Una aproximación al análisis cuantitativo de riesgo

Dr. Viterbo H. Berberena González
Profesor de Ciencias Analíticas
Centro de Alta Dirección en Ingeniería y Tecnología, Universidad Anáhuac

Conferencia del Forum Scoring 2008, Hotel Nikko 24 de junio 2008, Ciudad de México

Agenda

1. La función binomial como modelo estadístico de eventos dicotómicos.
2. Medidas de asociación estadística.
   - Tablas de contingencias
   - Análisis de correlación.
3. Medidas de riesgo.
   - Reducción absoluta del riesgo.
   - Reducción relativa del riesgo.
   - Riesgo relativo.
   - Odd - Momio.
   - Odds ratio - Relación de momio.
   - WOE - Weight of Evidence.
4. La regresión logística.
   - Matriz de confusión.
   - Curva ROC.

Agenda

5. Construcción de “ScoreCards” de riesgo.
6. Medidas de desempeño de los “SoreCards”.
   - Medidas de Tipo I (No tienen en cuenta el punto de corte).
     - Gini
     - KS
     - Information Value.
     - Mean Difference.
     - Divergence.
   - Medidas de Tipo II (Tienen en cuenta el punto de corte).

Función Binomial

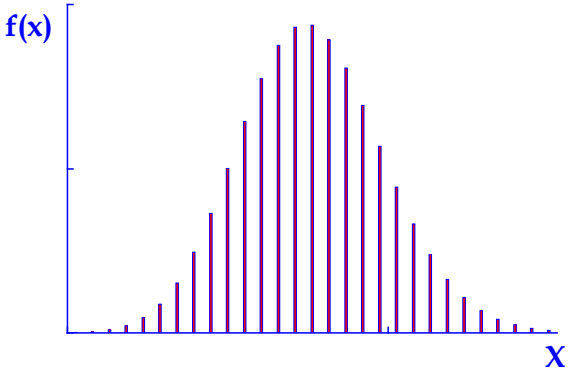
Modelo probabilístico para eventos dicotómicos:

- Éxito o evento (1)
- Fracaso o no evento (0)
- La probabilidad de cada uno de ellos constante en una serie de repeticiones.

La función de cuantía o masa de la probabilidad:

$$f(x) = \binom{n}{x} p^{x} q^{n-x}$$

$$0 \le x \le n$$



Medidas de Asociación Estadística

Supongamos que una Institución Bancaria está interesada en: "***evaluar la capacidad de un modelo de prevención en línea para disminuir el número de fraudes en cheques***".

Se selecciona una muestra aleatoria de 400 clientes y se divide en dos: a una parte se le aplica el modelo y a la otra no. Se contabilizan los cheques falsos obtenidos y los resultados se presentan a continuación.

Medidas de Asociación Estadística

		Variable Independiente	
		Con Modelo **X=1**	Sin Modelo **X=0**
Variable Objetivo	No Fraude (NoEvento) **Y=0**	**180**	**130**
	Fraude (Evento) **Y=1**	**20**	**70**
		200	**200**

Medidas de Asociación Estadística

El chi-cuadrado:

$$Q = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \; aproxim. \sim \chi^2\left(k\text{-}c\text{-}1\right)$$

$$Pearson \;\; \chi^2 = 35.84 \qquad P\left(\chi^2\right) \langle\, 0.0001$$

$$Phi \;\; \phi = -0.299 \qquad P\left(\phi\right) \langle\, 0.0001$$

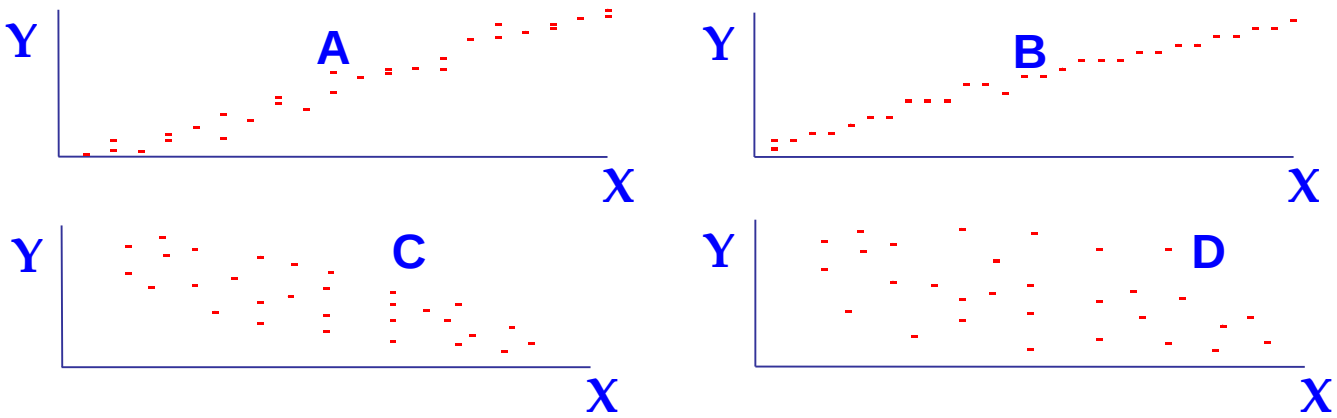
$$Cramer \; V = \;\; 0.299 \qquad P\left(V\right) \langle\, 0.0001$$

