluar un sistema diagnóstico: precisión, eficacia, fidelidad y consistencia. Comenzando por estos dos últimos:

La fidelidad de un sistema diagnóstico se puede evaluar mediante patrones de prueba construidos *ad-hoc*, de tal modo que indiquen hasta qué punto el sistema representa los resultados esperados. La evaluación podría proporcionar los artefactos producidos en el sistema, de alguna manera podríamos tener una “función de transferencia” del sistema. Es éste uno de los objetivos de las pruebas de diagnóstico por la imagen que hemos comentado en el capítulo 4.

La consistencia sería una propiedad relacionada con el hecho de tener diferentes evaluadores, decisores o momentos de decisión o diagnóstico, y se podría expresar como el porcentaje de ensayos en los que los juicios coinciden, o también se podría calcular como un índice de correlación. Es el objetivo del ejemplo del índice *kappa* que hemos expuesto en la introducción a este capítulo.

La falta de consistencia determinaría que la falta de precisión de un sistema diagnóstico o predictivo provendría del error en los procesos de toma de decisiones, y Swets señala que puede servir como una indicación de límite superior de precisión.

La precisión de un sistema diagnóstico, a diferencia de la consistencia, intenta reflejar el valor diagnóstico de la información que se transmite. Esto es, no se trata de evaluar las diferencias entre decisores o sistemas, sino extraer una medición “pura” de la

capacidad de un sistema para servir de base a la toma de decisiones o a la clasificación o al diagnóstico. Se trata también de proporcionar una base para estimar el valor práctico que puede proporcionar la información que genera el sistema.

Por último, **la eficacia** de un sistema sería una propiedad global, relacionada con los beneficios, riesgos, y costes del diagnóstico. La gran dificultad para poder determinar la eficacia de un sistema en la práctica será la de asignar valores a la información o a los resultados que maneje el sistema.

Swets y Picket (1982) finalizan estas distinciones con una tabla de las cuatro propiedades, según su valor diagnóstico y valor práctico:

Tabla 5.2. Propiedades de cuatro posibles medidas de un sistema diagnóstico

	Valor diagnóstico	**Valor práctico**
Fidelidad	Medidas de relevancia cuestionable	No proporciona medida
Consistencia	Mide sólo el valor potencial	No proporciona medida
Precisión	Mide el valor real	Mide sólo el valor potencial
Eficacia	Mide el valor real	Mide el valor real

En cierto modo, lo que vienen a sugerir Swets y Picket es que la eficacia recoge todas las propiedades anteriores. Sin embargo, necesitaremos conocer los beneficios y costes de la aplicación del sistema diagnóstico, y todo lo relacionado con la aplicación del mismo para poder estimarla. En este estudio abordaremos la medida de la eficacia diagnóstica, si nos es posible, y en cualquier caso abordaremos la precisión, tal y como está definida antes. Todo el empeño de J. A. Swets a lo largo de su dilatada carrera ha sido demostrar que la curva ROC supone la mejor forma de medir la precisión de un sistema diagnóstico y que además proporciona el mejor punto de partida para poder medir la eficacia de su aplicación en la práctica.

El apartado 5.2 es un resumen de la recopilación del mismo Swets, publicada en 1986 en *Psychological Bulletin* en la que compara diferentes índices de medida con los que proporciona la curva ROC.

5.2 Curva ROC y otros índices de medida de un sistema diagnóstico

Supongamos que disponemos de un indicador predictivo o diagnóstico, y que nuestra decisión debe hacerse sobre dos posibes estados únicamente (ausencia o presencia de enfermedad, aceptación de una oferta o no, abandono de una suscripción o no). El punto de partida, para diagnósticos o decisiones sobre estos dos posibles estados, es en la mayoría de los casos una tabla de doble entrada como la expuesta en la tabla 4.1.

Todos los datos en la tabla dependen del punto de corte en la escala del indicador que establezcamos. Para cada punto de corte tendremos unos resultados diferentes con respecto a la predicción que hagamos. En su momento ya establecimos que hay dos medidas que permiten obtener una representación completa del comportamiento diagnóstico. Estas dos medidas son la sensibilidad y la especificidad.

Ambas medidas están relacionadas, y dependen a su vez del punto de corte o umbral que se establezca a partir del indicador cuantitativo, de tal manera que si la regla de decisión impuesta establece un umbral muy bajo para la decisión, tendremos alta especificidad pero baja sensibilidad, y si, por el contrario establecemos un umbral alto tendremos al contrario, alta sensibilidad pero baja especificidad.

Ambos indicadores dependen, pues, de la regla de decisión que se establezca, esto es, del umbral o punto de corte en la función o modelo estimado a partir del cual los sujetos serán diagnosticados como positivos o negativos.

Cómo resume cada punto de la curva ROC todos los elementos de la tabla tomados en probabilidades:

Swets y Picket (1982) muestran cómo cada punto de la curva ROC (que, recordemos, es el correspondiente a un punto de corte en particular) resume toda la información en la tabla de clasificación. Efectivamente, cada punto (x, y) representa:

$$x = P(\text{detectado como positivo} \mid \text{verdadero positivo}) \qquad (5.1)$$

que se podría estimar por a/(a+c), e

$$y = P(\text{detectado como positivo} \mid \text{verdadero negativo}) \qquad (5.2)$$

que se puede estimar por b/(b+d).

Y todas las restantes probabilidades condicionales en la tabla son sus complementos. La primera es la sensibilidad y la segunda (1-especificidad). Por tradición con sus orígenes en Psicofísica, también hablaremos de los términos "*hit*" (éxito) y falsa alarma. De tal manera que P(*hit*)=sensibilidad, y P(falsa alarma)= (1-especificidad).

Para realizar las comparaciones que Swets propone con otros índices de evaluación de sistemas diagnósticos, se utiliza la siguiente notación:

h= *hits* (éxito) ó Verdaderos Positivos (VP) (5.3)

f= falsa alarma ó Falsos Positivos (FP) (5.4)

N= total de casos (5.5)

s=(a+c)/N (tasa de prevalencia) (5.6)

Esta última definición requiere aclarar que este indicador es poblacional. Por tanto, para que sea obtenido de nuestros datos de frecuencias en la tabla de doble entrada, debe tratarse de una muestra suficientemente representativa. O, también, puede estar obtenido por medios ajenos al estudio que estemos llevando a abo.

También estableceremos las siguientes relaciones:

$$a/(a+c)=h \quad (5.7)$$

$$(a+c)/N=s \quad (5.8)$$

$$a=hsN \quad (5.9)$$

$$b/(b+d)=f \quad (5.10)$$

$$(b+d)/N= 1 -s \quad (5.11)$$

$$y \;\; b = f(1 - s)N \quad (5.12)$$

Swets (1986) expone en primer lugar seis índices que implican modelos de umbral fijo (tabla 5.3). Swets señala su inadecuación para el objetivo de medir la precisión diagnóstica de un sistema, puesto que están sujetos a falta de fiabilidad o inestabilidad, al depender de la decisión (más o menos arbitraria) de establecer un umbral fijo para tomar la decisión.

Tabla 5.3. Definiciones y fórmulas en términos de componentes ROC para índices que implican un modelo de umbral fijo

Nombre del índice y símbolo más usual	Definiciones y su equivalencia en términos de componentes de curva ROC
1. Probabilidad corregida de éxito H_c	$H_c = [a/(a+c)] - [b/(b+d)]/[1-[b/(b-d]] =$ $= (h-f)/(1-f)$ (5.13) $h = H_c + f(1-H_c)$
2. Probabilidad corregida de éxito H'_c	$H'_c = (ad-bc)/(a+c)(b+d)$ $= h - f$ (5.14) $h = H'_c - f$
3. Proporción de correctos PC	$PC = (a+d)/N$ $= (1-s)(1-f) + sh$ (5.15) $h = [PC - (1-s)(1-f)]/s$
4. Prueba de capacidad (*skill test*) Z	$Z = 4(ad-bc)/N^2$ $= 4s(1-s)(h-f)$ (5.16) $h = f + [Z/4s(1-s)]$

Tabla 5.3. (cont.) Definiciones y fórmulas en términos de componentes ROC para índices que implican un modelo de umbral fijo

Nombre del índice y símbolo más usual	Definiciones y su equivalencia en términos de componentes de curva ROC
5. Estadístico *kappa* K	$K = \frac{2(ad-bc)}{2(ad-bc)+N(b+c)}$ $= \frac{2s(1-s)[h-f]}{(1-2s)[hs+f(1-s)]+s}$ $h = \frac{f(1-s)[1-(1-K)(1-2s)]-sK}{s[1+(1-K)(1-2s)]}$ (5.17)
6. Coeficiente phi ϕ	$\phi = (ad-bc)/[(a+c)(b+d)(a+b)(c+d)]^{1/2}$ $= \frac{[(1-s)s]^{1/2}[h-f]}{[(sh+(1-s)f)(1-(sh+(1-s)f))]^{1/2}}$ (5.18) $h = [\phi^2 + 2(1-s)(1-\phi^2)]f +$ $+ \left[\phi[\phi^2 + 4(1-s)(1-f)/s]^{1/2}\right]/[2((1-s)+s\phi^2)]$
7. RIOC (*Relative Improvement over chance*) - Loeber y Dishion (1983)	$RIOC = \frac{PC-s}{\left(1-\left\lvert\frac{a}{a+b}\right\rvert\right)-s}$ (5.19) *donde PC es la proporción de correctos*

Los dos primeros índices listados son dos formas de "probabilidad corregida de éxito". Nótese que tienen el mismo nombre (hemos comprobado que es así tanto en el artículo original de 1986 como en la reimpresión del mismo en 1996), y sólo se distinguen en el símbolo en un simple apóstrofo. Ambos se concentran en la tasa de éxito (proporción de aciertos entre el total de casos detectados), y ambos intentan corregirlo de componentes espúreos que puedan, por ejemplo, inducirlos hacia las falsas alarmas, representado en las fórmulas por f.

