Diseño y Construcción de Scorecards con Herramientas de Data Mining – Clase 1

Cristián Bravo R.
cbravo@dii.uchile.cl

Banco de Crédito e Inversiones
2 al 5 de Julio, 2011

Agenda del Módulo

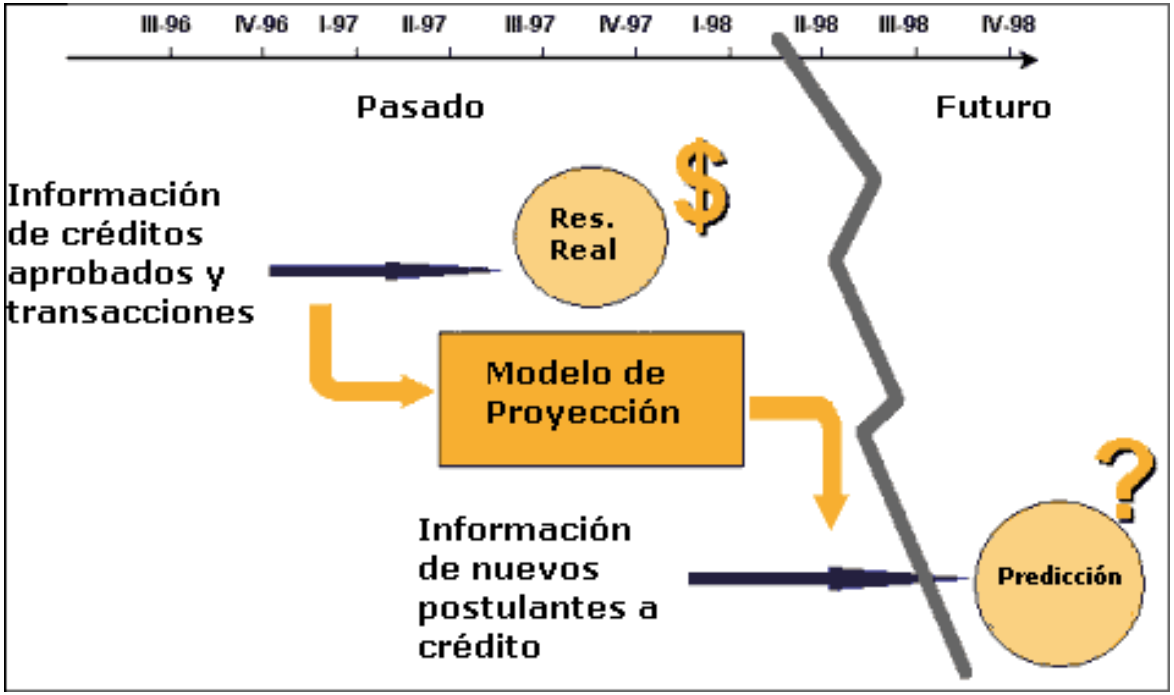
- Preparación de datos para generación de scorecards.
  - Selección de Variables.
  - Transformaciones Notables.
  - Segmentación de Universos.
- Modelos avanzados:
  - Análisis de Supervivencia.
  - Redes neuronales en Credit Scoring.
  - Modelos de optimización.
  - SVMs lineales y no lineales.
    - Transformación a probabilidad.
  - Consideraciones de Basilea II y uso avanzado de modelos.

Agenda del Módulo

- Inferencia en los Rechazados.
- Construcción de un Scorecard.
  - Transformación a log-odds.
- Puntos de Corte.
- Stress Testing.
- Seguimiento.

Introducción

Funcionamiento de Credit Scoring



$$f(x) = y$$

x: Variables descriptoras.
y: Variable objetivo (Default).

Scorecard

- Corresponde a una tabla, con puntos asociados a cada tramo, por cada variable.