$$Coef.Cont. = \;\; 0.299 \qquad P\left(CC\right) \langle\, 0.0001$$

Medidas de Asociación Estadística

El coeficiente de correlación muestral:

$$r = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\left(n-1\right) s_x^2 \; s_x^2}$$



Medidas de Riesgo

Diferencia de riesgo o Reducción Absoluta del Riesgo (RAR):

$$RAR = \frac{180}{200} - \frac{130}{200} = \frac{180-130}{200} = \frac{50}{200} = .25$$

P[(Y=0)|(X=1)]=0.9

P[(Y=0)|(X=0)]=0.65

Esta probabilidad es la disminución que se produce en el riesgo de fraude al utilizar el modelo

El RAR puede oscilar entre -1 y 1; 0 indica no asociación.

Medidas de Riesgo

Reducción Relativa del Riesgo (RRR): Se define como la reducción absoluta del riesgo dividida por el riesgo basal o riesgo del grupo de referencia.

$$RRR = \frac{.25}{.65} = .385$$

Esto significa que la disminución del riesgo, teniendo en cuenta la utilización de modelo contra la no utilización, es el 38.5% del riesgo que se produce cuando no se usa el modelo.

Medidas de Riesgo

Riesgo Relativo (RR): Es el cociente entre los riesgos.

$$RR = \frac{.90}{.65} = 1.4$$

La probabilidad de éxito (no fraude) es 1.38 veces mayor cuando se usa el modelo de prevención, en la relación con el caso en que se prescinde de este.

El RR puede oscilar entre 0 y ∞; 1 indica no asociación. Es el estadístico preferido.

Medidas de Riesgo

Odd (Momio): Es el cociente p/q.

$$Odd_{con\,modelo\,X=1} = \frac{.90}{.10} = 9.0$$

$$Odd_{sin\,modelo\,X=0} = \frac{.65}{.35} = 1.9$$

Es 9 veces más probable que no ocurra fraude cuando se utiliza el modelo de prevención, mientras que cae a sólo 1.9 veces cuando no se utiliza.

Medidas de Riesgo

Odds Ratio (OR): Es el cociente de odds de dos atributos (***X=1*** y ***X=0***). Cuantifica cuánto más probable es la aparición del Evento cuando está presente el atributo ***X=1*** con respecto a cuando está presente el atributo ***X=0***.

$$Odds\ Ratio = \frac{9}{1.9} = 4.8$$

Es 4.8 veces más probable que no se presenten fraudes cuando se usa el modelo de prevención en relación con la no utilización de este.

Los Odds puede oscilar entre 0 y ∞; 1 indica no asociación.

Weight of Evidence (WOE)

El **WOE** mide el riesgo relativo de un atributo o el nivel de grupo. Si se tiene una variable objetivo binaria de forma que: Bad, (Event=1) y Good (NoEvent=0), el WOE se define como:

$$WOE_{Attribute} = \ln \frac{P_{Attribute}^{NoEvent}}{P_{Attribute}^{Event}}$$

$$P_{Attribute}^{NoEvent} = \frac{n_{Attribute}^{NoEvent}}{N_{Charact.}^{NoEvent}}$$

$$P_{Attribute}^{Event} = \frac{n_{Attribute}^{Event}}{N_{Charact.}^{Event}}$$

Weight of Evidence (WOE)

Donde,

$n_i^{NoEvent}$ - Número de no eventos en el atributo (grupo o clase).

n_i^{Event} - Número de eventos en el atributo (grupo o clase).

$N_{Charact.}^{NoEvent}$ - Número de no eventos totales en la característica (variable).

$N_{Charact.}^{Event}$ - Número de eventos totales en la característica (variables).