Tabla 5.4. Índices Hc y H'c para un intervalo de puntuaciones de factor 2 del instrumento de maltrato

Puntuación	h	f	Hc	H'c
41.00	0.87	0.18	0.84	0.69
42.00	0.86	0.17	0.84	0.69
43.00	0.85	0.17	0.81	0.68
44.00	0.83	0.16	0.80	0.67
45.00	0.82	0.16	0.79	0.66
46.00	0.80	0.15	0.76	0.65
47.00	0.78	0.14	0.75	0.64
48.00	0.77	0.14	0.74	0.64
49.00	0.76	0.13	0.73	0.63
50.00	0.74	0.12	0.71	0.62

El primer índice, Hc, normaliza el rango del valor corregido, y Swets indica que se utilizaba normalmente en estudios de funciones sensoriales. El segundo índice H'c se ha utilizado en memoria, y también en predicción del tiempo.

El índice "proporción correctos" (también denominado "porcentaje de correctos") es el más viejo de todos. Swets señala que se propuso en 1884 para medir la predicción de tornados, y todavía se usa en muchos campos, incluyendo ese mismo de predicción meteorológica, así como en diagnóstico médico. Ya desde 1885 se sabe que este índice depende enormemente de s, de la tasa base o prevalencia, y PC puede llegar a ser tan alto como la propia tasa base, o (1-s) por puro azar, sin discriminar. Este hecho muchas veces no se tiene en cuenta, y se sigue utilizando alegremente, sobre todo en la evaluación de la capacidad predictiva de modelos estadísticos o de minería de datos, quizá porque al final muchas de las pruebas tratan de detectar "tanto como sea posible" sin tener en cuenta que tan importante como detectar verdadero positivos es hacerlo con el mínimo de falsas alarmas.

El índice RIOC aparece mencionado por Paul Barret en sus apuntes disponibles en Internet (http://www.pbarrett.net) y lo hemos incluido en la tabla. Se trata de otro índice relacionado con el porcentaje de correctos, que hace unas correcciones basadas en la tasa base o prevalencia.

En el campo meteorológico, existe un índice que intenta medir hasta qué punto la discriminación se produce por encima del azar, y se denomina "prueba o puntuación de capacidad" (*skill test*). En la tabla 5.5 mostramos estos dos índices para el mismo rango que antes para poder hacer comparaciones.

El estadístico *kappa* se conoce bien en Medicina clínica y otros ámbitos como medida de acuerdo entre observadores. Se trata en suma de otra forma de corrección de la proporción total de correctos. Por último, el índice Phi, o su raíz cuadrada, se han utilizado como índices de precisión discriminativa en Psicología Experimental, predicción del tiempo y pruebas de calidad no destructivas. Todas las referencias se pueden consultar en Swets (1986, 1996).

En suma podemos ver que todos estos índices varían enormemente en función de sus correcciones por azar o por otros motivos. El principal inconveniente que ofrecen como medida de precisión de todo el sistema (supongamos por un momento que sólo disponemos del intervalo de ejemplo del factor 2 de detección de maltrato) es que dependen totalmente del punto de corte o umbral elegido, con lo cual nunca tendremos una única medida. En función de esa decisión, que puede ser más o menos arbitraria, tendremos un valor diferente.

Tabla 5.5. Índices PC, puntuación de capacidad, phi y kappa para un intervalo de puntuaciones de factor 2 del instrumento de maltrato

Puntuación	pc	skill test	phi	kappa
41.00	0.83	0.59	0.65	0.64
42.00	0.84	0.60	0.65	0.64
43.00	0.84	0.59	0.64	0.64
44.00	0.84	0.58	0.64	0.63
45.00	0.84	0.58	0.64	0.63
46.00	0.83	0.57	0.63	0.62
47.00	0.84	0.56	0.63	0.62
48.00	0.84	0.56	0.62	0.62
49.00	0.84	0.55	0.62	0.62
50.00	0.84	0.54	0.61	0.61

Existen más indicadores de este tipo de acuerdo entre jueces, y resulta muy recomendable el software DICHOT 3.0, disponible en internet de forma gratuita en la el sitio *web* de Paul Barrett, que calcula unos cuantos de ellos, además de proporcionar una buena ayuda con referencias a sus autores. La figura 5.1 muestra una imagen de la pantalla de este software en el que aparecen los cálculos para el punto de corte 41 de nuestro ejemplo. En resumen, todos ellos dependen del punto de corte y por tanto no son válidos para nuestro propósito de tener una medida única de precisión diagnóstica.

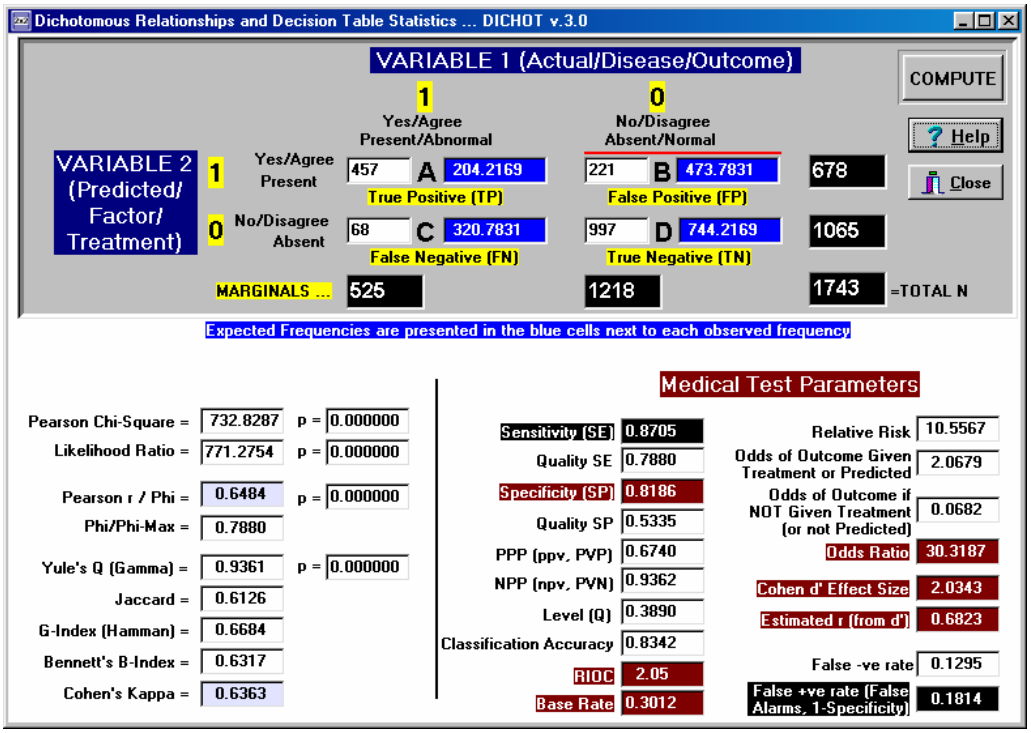


Figura 5.1. Cálculo de índices de acuerdo y otros para un umbral fijo con el software DICHOT 3.0

La tabla 5.6 lista cuatro índices que son consistentes con modelos con umbral o punto de corte variable, esto es, recogen en una única medida la capacidad diagnóstica según se varía el umbral o punto de corte. Su principal problema es que requieren parámetros distribucionales fijos, esto es, requieren normalidad de las distribuciones subyacentes, así como igualdad de varianzas.

El primero de ellos es el clásico *d'* que proviene de la TDS, como ya hemos mencionado. Se define en términos de las puntuaciones (típicas, o z) de la distribución normal. El valor de *d'* es la puntuación z correspondiente a las falsas alarmas, menos la puntuación z correspondiente a los éxitos o aciertos.

Tabla 5.6. Definiciones y fórmulas en términos de componentes ROC para índices que implican un modelo de umbral variable

Nombre del índice y símbolo más usual	Definiciones y su equivalencia en términos de componentes de curva ROC
7. Detectabilidad d'	$d' = z_{b/(b+d)} - z_{a/(a+c)} =$ $= z_f - z_h$ (5.19) $z_h = z_f - d'$
8. Medida de la teoría de la elección (*choice theory*) η	$\eta = (bc/ad)^{1/2} =$ $= [[f(1-h)]/[h(1-f)]]^{1/2}$ (5.20) $h = f/[f + \eta^2(1-f)]$
9. Logaritmo de la razón de *odds* LOR	$LOR = \ln(ad/bc) =$ $= \ln[h(1-f)/f(1-h)]$ (5.21) $h = f/[f + e^{-LOR(1-f)}]$ $\eta^2 = e^{-LOR}$ (5.22)
10. Yule's Q Q	$Q = (ad - bc)/(ad + bc) =$ $= (h - f)/(h - 2fh + f)$ (5.23) $h = \dfrac{f}{\left[f + \dfrac{1-Q}{1+Q}(1-f)\right]}$

El índice η fue propuesto por Luce en su teoría general de la elección. Tanto este índice como el logaritmo de la razón de *odds* (LOR) son similares a *d'*, pero dependen para su cálculo de distribuciones logísticas, más que normales o gaussianas.