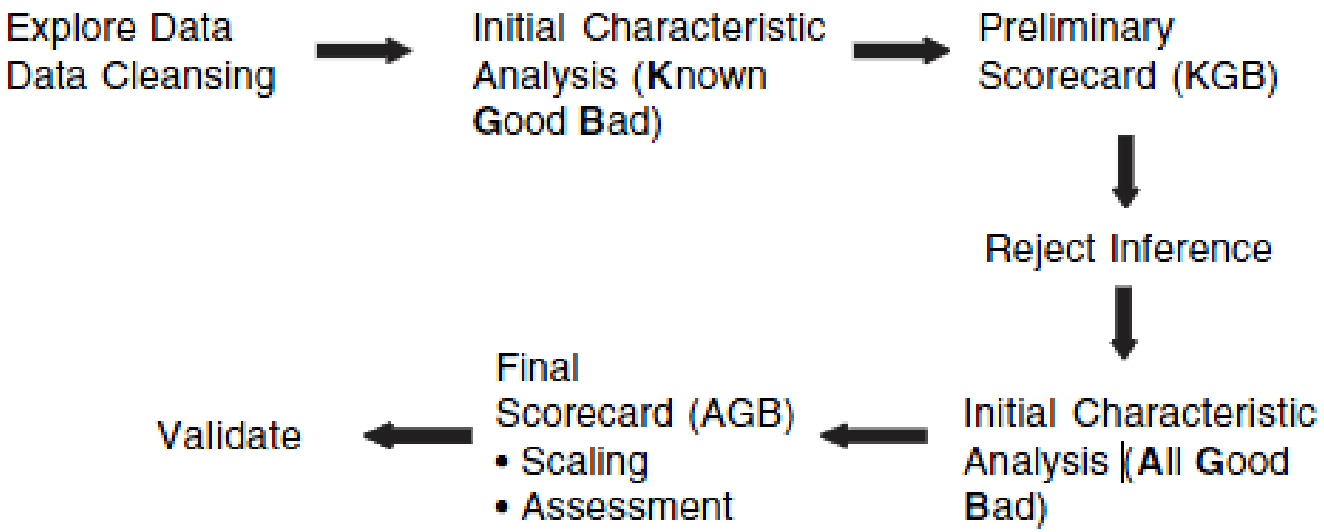
Characteristic Name	Attribute	Scorecard Points
AGE	. -> 23	63
AGE	23 -> 25	76
AGE	25 -> 28	79
AGE	28 -> 34	85
AGE	34 -> 46	94
AGE	46 -> 51	103
AGE	51 -> .	105
CARDS	"AMERICAN EXPRESS," "VISA OTHERS," "VISA MYBANK," "NO CREDIT CARDS"	80
CARDS	"CHEQUE CARD," "MASTERCARD/EUROC," "OTHER CREDIT CARD"	99
EC_CARD	0	86
EC_CARD	1	83
INCOME	. -> 500	93
INCOME	500 -> 1,550	81
INCOME	1,550 -> 1,850	75
INCOME	1,850 -> 2,550	80
INCOME	2,550 -> .	88
STATUS	"E," "T," "U"	79

¿Por qué utilizar un Scorecard?

- **Fácilmente comprensible** para personas que no tienen trasfondo en Data Mining.
  - Score final: Suma de los scores particulares.
- Las **razones para rechazar** un crédito pueden ser explicadas a partir de los resultados (puntaje alto en una variable particular).
- El desarrollo **NO es una caja negra**, y es ampliamente conocido.
- Es más **sencillo de monitorizar**.

Diseño de un Scorecard

- Los pasos para diseñar un scorecard.



Fuente: “Designing Credit Scorecards”. Naeem Sidiqi. 2006.

Requisitos

- Se necesita:
  - Base de datos con todos los créditos entregados a lo largo de N años, más todas las cuotas pagadas con sus fechas.
    - N: Normalmente son 5 años. Basilea II recomienda 20 (!).
    - Base de datos **incluye** las solicitudes rechazadas (fase Inferencia de Rechazados).
  - Definición de Buenos y Malos.
  - Técnica de Modelación (Regresión Logística).

Definición de Variable Objetivo

- La variable objetivo debe considerar:
  - **Evento** "Default".
    - Máximo legal: 90 días de mora en una cuota, para créditos de consumo. A este evento se le conoce como caer en Cartera Vencida.
    - En general se utiliza 90 en el mundo, pero hay bancos que tienen horizontes de 45.
    - Decisión **estratégica**. Define riesgo de forma estructural en la compañía.