Regresión Logística

Partiendo de la definición de Odds (Momios):

$$Odds = \frac{P(Event)}{P(NoEvent)} = \frac{P(Event)}{1 - P(Event)}$$

Aplicando ***Ln*** e igualando a un modelo lineal:

$$\ln\left(\frac{P(Event)}{1 - P(Event)}\right) = \alpha + \beta x$$

Extrayendo exponencial a ambos términos:

$$\frac{P(Event)}{1 - P(Event)} = e^{\alpha + \beta x}$$

Regresión Logística

Despejando:

$$P(Event) = e^{\alpha+\beta x} - P(Event)e^{\alpha+\beta x}$$

$$P(Event) + P(Event)e^{\alpha+\beta x} = e^{\alpha+\beta x}$$

$$P(Event)\left(1 + e^{\alpha+\beta x}\right) = e^{\alpha+\beta x}$$

$$P(Event) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

Dividiendo por $e^{\alpha+\beta x}$:

$$P(Event) = \frac{1}{\dfrac{1 + e^{\alpha + \beta x}}{e^{\alpha + \beta x}}}$$

Regresión Logística

Ordenando:

$$P(Event) = \frac{1}{\frac{1 + e^{\alpha + \beta x}}{e^{\alpha + \beta x}}}$$

$$P(Event) = \frac{1}{1 + \frac{1}{e^{\alpha + \beta x}}}$$

$$\boxed{P(Event) = \frac{1}{1 + e^{-\alpha - \beta x}}}$$

Regresión Logística

La matriz de confusión:

		PREDICTED		
		0	1	
OBSERVED	0	TRUE NEGATIVES	FALSE POSITIVES	TN + FP
	1	FALSE NEGATIVES	TRUE POSITIVES	FN + TP

Regresión Logística

La Sensibilidad:

- Probabilidad de clasificar correctamente un caso cuyo estado original es ‘positive’ (1).
- Proporción de ‘True Positives’, 1s observados predichos como 1s (HITS).
- A mayor sensibilidad, mayor potencia para clasificar o identificar correctamente a los 1s.

$$Sensibilidad = \frac{TP}{FN + TP}$$

Regresión Logística

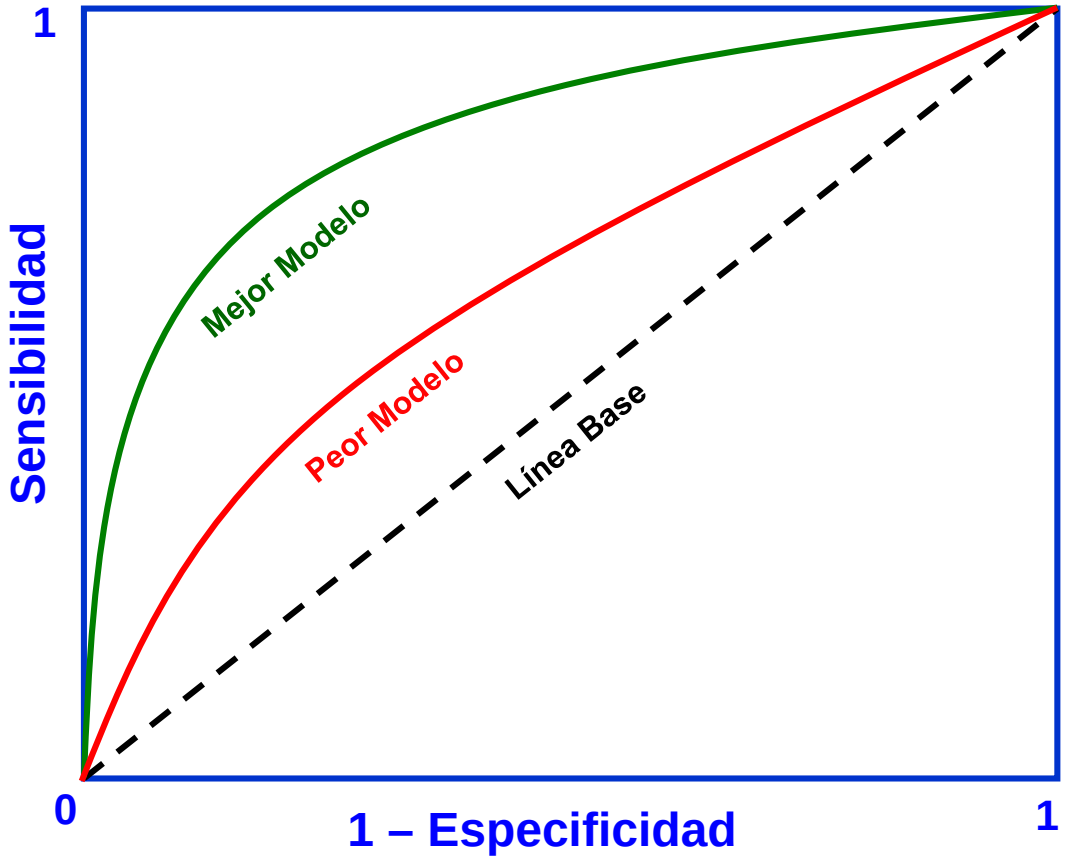
La Especificidad:

- Es la probabilidad de clasificar correctamente un caso cuyo estado original es ‘negative’ (0).
- Proporción de ‘True Negatives’, 0s observados predichos como 0s (Accurate rejection).
- A mayor especificidad, mayor potencia para clasificar o identificar correctamente a los 0s.

$$Especificidad = \frac{TN}{TN + FP}$$

Regresión Logística

La Curva ROC:



Construcción del Scorecard

El WOE de cada atributo es multiplicado por el coeficiente de regresión de su característica para obtener el puntaje del scorecard del atributo. El puntaje total del solicitante es proporcional al logaritmo de la probabilidad de su razón good/bad estimada. A continuación se presenta la ecuación para el cálculo de los puntajes del scorecard para cada solicitante.

Construcción del Scorecard

$$score = \log(odds) * factor + offset =$$

$$= (-\sum_{i=1}^{n}(woe_i * \beta_i) + \alpha) * factor + offset =$$

$$= (-\sum_{i=1}^{n}(woe_i * \beta_i + \frac{\alpha}{n})) * factor + offset =$$

$$= \sum_{i=1}^{n}(-(woe_i * \beta_i + \frac{\alpha}{n}) * factor + \frac{offset}{n})$$

Los puntajes están en escala lineal de enteros conforme a las normas de esta industria.

Construcción del Scorecard

Se tomó una escala de puntajes tal que el valor de 600 corresponde a una relación good/bad de 50/1 y que un incremento en el puntaje de 20 unidades coincide con el doble de la relación good/bad. Para la obtención de la regla de escalamiento que transforme los puntajes de cada atributo se usan las ecuaciones:

$$600 = \log(50) * factor + offset$$

$$620 = \log(100) * factor + offset$$

Construcción del Scorecard

$$factor = 20 / \log(2)$$

$$offset = 600 - factor * \log(50)$$

El scorecard resultante es una tabla, opcionalmente en formato HTML. Un fragmento de esta se muestra a continuación. Se aprecia como los puntajes de las características cubren diferentes rangos.

Construcción del Scorecard

Characteristic Name	Attribute	Scorecard Points
EDAD	. -> 27	56
EDAD	27 -> 30	70
EDAD	30 -> 31	61
EDAD	31 -> 38	60
EDAD	38 -> 44	60
EDAD	44 -> .	70
ESTADO	"CHIAPAS", "COAHUILA", "JALISCO", "NAYARIT", "NL", "SINALOA", "SONORA"	49
ESTADO	"CHIHUAHUA", "GUERRERO", "SLP"	55
ESTADO	"AGS", "DF"	56
ESTADO	"GUANAJUATO", "MICHOACAN", "PUEBLA", "QUERETARO", "YUCATAN", "ZACATECAS"	67
ESTADO	"HIDALGO", "MEXICO", "TAMAULIPAS"	70
ESTADO	"BCALIFORNIA", "CAMPECHE", "DURANGO", "MORELOS", "OAXACA", "QUINTANAROO", "TABASCO", "TLAXCALA", "VERACRUZ"	80
TIPO TRABAJO	"0", "8"	56
TIPO TRABAJO	"1", "2"	74
TIPO DEPTO	. -> 4	76
TIPO DEPTO	4 -> .	-102

Construcción del Scorecard

INGRESOS CTE	. -> 55	54
INGRESOS CTE	55 -> 11950	91
INGRESOS CTE	11950 -> .	55
INGRESOS CONYUGE	. -> 28	66
INGRESOS CONYUGE	28 -> .	59
IMPORTE VENTA	. -> 2173	73
IMPORTE VENTA	2173 -> 2840	69
IMPORTE VENTA	2840 -> 3818	63
IMPORTE VENTA	3818 -> 4646	64
IMPORTE VENTA	4646 -> 6596	65
IMPORTE VENTA	6596 -> 7963	63
IMPORTE VENTA	7963 -> .	52
INGRESO FIADOR	. -> 1701	75
INGRESO FIADOR	1701 -> .	61
PLAZO	. -> 28	87
PLAZO	28 -> 37	69
PLAZO	37 -> 41	60
PLAZO	41 -> .	50

Medidas de Desempeño

Existen dos tipos de medidas para evaluar el desempeño de los scorecards:

Tipo I. Son las que comparan la distribución de scores de los clientes en clase=0 con la distribución de los clientes en clase=1. Estas no tienen en cuenta el punto de corte (break point).