El índice LOR resulta muy interesante y se usa muy frecuentemente. Los *odds* (no encontramos una buena traducción española, además de que en este caso el término está tan unido al concepto que resulta inútil proponer una) serían como apuestas, el número de veces que teóricamente se tendría un resultado correcto (ad) frente a una elección incorrecta (bc). Paul Barrett, citado antes, ofrece una excelente exposición de lo que significan los *odds*: los *odds* de la ocurrencia de un evento E son O(E)=P(E)/(1-P(E)), esto es, la razón de su probabilidad de ocurrencia frente a la ocurrencia de todos los demás eventos menos el suyo. Por ejemplo, si la tasa base de asistencia a un médico por tener una gripe es, para la población de un determinado sitio, 0.20, eso significa que tenemos un *odds* de 0.25, ó 25:100, ó de 4:1, de que un paciente que entre por la puerta

de una consulta tenga la gripe. También se puede expresar de la siguiente manera: tenemos 4 veces más probabilidad de que un paciente que entre tenga la gripe que que no la tenga.

La fórmula (5.22) establece la relación que existe entre el índice logarítmico de la razón de *odds* y el de la teoría de la elección de Luce.

El índice Q de Yule fue propuesto por su autor en 1912. Este índice fue propuesto por como un índice de precisión en Psicología por Nelson (1984), porque sus supuestos parecen menores que los de los otros índices. Sin embargo, Swets (1986) muestra que este índice es consistente con (aunque no supone) distribuciones logísticas subyacentes.

En cuanto a estos índices que asumen umbrales variables, se puede mostrar que, teóricamente, hay una relación uno-a-uno y a su vez con el valor de d' tomado en la diagonal negativa de la curva ROC, denominado d'_e. En la práctica, las aproximaciones son muy buenas para $d'_e < 2$, y se pueden corregir fácilmente para valores mayores.

¿Cuál es el propósito de Swets en su larga y compleja exposición de estos índices? Mostrar que los cuatro índices tienen raíces teóricas comunes, y que en una mayoría de condiciones, llevarán a las mismas conclusiones en la práctica.

Swets también aclara que no hay ningún enfoque que no haga un conjunto previo de supuestos, o que esté libre de tener un modelo subyacente, y que el modelo más general, puesto que todos los demás se pueden expresar con él, es el la curva ROC. De hecho, Swets sugiere que un índice candidato se puede evaluar dibujando la familia de ROCs que implica. Se dibujan las curvas que conectan los puntos en los cuales el índice tiene un valor constante, y de ese modo se obtiene la ROC implícita en el índice para ese valor. Todo el propósito de Swets es mostrar que un índice sólo puede ser válido, y tiene alta probabilidad de ser fiable, si sus ROCs implícitas tienen la misma forma que las ROCs empíricas que se hayan encontrado para el problema de discriminación (observador, tarea o contexto) en cuestión. Sin embargo, no dice cómo se puede hacer esa comprobación.

5.3 Ventajas y desventajas del análisis de curvas ROC

Una vez que sabemos que el área bajo la curva ROC es, quizá, el mejor índice de precisión de un sistema diagnóstico, ¿qué otras ventajas tiene? Según Swets y Picket (1982) el análisis de curvas ROC:

1. Proporciona un índice de eficacia diagnóstica puro, cualquiera que sea el criterio o punto de corte en el indicador cuantitativo en que se basa la decisión, e incluso independientemente de que dicho indicador esté sesgado. Esta medida, que es el área bajo la curva ROC, resulta extremadamente útil para poder elegir una técnica entre varias en competencia.
2. Estima la probabilidad de diferentes resultados en la tabla de 4 resultados posibles. De este modo, permiten aislar causas de error en sistemas diagnósticos (por ejemplo, falsas alarmas y omisiones).
3. Proporciona la base para la decisión sobre un criterio de decisión (punto de corte o umbral para la toma de decisión), que permite incluir probabilidades (incluso estimaciones subjetivas de la probabilidad) y costes o utilidades. Este método permite establecer razonadamente reglas o mecanismos óptimos de toma de decisión dentro de un sistema de diagnóstico.

¿Qué desventajas podemos señalar? Aquí debemos volver a la distinción entre curvas ROC empíricas y "binormales". Hemos de señalar que el trabajo mostrado sobre las curvas ROC "binormales" es únicamente un subconjunto de las propuestas que existen en este tema. Hemos expuesto el enfoque más aplicable, pero por ejemplo, Charles Metz sigue desarrollando un modelo de curvas ROC "binormales propias" (véase, Metz y Pan, 1999), que requiere a su vez de determinados programas para su estimación (disponibles de este autor bajo petición).

Las curvas ROC empíricas son un instrumento muy poderoso, en tanto en cuanto son muy fáciles de calcular, proporcionan un enfoque completo de contrastes estadísticos que es aplicable en una gran cantidad de situaciones, por sus pocos supuestos paramétricos, y proporcionan, como dice Swets, una medida única de la precisión diagnóstica de un sistema. Sin embargo, tienden a subestimar sistemáticamente esta

capacidad predictiva y sus índices (especificidad y sensibilidad) pueden ser muy variables en función de las condiciones de nuestro estudio.

Las curvas ROC binormales son más potentes, pero ya imponen supuestos que pueden ser muy difíciles de cumplir, e incluso de comprobar, en condiciones reales de un estudio. Sin embargo, proporcionan índices más estables, además de permitir un conjunto de contrastes estadísticos, que aunque equivalente a las no paramétricas, resulta más potente.

5.4 Uso de las curvas ROC para la medida de la eficacia de un sistema diagnóstico

Mc. Fall y Treat (1999) hacen referencia al artículo clásico de Meehl y Rosen (1955) quienes identificaron tres problemas clásicos que complican mucho la tarea de evaluar el valor de la información presente en un indicador diagnóstico conjuntos de datos. El primero de estos problemas es el de elegir el punto de corte óptimo para diferenciar entre dos posibles grupos. Ya en ese momento señalan que en todos los casos en los que haya que elegir un punto de corte para tomar una decisión entre dos posibles estados finales o decisiones, es inevitable tener que hacer un intercambio entre sensibilidad y especificidad en función de los cambios en el punto de corte entre las dos distribuciones.

El segundo problema identificado por Meehl y Rosen (1955) fue el problema de la tasa base. Brevemente, la potencia discriminatoria de una medida particular variará en función de la tasa base de la variable que se intenta predecir en la población que se evalúa. Ya hemos mostrado en el capítulo 4 cómo diferentes tasas base, o prevalencia en la población, producen variaciones muy importantes en la capacidad predictiva de la prueba.

El problema de elegir un punto de corte en esta situación dependerá de la tasa base. Para un punto de corte fijo, los índices de sensibilidad y especificidad no cambiarán, pero la utilidad práctica de la medida cambiará como resultado de los cambios en la tasa base o prevalencia del desorden. El desequilibrio entre la razón de probabilidades entre

los dos grupos significa que cuando las dos distribuciones se superponen, los errores de clasificación serán mucho más frecuentes para el caso con mayor prevalencia.

El tercer problema identificado por Meehl y Rosen (1955) y Meehl (1973) fue la falacia lógica de usar distribuciones normativas de probabilidad para tomar decisones de naturaleza idiográfica o para hacer predicciones. A esto también se lo denomina el problema de la probabilidad inversa. Este problema surge cuando el evaluador confunde dos tipos de probabilidad (adaptando la notación y el concepto a nuestro enfoque):

p(S|s), la probabilidad de tener una puntuación dada en una prueba diagnóstica, dado que se pertenece al grupo de verdaderos positivos,
y
p(s|S), la probabilidad de ser verdadero positivo, dada una puntuación determinada en una prueba diagnóstica.

El problema de la probabilidad inversa interactúa con el de la tasa base. Cuando la tasa base, o prevalencia, de un desorden en la muestra es exactamente 0.5, entonces p(S|s)= p(s|S). Cuando la prevalencia no es 0.5, sin embargo, estas dos probabilidades no son iguales. Y además, cuanto más diferentes sean estas tasas de prevalencia, mayor será la desigualdad.

Meehl y Rosen (1955) mostraron cómo el teorema de Bayes soluciona el problema mediante el control para las tasas base. El teorema de Bayes es:

$$p(Hipótesis \mid Datos) = \frac{p(Hip)^* p(Datos \mid Hip)}{p(Datos)} \qquad (5.24)$$

o, utilizando nuestra notación:

$$p(s \mid S) = \frac{p(s)^* \, p(S \mid s)}{p(S)} \qquad (5.25)$$

El enfoque bayesiano permite solucionar dos de los tres problemas señalados por Meehl y Rosen, el de la tasa base y el de las probabilidades inversas, pero no el primero de los expuestos, el de los puntos de corte. Seleccionar puntos de corte óptimos siempre

requiere juicios subjetivos sobre cómo mejor resolver los inevitables intercambios entre sensibilidad y especificidad, o entre los costos y beneficios relativos de los diferentes tipos de errores de clasificación. Dado que las decisiones sobre los puntos de corte nunca pueden estar libres de valor, no hay una fórmula mágica para encontrar un punto de corte óptimo y válido para todo propósito, de tal modo que cada fórmula para los puntos de corte está basada en supuestos o valores asumidos.