Definición de Variable Objetivo

- **Horizonte de tiempo** de espera.
  - Por razones estadísticas (sesgo) y operacional (no descartar solicitudes), se decide cuánto tiempo esperar como máximo.
  - Usual: 12 meses.
  - Mejor opción: Segmentar por plazos y utilizar el mínimo plazo del segmento.
  - Problemas: Concept Drift. A medida que pasa el tiempo las distribuciones cambian de manera drástica.

Limpieza de Datos y Transformación

Exploración y Selección de Datos

- Exploración: Determinar aquellas **variables relevantes** para el problema.
  - Historia crediticia en la empresa.
    - Productos que posee, si tuvo créditos anteriores, si estos están cerrados o no, si pago cuotas a tiempo, etc.
    - Información del crédito actual **sirve sólo para segmentar**.
  - Monto adeudado en sistema (Base de Datos SBIF).
  - Moras y protestos (Base de Datos DICOM).
  - Características del cliente:
    - Estado civil, ingreso, bienes (colaterales!), vivienda, etc.

Exploración y Selección de Datos

- A considerar:
  - Poder predictivo esperado.
  - Credibilidad y confianza de la variable.
  - Facilidad de recolección.
  - Interpretabilidad.
  - Efectos de la intervención humana (¡¡monto!!).
  - Restricciones legales: NO usar etnia, sexo.
  - Creación de ratios basado en el conocimiento experto. (Ej: Variaciones en la deuda semestral).
  - Disponibilidad futura.

Limpieza de Datos

- Es necesario realizar limpieza de los datos en tres niveles.
  - A nivel de datos/variables: **Métodos clásicos** del Data Mining, limpiando datos nulos y eliminando variables concentradas.
  - A nivel de casos: Es necesario **excluir** algunos elementos que son dañinos para el modelo.
    - Ej: Cuentas VIP (compra de un edificio de 1MM UF), productos con muy poco mercado (créditos a 100 años).
  - A nivel de épocas: Se deben eliminar casos de épocas con **variabilidades de mercado muy notorias**.
    - Ej: Se cambió política para otorgar crédito a personas entre 18-80 y ahora es de 25-80. Eliminar casos entre 18-24.

Segmentación

- Con la base de datos limpia es posible comenzar a diseñar el scorecard.
- El primer caso es la **segmentación del mercado**.
- Se segmenta el mercado debido a:
  - No todos los clientes tienen el mismo perfil de riesgo.
  - No todos los productos tienen las mismas características.
- Regla de oro: Dividir para diferenciar el **funcionamiento en base a riesgo**.
  - Es decir, dividir para separar aquellos grupos que se caracterizan intrínsecamente de forma distinta, desde un punto de vista del riesgo que se corre.

Segmentación (II)

- Métodos:
  - **Por antigüedad de cliente:** Diferencia importante. Dos tipos de scores.
    - **Application Scoring:** No tengo información pasada del cliente.
      - Variables asociadas a quién el cliente es (sociodemográficas) junto a moras, protestos y variaciones (externas) de la deuda.
    - **Behavioral Scoring:** Tengo información de créditos pasados.
      - Las mismas variables anteriores, más toda la información de pagos, moras, créditos, etcétera, que posea la compañía.

Segmentación (III)