Tipo II. Son las que reconocen que el foco es la realización de una acción con el cliente. Teniendo en cuenta que se realizan acciones diferentes para los que son predichos en clase=0 y para los que son predichos en clase=1. Estas tienen en cuenta el punto de corte (break point).

Medidas de Desempeño

Las medidas de Tipo I tienen un enfoque Teórico Estadístico y las del Tipo II más bien, Práctico, de Uso.

El Criterio II es el más poderoso de los dos, en el sentido de refleja más aproximadamente el uso de los scorecards y sus reglas de clasificación. Sin embargo, se requiere, no sólo, la construcción del scorecard, sino la selección del punto de corte.

Pero ocurre que no esta bien claro cuál punto de corte seleccionar en el momento que se construye el scorecard. De hecho el punto de corte, puede variar con el transcurso del tiempo, así como las condiciones económicas cambien.

Medidas de Desempeño

Esto significa que se presenta la situación en que se quiere evaluar la efectividad del scorecard **si tiene un punto de corte explícito**. Entonces, los criterios Tipo I son apropiados para este caso.

Los criterios Tipos I son:

- Coeficiente Gini.
- Estadígrafo de Kolmogorov-Smirnov.
- Information Value (IV).
- Diferencia de Medias.
- Divergence (Fair Isaac)

Gini Coefficient

Se define usualmente como el doble del área entre la curva ROC y la diagonal del cuadrado del gráfico de ROC. Es equivalente al área bajo la curva ROC. A estas medidas se les puede dar un interpretación natural, ya que son equivalentes al estadígrafo de Mann-Whitney-Wilcoxon para comparación de dos muestras independientes, el cual estima la probabilidad de que un elemento aleatoriamente escogido de la clase 0, tendrá un score más bajo de uno escogido aleatoriamente de la clase 1.

Todas estas medidas son equivalentes en el sentido que hay funciones matemáticas directas que la relacionan.

Gini Coefficient

Es decir, dado unos se pueden calcular los otros. Se ve que ninguna de estas definiciones hace referencia al punto de corte.

Puede demostrarse que estas medidas son equivalentes al integrar la proporción de mal clasificados sobre toda la gama posible de puntos de corte, de modo que estas medidas, en un sentido real, no requieren de que sea escogido un punto de corte.

El coeficiente GINI mide el poder predictivo del scorecard. Mide la habilidad de un scorecard o de una característica, para discriminar el riesgo (to rank order risk).

Gini Coefficient

El coeficiente GINI se calcula como:

$$GC = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(1 - \hat{P}_{ij}^{2}\right)$$

$i = \overline{1,n}$ Número de observaciones en el grupo.

$j = \overline{1,m}$ Número de grupos.

$\hat{P}_{ij}$ Probabilidad posterior de la observación i en el grupo j.

Gini Coefficient

Algunos valores típicos de Coeficiente GINI:

GC = 0%	La característica/scorecard no puede distinguir los buenos de los malos.
GC = 100%	La característica/scorecard distingue perfectamente los buenos de los malos.
GC = 40% - 60%	Scorecard de crédito típico.
GC = 70% - 80%	Scorecard de comportamiento de crédito (Behaviour scorecards).
GC = 25%	Puede tener una característica muy poderosa.

Prueba de K-S

El estadígrafo de Kolmogorov-Smirnov es un estimado de la **<u>máxima diferencia</u>** entre las funciones de distribución acumulativas de los scores de los clientes clasificados como “0” y los scores de los clientes clasificados como “1”.