Dado un conjunto de datos y una tasa base en la población, por ejemplo, podríamos elegir una puntuación de corte que maximizara el porcentaje correcto total, sin embargo, esta elección supone que la solución óptima debiera asignar pesos iguales a los dos tipos de error (por ejemplo, falsos positivos y falsos negativos). A menudo los costes de estos dos errores son desiguales, sin embargo. Mc. Fall y Treat (1999) ponen como ejemplo cómo, para prevenir terrorismo aéreo, la sociedad tolera unas tasas muy altas de falsos positivos y amenaza a todo el mundo como terroristas potenciales, porque la sociedad impone un mayor valor en asegurar la tasa más alta posible de verdaderos positivos (aunque probablemente, como se ha demostrado en los atentados del 11 de septiembre de 2001, la proporción de omisiones era demasiado alta).

Mc. Fall y Treat destacan que, aunque el artículo de Meehl y Roses ha circulado durante medio siglo, ha tenido poco impacto en la práctica de la evaluación clínica en Psicología, a diferencia de en la Medicina como hemos visto. Estos autores proponen estudiar la precisión de los estimadores de probabilidad (prevalencia, etc.) mediante el análisis sistemático de la información de diferentes pruebas para que pueda ser aplicado por los evaluadores.

Desafortunadamente, sin embargo, el enfoque bayesiano proporciona una solución incompleta a las necesidades del evaluador clínico. Todavía no proporciona una métrica común con la que cuantificar el valor de la información de los datos de evaluación que sea independiente de los cambios en los puntos de corte y en las tasas de prevalencia. Dicha escala estándar es esencial si los evaluadores desean comparar la validez incremental, o utilidad relativa, de diferentes métodos de evaluación. Esta medida vendrá dada, como hemos mostrado, por la curva ROC, que hemos mostrado que cumple el requisito impuesto de ser una propiedad únicamente del método, no de la

prevalencia de la característica en la muestra a la que el método se aplicó, o de los sesgos de decisión o elecciones de criterios de los evaluadores que usen el método.

Cómo utilizar la curva ROC para medir la eficacia diagnóstica

La curva ROC básica tiene muchas ventajas, pero entre ellas no está el tomar en cuenta la prevalencia de la enfermedad. Esto es, los índices de sensibilidad y especificidad que definen la curva ROC no dependen ni varían con la prevalencia. Hemos mostrado cómo el análisis de curvas ROC proporciona el mejor esquema para realizar el análisis coste-beneficio, pero ¿cómo interactúa con las necesarias correcciones para tener en cuenta la prevalencia en la población? Este problema es especialmente importante en el caso de tener tasas base muy bajas, puesto que, como hemos visto en el capítulo 4, deberemos incrementar el punto de corte para garantizar un valor equivalente de poder predictivo que si tuviéramos tasas base más altas.

Ilustraremos algunas de estas dudas mediante uno de los mejores ejemplos disponibles, como es el diseño de las pruebas ELISA ("enzyme-linked immunosorbent assay") para la detección del VIH. Altman presenta este ejemplo a partir de los resultados de Weiss *et al.*, (1985):

Tabla 5.7. Resultado de la clasificación en pruebas de detección del VIH (Weiss et al., 1985, recogido en Altman, 1991)

Razón de la absorción media de par de muestras dividido por absorción media de 8 controles negativos	Donantes sanos	Pacientes con SIDA
<2.0	202 (68%)	0 (0%)
2.0 - 2.99	73 (25%)	2 (2%)
3.0 - 3.99	15 (5%)	7 (8%)
4.0 - 4.99	3 (1%)	7 (8%)
5.0 - 5.99	2 (1%)	15 (17%)
6.0 - 11.99	2 (1%)	36 (41%)
>= 12.0	0 (0%)	21 (24%)

A partir de estos datos se pueden calcular especificidad y sensibilidad, pero también necesitamos como hemos visto, los valores predictivos positivo y negativo, para lo que

Altman utiliza dos razones de prevalencia para "seropositividad al HIV" (básicamente desconocida en la población, y mucho más en la época en que se hizo el estudio), 1% y 10%.

Tabla 5.8. Valores predictivos positivo y negativo utilizando dos valores de prevalencia de seropositividad VIH en la población (Weiss et al., 1985, recogido en Altman, 1991)

			Prevalencia de seropositividad HIV = 10%		**Prevalencia de seropositividad HIV = 1%**	
Punto de corte	Sensibilidad	Especificidad	Valor pred. positivo	Valor pred. negativo	Valor pred. positivo	Valor pred. negativo
2	1.00	0.68	0.26	1.00	0.03	1.00
3	0.98	0.93	0.59	0.997	0.12	0.9997
4	0.90	0.98	0.81	0.99	0.28	0.999
5	0.82	0.99	0.87	0.98	0.38	0.998
6	0.65	0.99	0.91	0.96	0.49	0.996
12	0.24	1.00	1.00	0.92	1.00	0.992

Altman señala que no hay razón para usar la prevalencia del estudio, un 23% (propiamente no es prevalencia, sino la proporción de casos VP en la muestra de estudio) porque se trata de un estudio de validación de una prueba. Además, los sujetos se han reclutado de forma completamente independiente.

Nuestro objetivo en estas pruebas será detectar cuantos más verdaderos positivos mejor, por razones obvias. Por tanto, desearemos maximizar el valor predictivo positivo. Podemos observar cómo en el caso de la prevalencia más baja, del 1%, para tener un valor predictivo positivo razonable tendremos que establecer el punto de corte en 12 (1 frente a 0.49 que se obtiene con 6), mientras que en el caso de una tasa de prevalencia del 10% podríamos llegar a tener un valor predictivo positivo más razonable con el punto de corte en 6 (0.91)[1].

[1] En los días en que se escribió esta tesis apareció publicada la prevalencia de seropositividad o infección por VIH en España, que es del 0.3% de la población, entre 120 y 150.000 personas. Datos del Plan Nacional sobre el Sida, publicados el 13 de junio de 2004. Pero el 25% de los seropositivos en la población no está diagnosticado.

Como conclusión de todo este apartado es importante señalar, como hace Altman y la mayoría de los autores con publicaciones en este tema, que la elección de un punto de corte no será (quizá nunca) una decisión estadística. En el caso de Medicina, tiene que ser útil desde el punto de vista clínico, y además se debe elegir el punto de corte teniendo en cuenta los costes relativos (no necesariamente financieros o monetarios) asociados con los resultados FP y FN que se pueden conseguir al aplicar dicho punto de corte.

Aquí también hay que diferenciar mucho entre lo que es un test de "detección" de lo que es un test "diagnóstico". Esta distinción también se llama "diagnosis" vs. "prognosis". Por ejemplo, podemos desear detectar la mayor parte de posibles positivos en una primera fase, utilizando pruebas que sean suficientemente rápidas y baratas (pruebas de detección rápida o "*screening*"), para dejar el estudio más detallado de casos, y por tanto, más costoso, para una segunda prueba (de diagnóstico). Éste es el enfoque actual en muchas pruebas médicas.

Medidas complementarias a las propias de la curva ROC

Altman propone dos enfoques que él dice que "son más informativos que simplemente mirar a la sensibilidad y especificidad":

La razón de verosimilitud (*likelihood* ratio, LR), que viene definida por

$$LR = \frac{\Pr(resultado\ positivo \mid verdadero\ positivo)}{\Pr(resultado\ positivo \mid verdadero\ negativo)} = \frac{\text{sensibilidad}}{1 - \text{especificidad}} \qquad (5.26)$$

Se puede considerar la razón de verosimilitud como indicador del valor de la prueba para incrementar la certeza sobre un diagnóstico positivo. La prevalencia es la probabilidad de la enfermedad antes de que se lleve a cabo la prueba. Los *odds* de tener la enfermedad se dan entonces como

$$odds = \frac{prevalencia}{(1 - prevalencia)} \qquad (5.27)$$

Entonces, si la prevalencia es de un 10%, los *odds* son 0.11, o 9 a 1 contra el hecho que la enfermedad esté presente. Este sería el valor de *odds* "pre-test" y los *odds* correspondientes al valor predictivo positivo serían los "*odds* post-test".

Se puede mostrar que:

$$Odds\ post\ test = odds\ pre-test\ x\ razón\ verosimilitud \qquad (5.28)$$

Este enfoque proporciona una mayor comprensión de la interpretación de los datos de la prueba diagnóstica, pero no añade información nueva porque se utilizan las mismas cantidades que antes.

Una alta razón de verosimilitud puede demostrar que la prueba es útil pero no necesariamente que un resultado positivo sea un buen indicador de la presencia de la enfermedad. Por ejemplo, para una prevalencia baja, p.ej., 0.25, un caso con un resultado positivo es todavía más probable que sea "negativo" que "positivo". Utilizar "*odds*" antes que probabilidades ayuda a ver la utilidad de la prueba tal y como la evalúa la razón de verosimilitud.