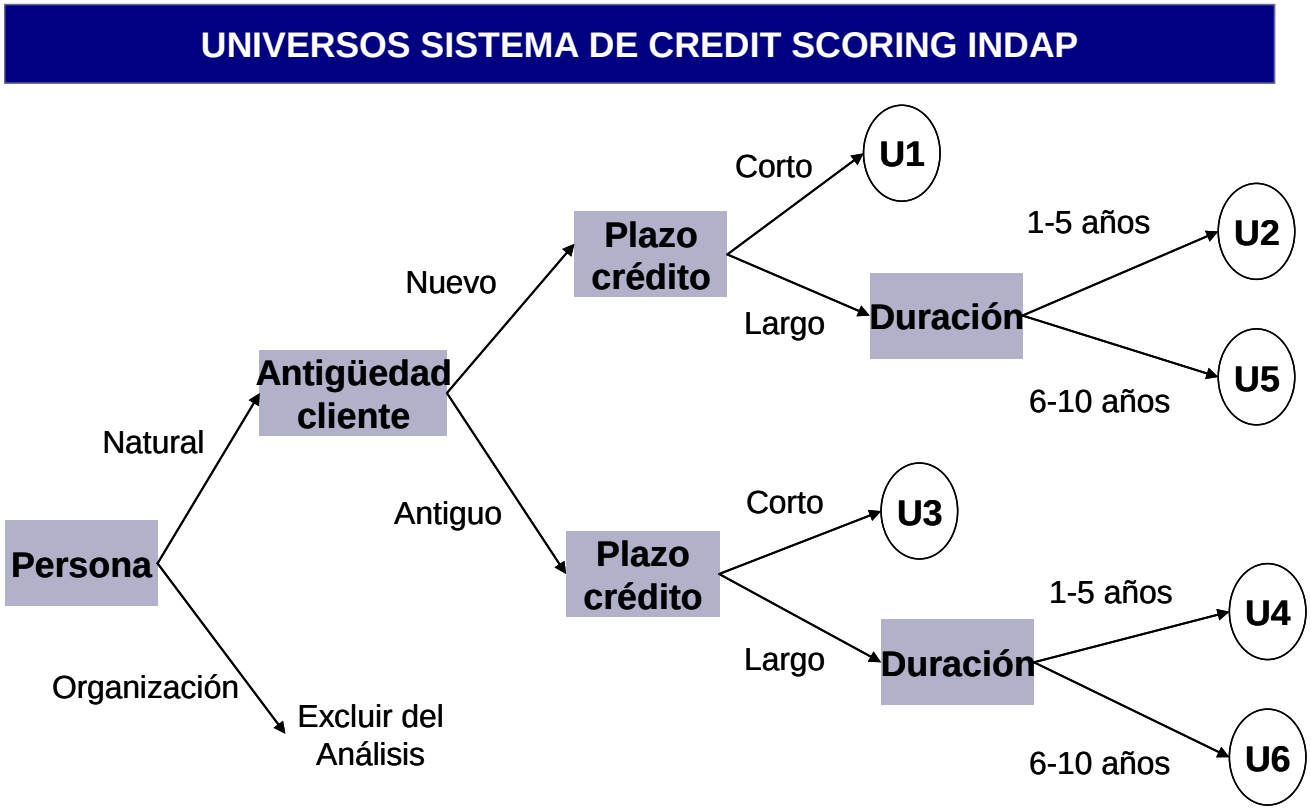
- Métodos:
  - **Por plazo y/o monto**: El monto de un crédito y el plazo modifican el comportamiento de los clientes.
    - No es lo mismo un crédito de MM$10 a 25 años, que uno a 6 meses.
    - En Chile, la SBIF obliga a utilizar el plazo como regresor o variable para segmentar.
  - Por métodos estadísticos:
    - Utilizar árboles de decisión.
  - **Por conocimiento de mercado**:
    - Se sabe que ciertos productos son "similares" desde un punto de vista del riesgo u operacional. Crear segmentos a partir de esto.

Segmentación: Consideraciones

- Existe un **trade-off** importante:
  - Más segmentos implican más modelos. Mayor costo asociado.
  - Menos segmentos implican menor discriminación. Mayor riesgo subyacente.
- Se deben armar segmentos que tengan un **tamaño suficiente** para la aplicación de modelos.
  - Regla general: por lo menos 2.000 casos por clase.

Segmentación: Ejemplo

- Créditos a Microempresarios.