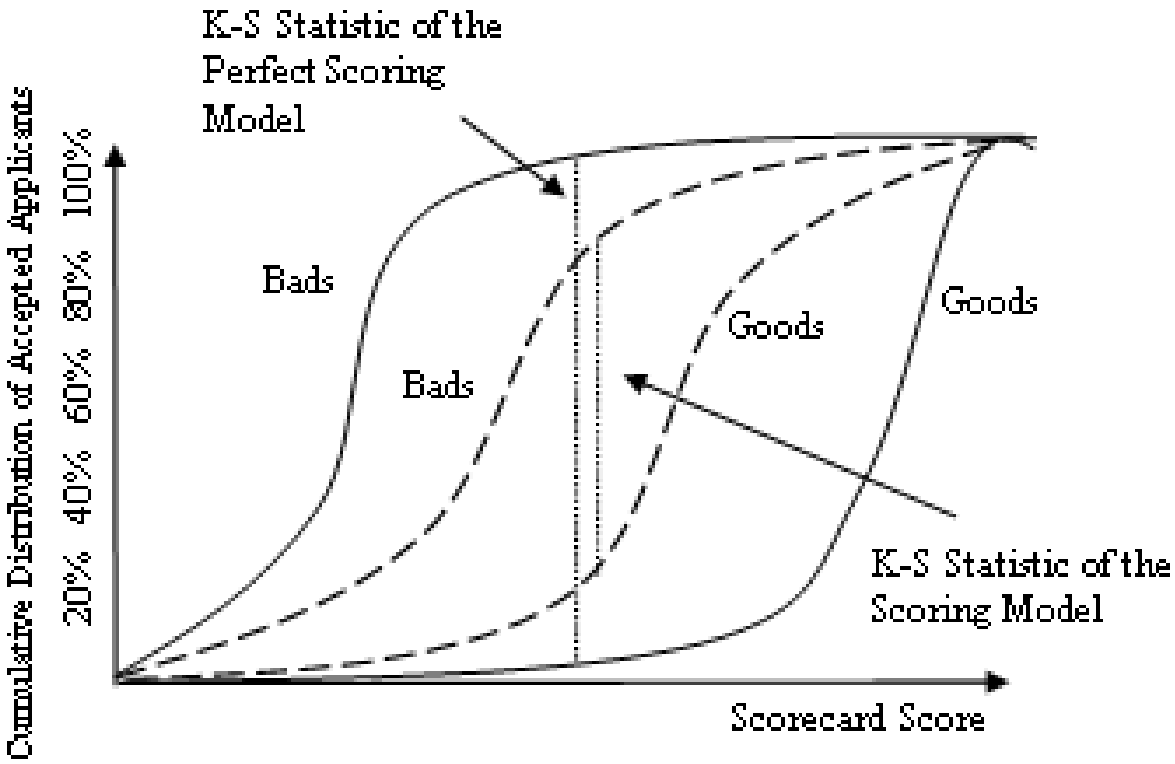
$$D = \max_{s} \left| B(s) - G(s) \right|$$

Donde, B(s) y G(s) son las funciones de distribución acumulativas de los scores de las poblaciones de bads y goods.