5.5 Evaluación de la eficacia de algoritmos de clasificación

En este punto presentamos los resultados de una (forzosamente) pequeña revisión de la medida de la eficacia en minería de datos, en especial sobre algoritmos de recomendación automática. Nuestro objetivo es mostrar, al menos desde un punto de vista descriptivo, que el enfoque de análisis ROC es generalizable a muchos casos de aplicación típica de la minería de datos.

Nuestro primer ejemplo representa un punto de partida en el solapamiento de las distribuciones predichos-reales, en una aplicación que proporciona sugerencias a compradores (Lawrence, Almasi, Kotlyar, Viveros, Duri, 2001). No nos cabe duda que en este estudio se podría evaluar sugeridos vs. aceptados utilizando el análisis ROC.

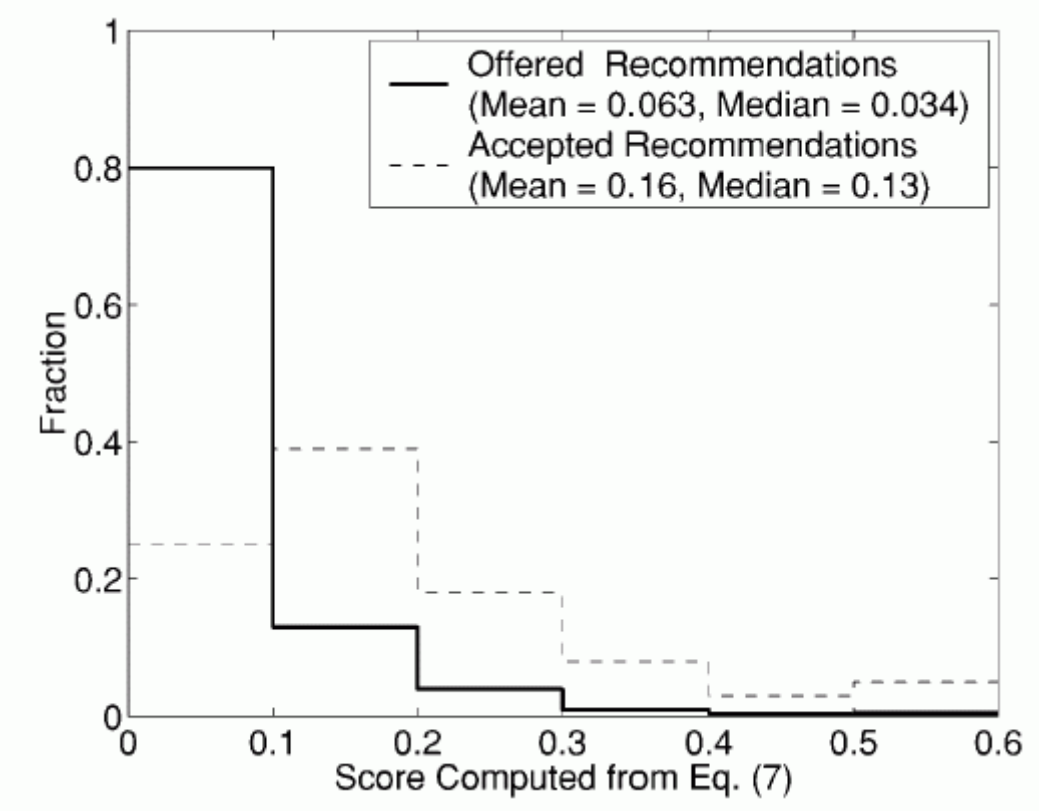


Figura 5.2. Distribuciones solapadas de puntuaciones para recomendaciones ofrecidas y aceptadas (tomado de Lawrence, Almasi, Kotlyar, Viveros, Duri, 2001)

Kleinberg, Papadimitriou y Raghavan (1998) presentan un enfoque para la evaluación de procedimientos de minería de datos en términos de su utilidad para la toma de decisiones, basado en la optimización. Este enfoque lleva a problemas computacionales relacionados con el análisis de sensibilidad, segmentación y la teoría de juegos. Estos autores destacan que la evaluación de las técnicas de minería sigue siendo un reto, puesto que muchas investigaciones se han concentrado en el reconocimiento o extracción de patrones pero no ha habido mucha investigación para determinar qué patrones son interesantes. Estos autores proponen procedimientos de optimización basados en la minimización o maximización de determinadas características económicas del uso que se hace de los modelos, según la teoría de decisión. Por otro lado, hablan de sensibilidad, pero en términos de teoría de decisión (análisis de sensibilidad), no de análisis ROC (puede estar relacionado pero no es directo):

La precisión de los sistemas automatizados de recomendación, y sobre todo la posibilidad de utilizar esa medida para compararlos, está presente en muchos artículos

recientes de minería de datos. Por ejemplo, Lin, Alvarez y Ruiz (2002) compararon varios algoritmos de redes neuronales y comparan su eficacia mediante una precisión basada en el porcentaje de asociaciones correctas. Exponen del siguiente modo su punto de partida:

"Utilizamos (...) ajuste (*accuracy*) como el porcentaje de artículos correctamente clasificados entre todos los clasificados por el sistema; "precisión" (*precision* en su original inglés) como el porcentaje de artículos recomendados a un usuario que efectivamente le gustan al usuario; recuerdo como el porcentaje de artículos que le gustan a un usuario que efectivamente son recomendados por el sistema."

Lin, Alvarez y Ruiz (2002) definen las siguientes relaciones (conservamos el original inglés para no inducir a confusión con su traducción):

accuracy = correctly classified articles / total articles classified
precision = correctly recommended articles / total recommended articles
recall = correctly recommended articles / total articles liked by users

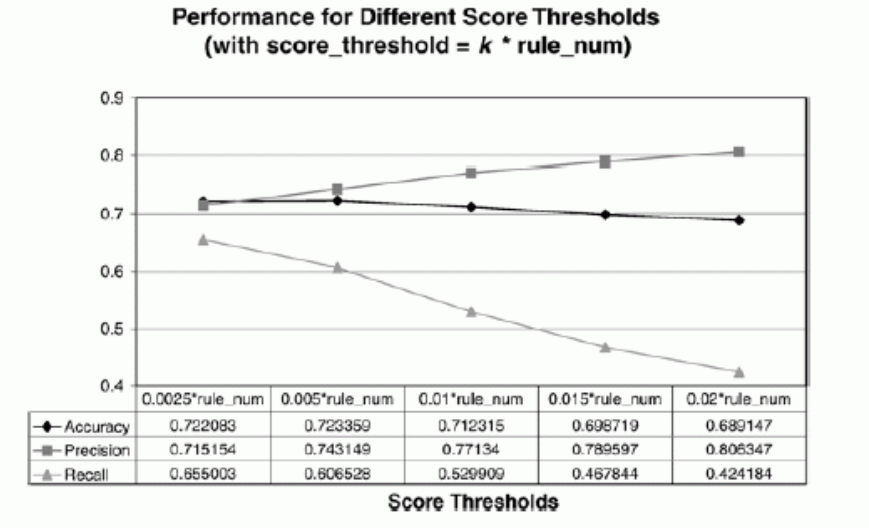
En ningún caso señalan las relaciones que puedan existir entre estos tres índices, y de hecho los tratan como independientes cuando muy seguramente no lo sean.

Las medida "*recall*" y "*precision*" son medidas clásicas en recuperación de información, y que se pueden expresar en los términos de ROC que estamos utilizando en esta tesis. La medida "*recall*" sería el porcentaje de documentos recuperados que son extraídos por el sistema, o también nuestra conocida "especificidad" mientras que "*precision*" sería el porcentaje de documentos que son relevantes, y sería el equivalente a la "sensibilidad" (Witten y Frank, 2000):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.29}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.30}$$

A continuación calculan los índices de medida definidos para diferentes umbrales de "gusto" o preferencia, y van viendo su efecto en las medidas de efectividad que han propuesto:



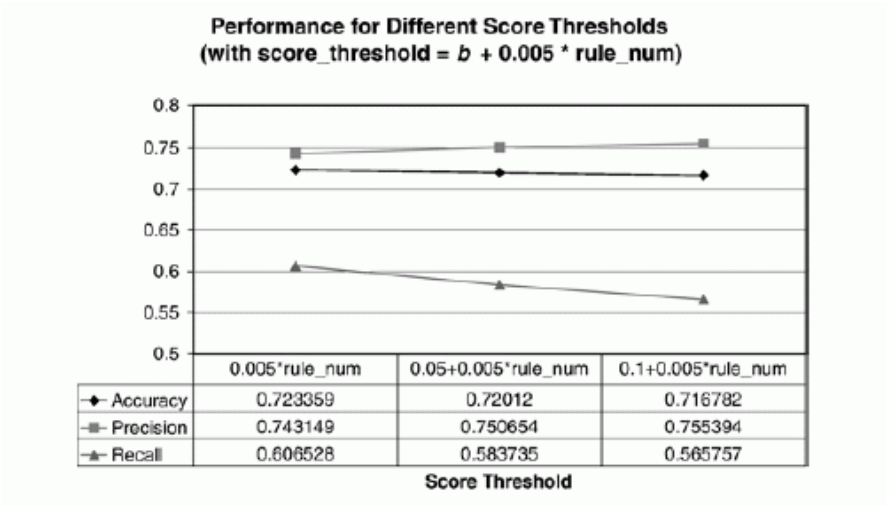


Figura 5.3. Resultados de clasificación de varios algoritmos de recomendación basados en reglas, en 3 medidas, y para diferentes puntos de corte tomadas de Lin, Alvarez y Ruiz (2002)