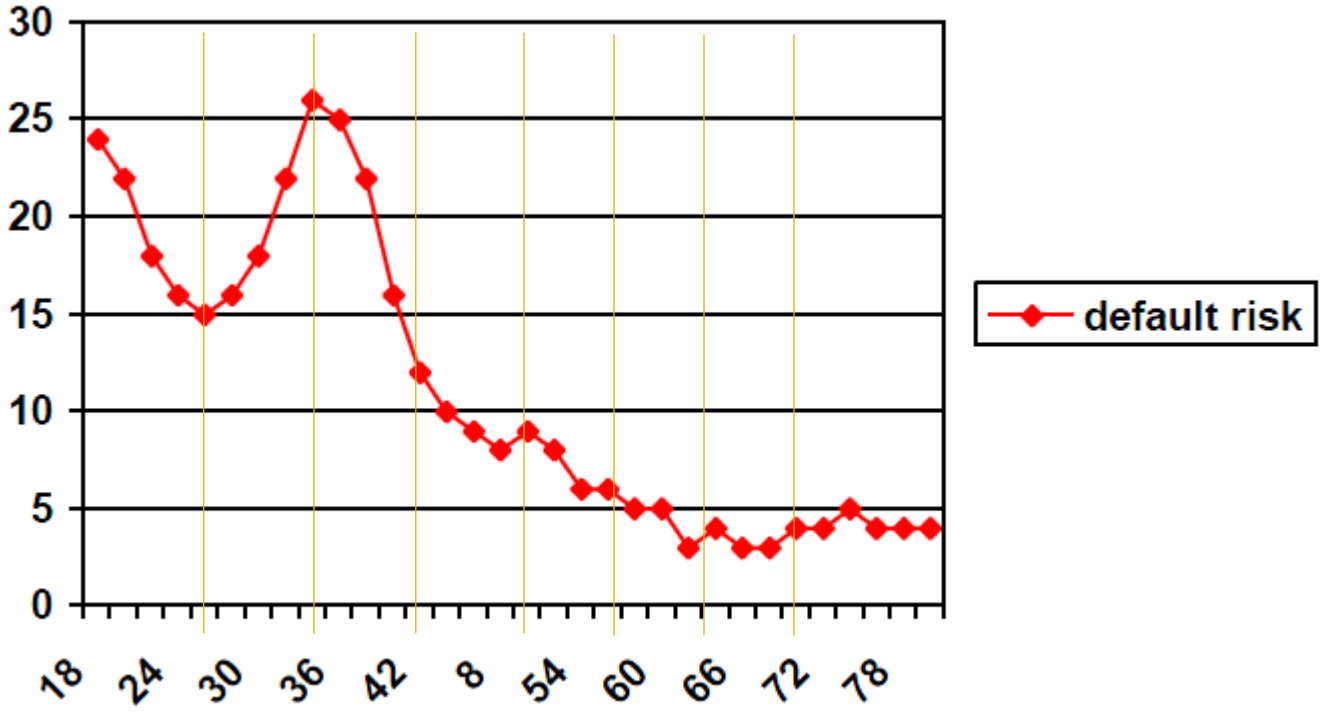


Transformación de Datos

- La transformación de datos sigue los lineamientos clásicos del KDD, salvo una característica adicional.
  - Las distintas variables deben ser **segmentadas en tramos**.
  - Razón: Por lo general los índices de riesgo no son continuos, sino asociados a tramos de variables. Además, se necesita para construir un scorecard.
- De minería de datos I: ¿Cómo se utilizan las variables categóricas?
  - Se debe generar una serie de **variables dummy**.

Transformación de Variables (II)

- Ejemplo: Edad.



Cómo Categorizar

- Hay dos maneras de categorizar:
  - Utilizando **variables dummy** (o ó 1) para cada categoría.
    - Método estándar, aunque no hace ninguna suposición sobre la diferencia entre un grupo y otro (deja eso al coeficiente de la regresión logística).
  - Utilizando el **Weight of Evidence**.

$$WOE = \ln\left(\frac{\%Buenos}{\%Malos}\right) * 100$$

    - Los porcentajes corresponde al total que caen en la categoría.