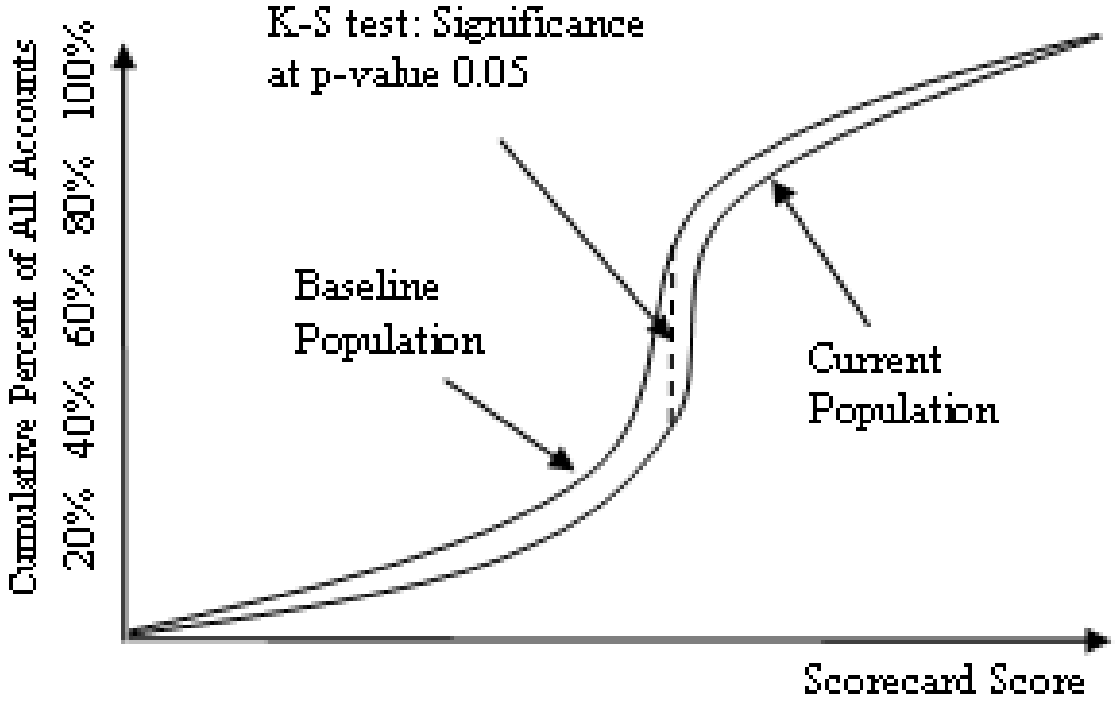
Prueba de K-S

Se demuestra que esta medida es equivalente a escoger un punto de corte que minimice la proporción de ceros mal clasificados y la proporción de unos mal clasificados, usando el punto de corte, el cual es un a función de los datos. Esto es, potencialmente muy engañoso, puesto que el punto de corte se debe elegir en base a los costos relativos de la mala clasificación.

Prueba de K-S


K-S Statistic of the Perfect Scoring Model
Bads
Bads
Goods
Goods
K-S Statistic of the Scoring Model
Cumulative Distribution of Accepted Applicants
20% 40% 60% 80% 100%
Scorecard Score

Prueba de K-S


K-S test: Significance
at p-value 0.05
Baseline
Population
Current
Population
Cumulative Percent of All Accounts
20% 40% 60% 80% 100%
Scorecard Score

Information Value

Esta medida es una distancia simétrica de Kullback-Leibler y como tal integra todos los valores posibles del score. Esto significa que ignora el punto de corte y usa información irrelevante acerca de los valores absolutos de los scores.

La potencia predictiva de una característica (variable), su habilidad para separar solicitudes de alto riesgo de las de bajo riesgo, se evalúa por:

$$IV = \sum_{Attribute} \left[\left(P_{Attribute}^{NoEvent} - P_{Attribute}^{Event} \right) * WOE_{Attribute} \right]$$

El valor de IV debe ser mayor que .02 para que una característica sea considerada para incluirla en el scorecard.

Information Value

Los valores de IV indican el nivel de importancia de la característica (variable). Por ejemplo:

IV < 0.1 → Weak

IV < 0.3 → Medium

IV < 0.5 → Strong

IV > 0.5 → Characteristic Overpredicting

El IV es la suma ponderada de las WOE de los atributos. El factor de ponderación es la diferencia entre la proporción de buenos y la de malos en el atributo respectivo.

Mean Difference Score

Es una prueba t de diferencias estandarizadas entre las medias de las distribuciones de ceros y unos. Esta ignora el punto de corte y el vector de ponderación (class priors) y usa información de los valores absolutos de los scores. Esta información es típicamente irrelevante, puesto que se tomará la misma acción para todos aquellos clientes con scores por encima del punto de corte, sin importar el score actual real y lo mismo aplica para los clientes que están por debajo del punto de corte.

Mean Difference Score

El estadígrafo Mean Difference (MD) se calcula por:

$$MD = \frac{\mu_G - \mu_B}{\sigma}$$

Donde,

μ_B es la media de la distribución de los scores de los goods.

μ_B es la media de la distribución de los scores de los bads.

σ es la media de las desviaciones estándar de ambas distribuciones.