Su propuesta acaba con una comparación entre su sistema de recomendación, una vez decidido un umbral a partir de los resultados anteriores, que comparan con la eficacia de la recomendación por azar:

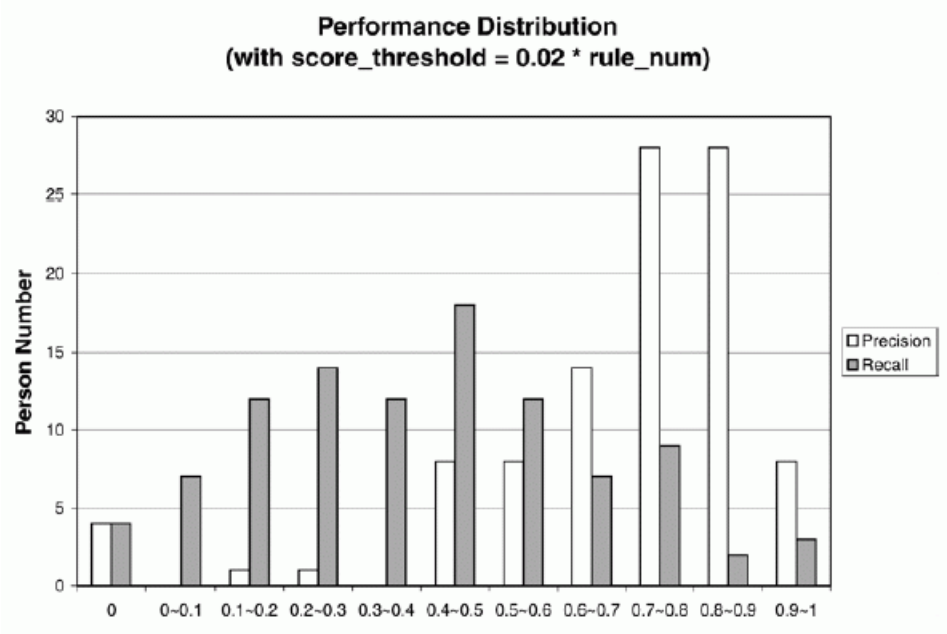


Figura 5.4. Distribución de una medida agregada de precisión en función de puntos de la función de preferencia, en comparación con la selección aleatoria de recomendaciones (tomadas de Lin, Alvarez y Ruiz, 2002)

Como podemos observar, al menos conceptualmente, este enfoque es muy similar al de un estudio susceptible de análisis ROC.

Mobasher, Dai, Luo y Nakagawa (2002) usan un enfoque más formal, y definen dos medidas de eficacia de los sistemas de recomendación. Supongamos que tenemos una transacción t (por ejemplo un conjunto de vistas de páginas en internet), y que utilizamos una ventana w para producir un conjunto de recomendaciones al visitante R. Entonces, la precisión de R con respecto a t se puede definir como:

$$precision(R,t) = \frac{|R \cap (t-w)|}{|R|} \qquad (5.31)$$

y la cobertura igualmente de R con respecto a t como:

$$cobertura(R,t) = \frac{|R \cap (t-w)|}{|t-w|} \qquad (5.32)$$

Estas medidas adaptan las medidas típicas de precisión y recuerdo que se utilizan normalmente en recuperación de información. En este contexto, la precisión mide el

grado en que los motores de recomendación producen recomendaciones ajustadas (por ejemplo, la proporción de las recomendaciones relevantes con respecto al número total de recomendaciones). Por otro lado, la cobertura mide la capacidad del sistema de recomendación de producir todas las páginas que serán probablemente visitadas por los usuarios (proporción de recomendaciones relevantes a todas las visiones que deberían ser recomendadas).

Aclaran que ninguna de estas medidas, por sí mismas son suficientes para evaluar el rendimiento del motor de recomendación.

Idealmente, se desearía la máxima precisión y la máxima cobertura. Una medida única que captura esto es la medida F1 (Lewis y Gale, 1994):

$$F1(R,t) = \frac{2\,precision(R,t)\,cobertura(R,t)}{precision(R,t) + cobertura(R,t)} \qquad (5.33)$$

Witten y Frank (2000) presentan esta misma fórmula denominada "medida-F", pero relacionando "precision" y "recall":

$$F(R,t) = \frac{2\,recall\ precision}{recall + precision} \qquad (5.34)$$

En el estudio de Mobasher et al. (2002) también adquieren una alta relevancia los umbrales a partir de los cuales se realizan las recomendaciones:

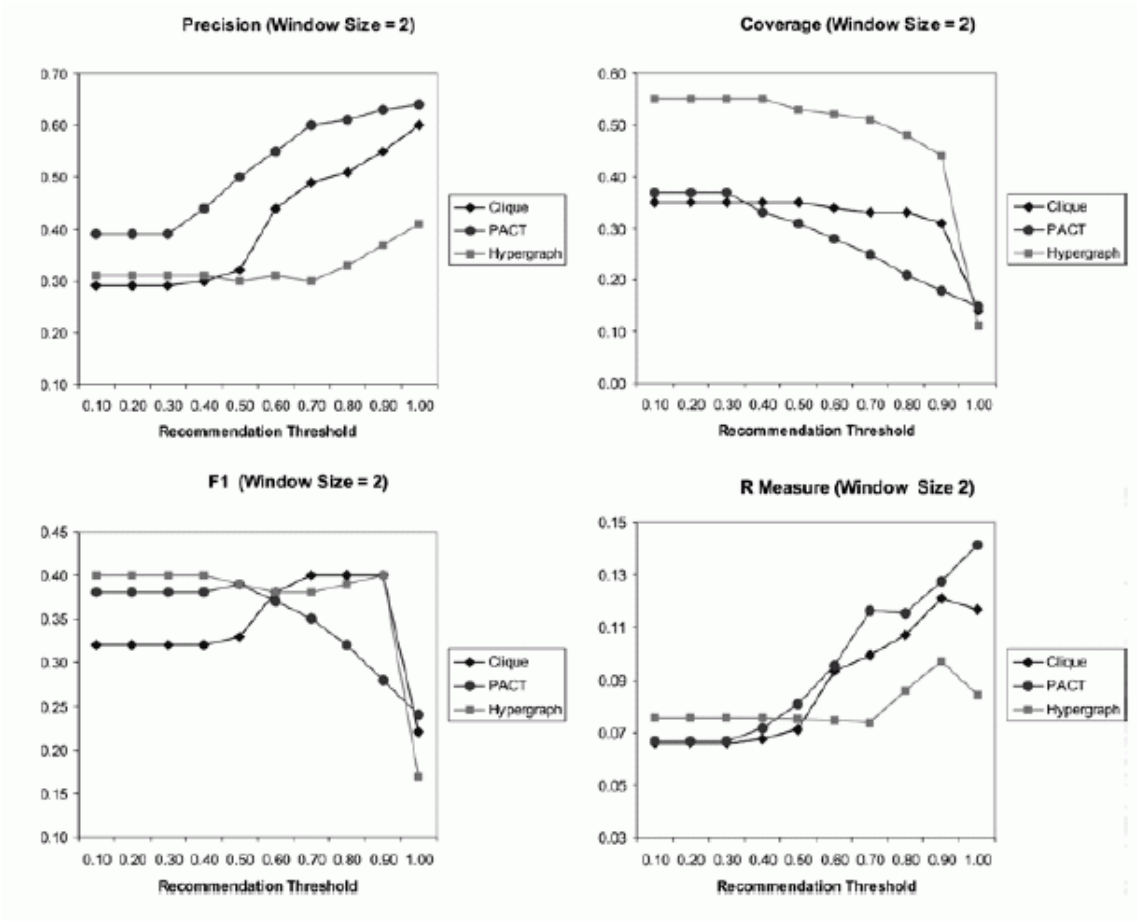


Figura 5.5. Comparación de varias medidas de eficacia de algoritmos de recomendación en función de umbrales (tomadas de Mobasher, Dai, Luo y Nakagawa, 2002)

Muchos estudios siguen este enfoque empírico de establecer unos índices de eficacia, y variar los umbrales para después tomar la decisión. Las medidas pueden ser estándar, como las mencionadas anteriormente, o realizadas ad-hoc para el estudio, como las que recogen Adomavicius y Tuzhilin (2001). Entre otras medidas objetivas mencionan:

- Confianza y apoyo (Agrawal *et al*., 1993),
- Ganancia (Fukuda *et al*., 1998),
- Varianza y valor del chi-cuadrado (Morishita, 1998),
- Medida de Gini (Morimoto *et al*., 1998),
- Fortaleza (Dhar and Tuzhilin, 1993),
- Convicción (Brin *et al*., 1997),
- Optimalidad (Bayardo and Agrawal, 1999),

Y entre las medidas subjetivas (con nombres ingleses de difícil traducción):

- Sorpresa ("*unexpectedness*")(Silberschatz and Tuzhilin, 1996; Liu and Hsu, 1996; Suzuki, 1997; Padmanabhan and Tuzhilin, 1999)
- Accionabilidad (Piatetsky-Shapiro and Matheus, 1994; Silberschatz and Tuzhilin, 1996; Adomavicius and Tuzhilin, 1997).