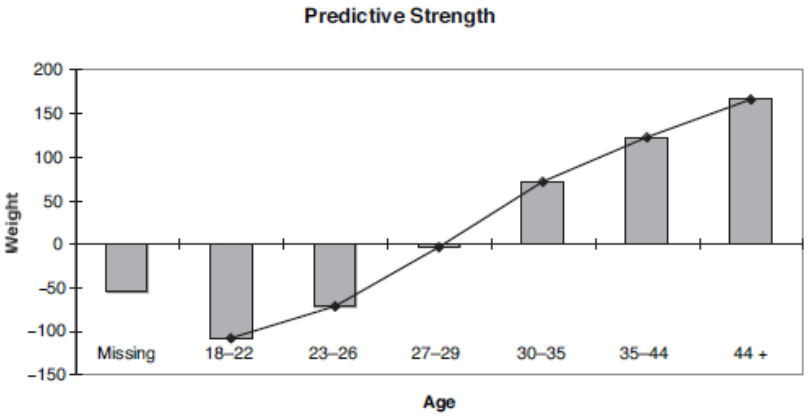
Cómo Categorizar: Reglas Generales

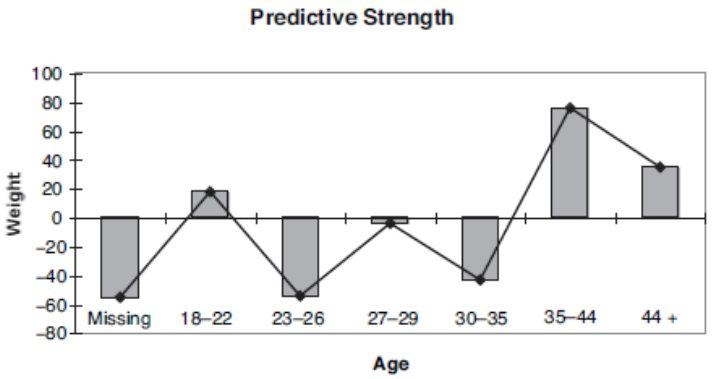
- **Agrupar por lógica** hasta que haya una cantidad razonable de casos por segmento.
  - 5% de los casos por lo menos en el segmento.
  - El segmento debes tener tanto casos buenos como malos.
- El WOE de cada grupo debe ser **significativamente distinto** de aquel correspondiente al grupo siguiente.

Weight of Evidence

Age	Count	Tot Distr	Goods	Distr Good	Bads	Distr Bad	Bad Rate	WOE
Missing	1,000	2.50%	860	2.38%	140	3.65%	14.00%	−42.719
18–22	4,000	10.00%	3,040	8.41%	960	25.00%	24.00%	−108.980
23–26	6,000	15.00%	4,920	13.61%	1,080	28.13%	18.00%	−72.613
27–29	9,000	22.50%	8,100	22.40%	900	23.44%	10.00%	−4.526
30–35	10,000	25.00%	9,500	26.27%	500	13.02%	5.00%	70.196
35–44	7,000	17.50%	6,800	18.81%	200	5.21%	2.86%	128.388
44+	3,000	7.50%	2,940	8.13%	60	1.56%	2.00%	164.934
Total	**40,000**	**100%**	**36,160**	**100%**	**3,840**	**100%**	**9.60%**	

- Este método también sirve para estimar si la variable es razonable o no.





Preparando la Muestra para Crear un Modelo

Antes de Crear los Modelos

- Con una base de datos limpia debemos :
  - Filtrar atributos irrelevantes **a nivel de problema**.
    - Este paso debe repetirse después a nivel de modelo.
  - Balancear base de datos.
    - Siempre existirá el problema que la base de datos esté desbalanceada (menos malos que buenos).
    - Ej: Caso INDAP.
      - Universo antiguos – corto plazo: 99% de buenos.
      - Universo muy largo plazo: 40% de buenos.

Selección de Atributos

- Con estos valores OK, es necesario crear el primer modelo.
- Procedimiento de Selección de atributos:
  - Test K-S y Chi-Cuadrado para variables.
    - K-S para variables continuas.
    - Chi-cuadrado para discretas.
    - Criterio: Si hay independencia ($p>0.05$) eliminar variable.
  - Árboles de decisión sobreajustados sirven también para seleccionar variables discriminantes.
    - Construir árbol con tamaño de hoja final 10 casos y todos los niveles que se puedan.
    - Criterio: Si variable NO aparece en el árbol, en ningún nivel, entonces se elimina.
    - Buen criterio para correlación multivariada.