Divergence

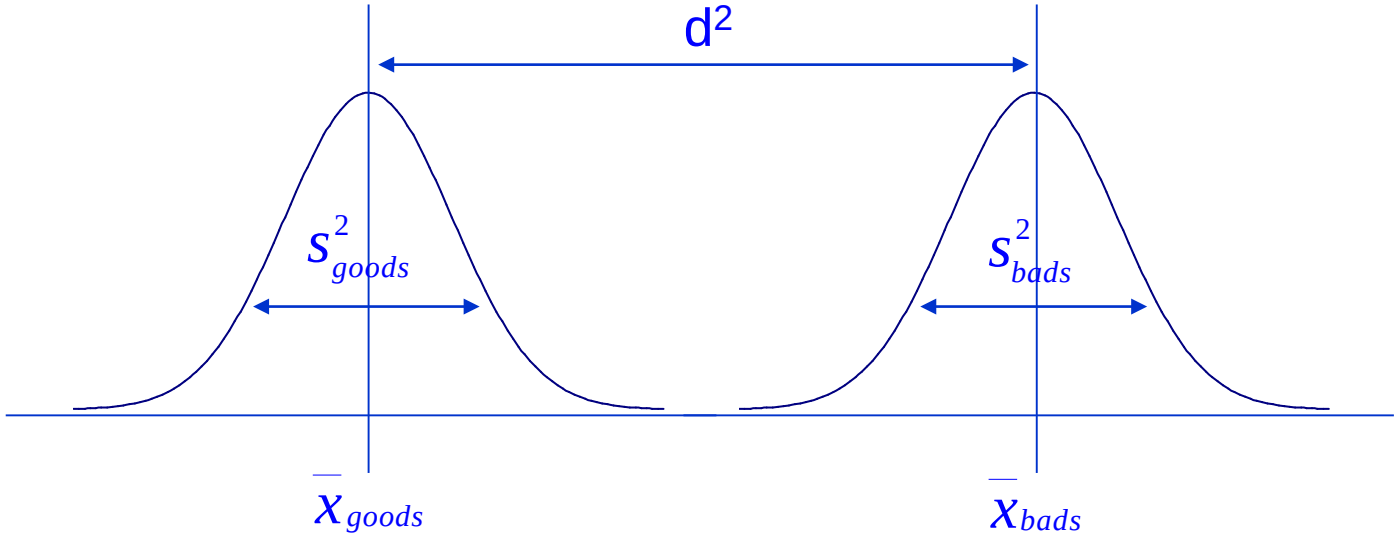
Es el cuadrado de la distancia entre las medias muestrales de las distribuciones de ceros y unos, referida a la dispersión promedio de las dos distribuciones:

$$DIV = \sum_{i=1}^{n} \frac{\left(\overline{x}_{sc.goods} - \overline{x}_{sc.bads}\right)^2}{\frac{1}{2}\left(s_{goods}^2 + s_{bads}^2\right)}$$

$i = \overline{1,n}$ - Número de características el scorecard.

Divergence

La mayoría de los scorecards tiene una divergence entre 0.5 y 3.0. Puede tomar valores [0, ∞]



Comparación medidas Tipo I

El Gini y el KS son invariantes a las transformaciones no lineales monótonas de la escala del score. La mean difference no es invariante. Esto significa que antes de ser usada esta medida tiene que ser transformada a una configuración estándar.

Una forma común de estandarizar es convertir la escala de modo que el Log(odds) de los están en la clase 1 sea una función lineal de los scores transformados. De hecho, si esto se hace sobre los scores de las escalas que son similares, los scorecards pueden ser comparados convenientemente, usando la pendiente de la línea Log(odds).

Comparación medidas Tipo I

El Gini y el KS y el IV no toman en cuenta los vectores de ponderación (class priors), simplemente comparan las distribuciones para los ceros y los unos, sin importar la frecuencia (cantidad) de unos y ceros.

Lo mismo ocurre para la mean difference, excepto que las dos clases se pueden ponderar diferenciadamente cuando se calcula la desviación estándar común para estandarizar la mean difference.

En particular esto significa que el estimado de la desviación estándar común es similar a la desviación estándar de la clase más grande, e influenciado relativamente poco por la clase más pequeña.

Comparación medidas Tipo I

Las medidas de desempeño del scorecard tipo I no toman en cuenta el tamaño de las clases y no está influenciadas por la falta de balance de las clases.

Comparación medidas Tipo II

Esta medidas usan solamente la clasificación verdadera de los clientes y si sus scores están por encima por debajo del punto de corte.

Una simplificación común es, suponer que las clasificaciones correctas no incurren en costos, porque se toman las acciones apropiadas.

Pero las clasificaciones incorrectas de “0” (a unos) incurre en un costo **Co**, debido a que se les aplica la acción de “1”. De la misma forma la clasificación incorrecta de “1” (a ceros) se incurre en un costo $\mathbf{C_1}$.

Comparación medidas Tipo II

Una medida apropiada del desempeño es el costo total de mal clasificados:

$$C_T = n_0 C_0 + n_1 C_1$$

Donde n_0 es el número de ceros mal clasificados y n_1 es el número de unos mal clasificados.

El costo total de la mala clasificación puede ser reinterpretado como un combinación de la “Especificidad” y la “Sensibilidad”, ponderadas apropiadamente por el tamaño de las clases de ceros y unos y los costos relativos de mala clasificación para los unos y los ceros.