Otros estudios utilizan técnicas de diseño experimental para la validación y comparación de modelos, como Salzberg (1997). Salzberg insiste en el ajuste del nivel de significación para todo el experimento, quizá con el ánimo de recordar a personas que no provienen de ese campo la necesidad de tener en cuenta este importante hecho.

5.6 El *lift chart*

Los gráficos conocidos como "*lift chart*" (gráfico de elevación) son una herramienta clásica en Marketing para comparar diferentes escenarios o también los resultados de la clasificación cruzada producto de modelos predictivos. En este contexto se le conoce también como gráfico de ganancias para un resultado binario (Parekkat, 2004). Hay varias posibilidades para dibujar *lift charts*, aquí sólo mostraremos las dos más comunes.

Qué es la "elevación"

Coppock (2002) señala que *lift* es probablemente la métrica más utilizada para medir el rendimiento de modelos de "*targeting*" (de selección de objetivos) en aplicaciones de Marketing. En este contexto particular Levy (2001) define el "*lift*" (podríamos traducirlo por elevación) como la medida de la eficacia de una campaña de Marketing que envía a individuos seleccionados mediante un modelo en comparación a una lista seleccionada aleatoriamente. Se trata de una medida típica del Marketing de bases de datos, dado que uno de sus principales objetivos es producir el mayor retorno de la inversión.

El propósito de un modelo simple de *targeting* es identificar un subgrupo (*target*) de una población más amplia. Los miembros *target* seleccionados son los que más probablemente responderán positivamente a una oferta de Marketing. Un modelo hace un buen trabajo si la respuesta dentro del *target* es mucho mejor que la media para la población como un todo. *Lift* o elevación es simplemente la razón de estos valores: respuesta del target dividido por la respuesta promedio.

La base de muchas medidas de eficacia de la clasificación es una matriz de confusión, que se produce a partir de la comparación entre los datos de entrenamiento y los de validación.

El *lift* normalmente se cuantifica dividiendo la población en deciles, en los que se posicionan los miembros de la población, basados en su probabilidad predicha de respuesta. Los respondientes más altos se ponen en el decil 1 y así sucesivamente. La tabla 5.9 muestra unos datos de ejemplo de respuesta a una promoción por correo en una población con una respuesta media del 5%.

Tabla 5.9. Ejemplo de cálculo del empuje con datos simulados

		Por decil			Acumulado por decil	
Decil	Ofertas por decil	Respuesta obtenida	Porcent. por decil	Elevac.	Respond. (acum.)	Porc. de tot. acum.
1	100	16	16.0%	3.20	16	32%
2	100	12	12.0%	2.40	28	56%
3	100	8	8.0%	1.60	36	72%
4	100	5	5.0%	1.00	41	82%
5	100	3	3.0%	0.60	44	88%
6	100	2	2.0%	0.40	46	92%
7	100	1	1.0%	0.20	47	94%
8	100	1	1.0%	0.20	48	96%
9	100	1	1.0%	0.20	49	98%
10	100	1	1.0%	0.20	50	100%
Totales	1000	50	5.0%			

El *lift chart*

Para cada persona objetivo hay un coste por hacerle la oferta, y un beneficio correspondiente si se obtiene una respuesta positiva. Se puede entonces calcular el beneficio de poner como target cada decil y simplemente incluir cada decil hasta el último que sea beneficioso. La tasa de respuesta acumulada y el *lift* mostrarán entonces el rendimiento medio del modelo para todos en el target. Si representamos esta información gráficamente, nos queda un gráfico decreciente como el de la figura 5.6.

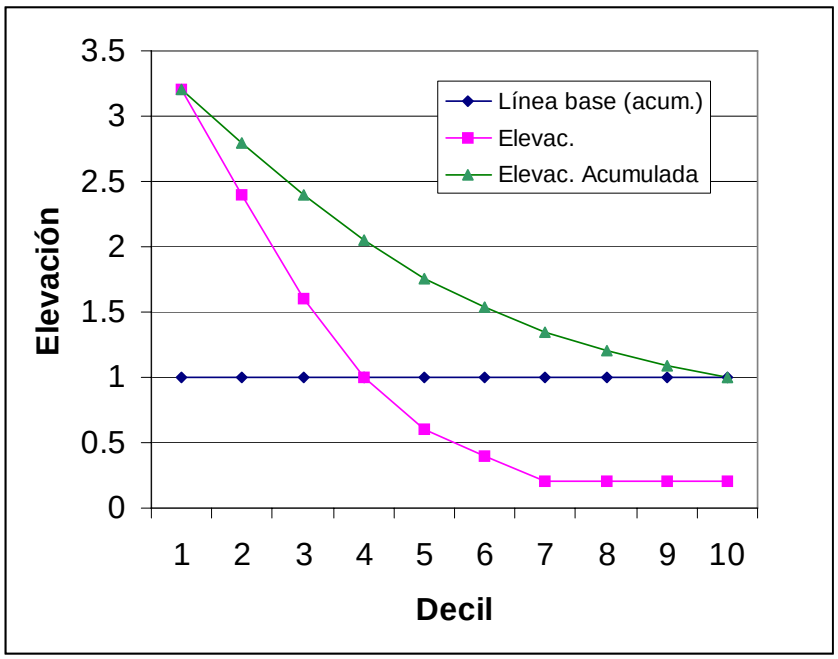


Figura 5.6. Gráfico de elevación con los datos del ejemplo en la tabla 5.9

Este gráfico nada tiene que ver con el gráfico de una curva ROC. Sin embargo existe otra forma útil del gráfico de elevación que compara el porcentaje acumulado de respuestas capturadas según se añade cada decil al target. En el ejemplo actual, los dos deciles más altos capturan alrededor del 55% de los respondientes. Esto se compara con una línea base aleatoria, en la que existe una relación perfectamente lineal entre decil y rendimiento (no existe diferencia en los deciles en cuanto a respuesta). Cuanto mayor sea el área entre las dos líneas, mayor capacidad del modelo de concentrar respondientes en los deciles más altos. Este gráfico se parece más al de una curva ROC.

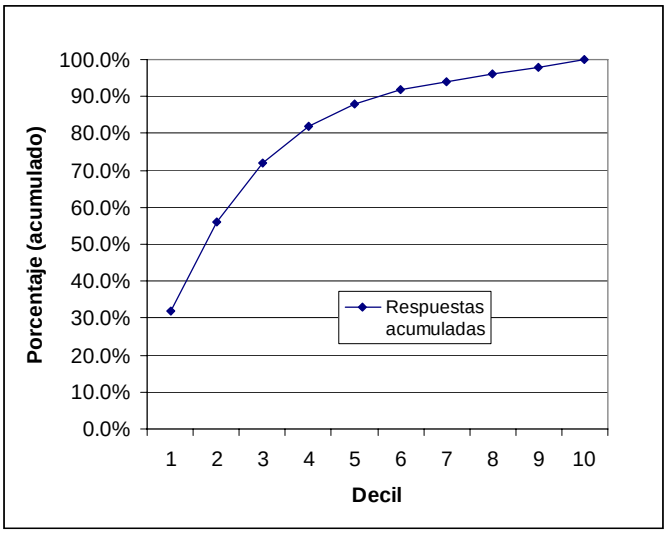


Figura 5.7. Gráfico de elevación para porcentajes acumulados con los datos del ejemplo en la tabla 5.9

Existe una relación entre curva ROC y gráfico de elevación, según describen Witten y Frank (2000). Tomemos los datos del factor 2 del instrumento de detección de maltrato que hemos utilizado en distintas partes de esta tesis. Si superponemos los gráficos de elevación y curva ROC según la modalidad anterior de porcentaje acumulado de respuesta (o pertenencia a grupo=1) y la curva ROC empírica encontraremos la representación del gráfico 5.8. Hay que señalar que esta superposición se ha hecho normalizando las escalas correspondientes al eje de abscisa puesto que no tienen la misma escala (porcentaje de falsas alarmas en curva ROC, vs. deciles o porcentaje del grupo de respondientes en el gráfico de elevación).

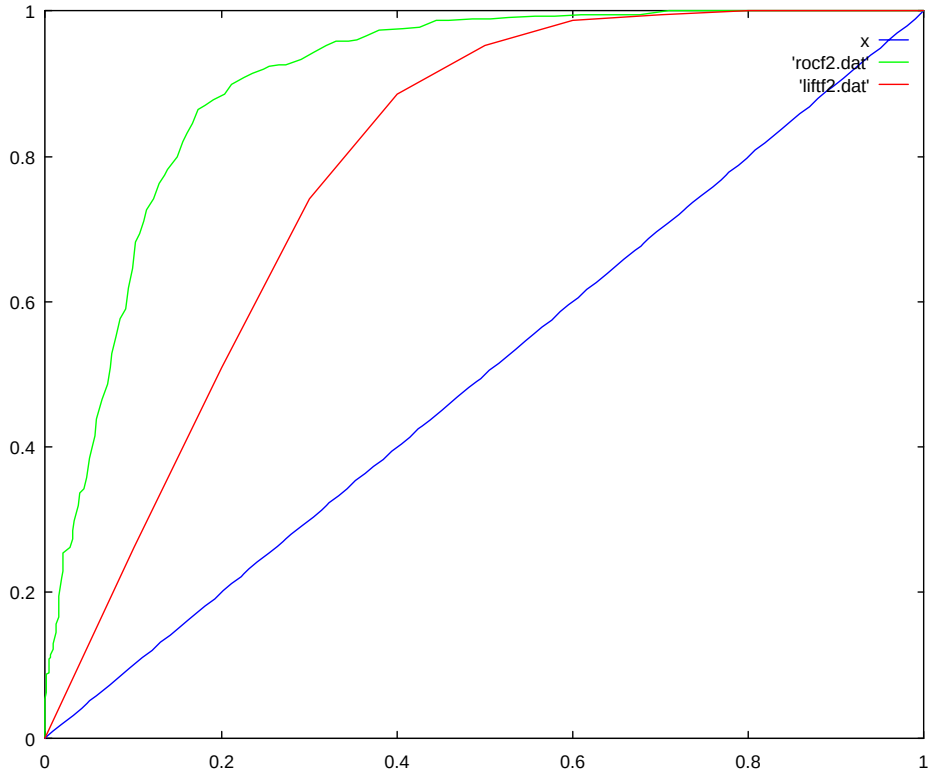


Figura 5.8. Gráficos de elevación y curva ROC superpuestos (los ejes no tienen la misma escala ni el mismo significado) para el factor 2 del instrumento de detección de maltrato

El eje de ordenadas en ambos tiene el mismo significado: es el porcentaje de éxito, el porcentaje del grupo de respuesta (o grupo=1), pero la diferencia esencial estriba en el eje x. Además de ser discreto (en este caso, pues los cálculos se han realizado manualmente), el significado no es el mismo, y por tanto es muy dependiente del contexto de aplicación concreta en la que se realice esta representación (sobre todo de la

prevalencia o tasa media de "respuesta" en la población). Aunque, por otro lado, suponen formas similares, no son el mismo ni poseen las mismas propiedades.

El gráfico de elevación es estándar para comparar modelos en algunos sectores. A continuación, en las figuras 5.8 y 5.9 mostramos cómo lo hacen dos programas de minería de datos, el SAS Enterprise Miner y el SPSS Clementine (véase la revisión de Haughton *et al*, 2003).

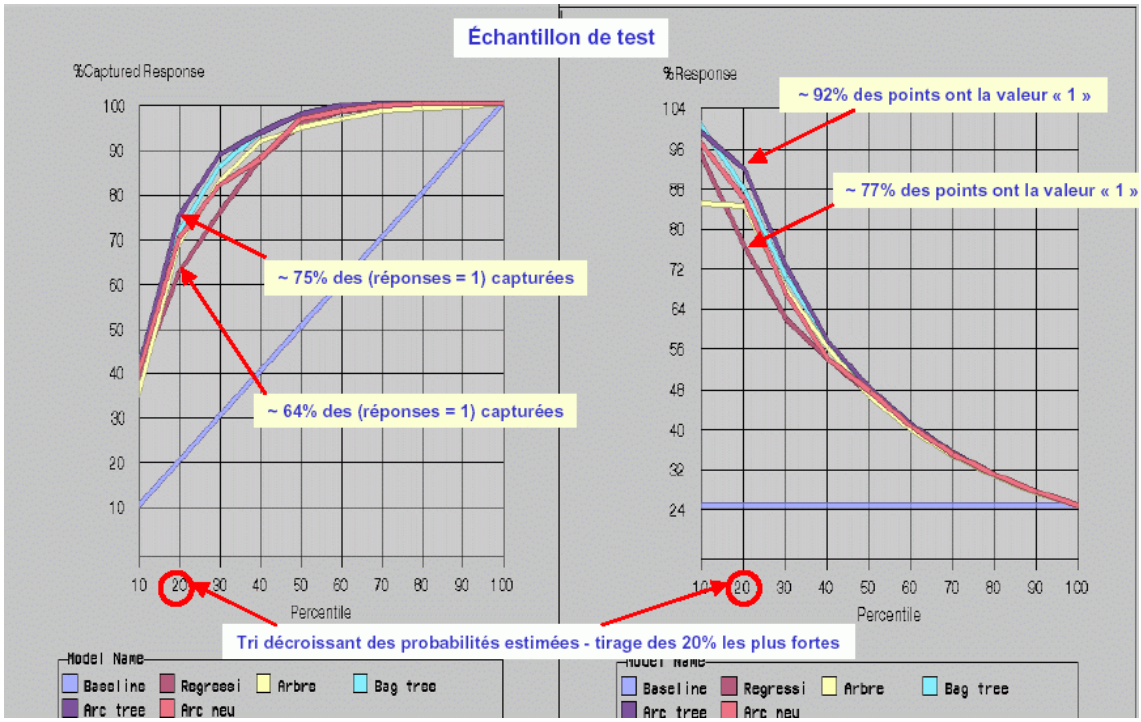


Figura 5.8. Ejemplos de lift chart convencional (dcha.) y acumulado, obtenidas con el programa SAS Enterprise Miner, tomado de Haughton et al, 2003.

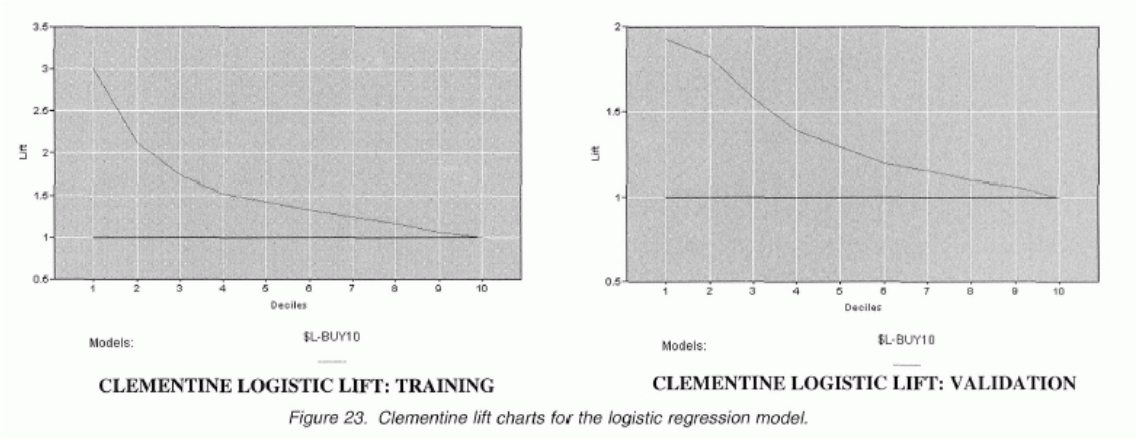


Figura 5.9. Ejemplos de lift chart acumulados, obtenidas con el programa SPSS Clementine, tomado de Haughton et al, 2003

Aunque sólo sea por curiosidad, hemos encontrado otra representación gráfica similar visualmente a la curva ROC, el denominado gráfico de Gini, que Haughton *et al* (2003) describen en su aplicación en el software Quadstone.

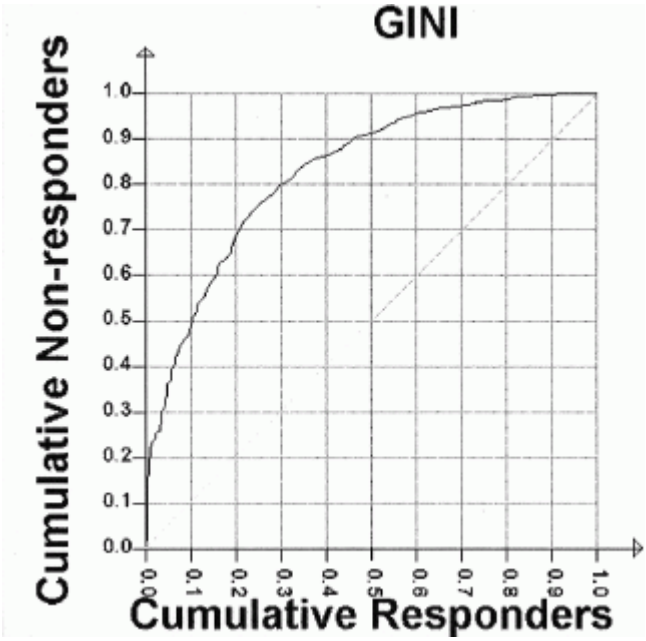


Figura 5.10. Ejemplo de gráfico de Gini, según lo calcula el programa Quadstone, tomado de Haughton et al, 2003

"El gráfico de Gini proporcionado por el software Quadstone, procedimiento "*scorecard*", está construido ordenando los casos en orden incremental de la puntuación del modelo (del menos probable para responder al más probable) y después yendo por los casos ordenados y haciendo un paso horizontal cuando un respondiente real aparece y un paso vertical cuando aparece un no-respondiente. En el modelo ideal, todos los

respondientes tendrían puntuaciones menores que todos los no respondientes, de tal modo que todos los verticales precederían todos los horizontales. Esto daría un gráfico de Gini con la forma de un triángulo con el vértice hacia arriba. Los autores notan que "no es una definición estándar de la medida de Gini y que el gráfico es más reminiscente, pero no igual, a los modelos típicos de curvas ROC para modelos logísticos".