Balanceo de Muestras

- Problema: Las muestras se encuentran desbalanceadas.
- Solución: **Depende del modelo**
  - En modelos lineales existen dos alternativas:
    - Balancear errores con un peso.
    - Mover el sesgo.
  - En modelos no lineales sólo es posible balancear errores.
    - Además, depende del modelo.
      - SVMs dispone de "Balanced Ridge".
      - Redes neuronales dispone de peso en error.
  - Siempre es posible balancear artificialmente muestra.

Balanceo de Muestras: Agregar Peso

- El peso (w_c) debe ser tal que:

$$\sum_{y_i=1} w_1 = \sum_{y_i=0} w_0$$

  - Además, el peso debe sumar un valor cercano a la cantidad de casos (problemas numéricos de la estimación).
- Pesos recomendados:
  - Peso de 1 a los la clase con menos casos.
  - Peso de "#malos/#buenos" para la clase con más casos.

Balanceo de Muestras: Mover el Sesgo

- La idea es mover la constante para igualar las muestras.
- Racionalidad: Las estimaciones (parámetros) no cambian **en modelos lineales** frente a ajuste de proporción.
  - En modelos no lineales esto NO aplica.
- Método SBIF: Si p_i es la proporción muestral y π_i la proporción poblacional (real):

$$\beta_0' = \beta_0 + \ln\left(\frac{\rho_1 * \pi_0}{\rho_0 * \pi_1}\right)$$

Modelos de Credit Scoring: Análisis de Supervivencia

Análisis de Supervivencia

- Otra forma de realizar credit scoring es cambiar el paradigma.
  - En vez de preguntarse si se es buen o mal pagador, preguntarse **cuánto tiempo pasará hasta que lo sea**.
- Ventajas:
  - Maneja bien datos censados (refinanciamientos, prepagos, etc.).
  - Evita tener que fijar un periodo de espera para el default.
  - Permite realizar análisis de utilidad (me da lo mismo si el default es en la cuota 59 de 60).
  - Facilita el aprovisionamiento.

Regresión de Cox

- Cox (1972) propuso un modelo para estimar este tipo de fenómenos.
- Supongamos que existe una función (hazards) dada por:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t}$$

  - Esto es, probabilidad que el cliente caiga en default en T dado que no cayó hasta t.
- Si sabemos vector de características **x** del cliente, ¿cómo las usamos?.

Regresión de Cox (II)

- Cox propuso el siguiente modelo:

$$h(t,x) = h_0(t)e^{\beta \cdot x}$$

  - $h_0(t)$: Riesgo base cuando regresores valen o.
- A este modelo se le llama **modelo de hazards proporcionales**, pues el riesgo va asociado a una escala fija base h.
- Este modelo puede ser estimado por máxima verosimilitud sin conocer los valores h_0.

Ajustando la Regresión para Credit Scoring

- El problema que surge es que en CS los tiempos para Default son fijos (sólo a fin de mes).
- Para ajustar el modelo se define:
  - t_i: Tiempos discretos (meses).
  - d_i: Número de defaulters en t_i.
  - D_i: Conjunto de clientes que fallaron en t_i.
  - $R(t_i, d_i)$: Conjunto de todos los subconjuntos de clientes tamaño d_i extraídos a partir del conjunto de clientes disponibles (defaulters y no defaulters) en t_i.

Ajustando la Regresión para Credit Scoring (II)

- Otros valores:
  - R: Elemento en $R(t_i, d_i)$, es decir, conjunto de tamaño d_i con clientes que pueden haber fallado en t_i.
  - S_R: Suma de los atributos **x** de cada cliente en R.
  - S_{Di}: Suma de los atributos **x** de cada cliente que falló en t_i.
- Al momento de entrenar, la regresión se balancear por **pesos en los errores**.

Ajustando la Regresión para Credit Scoring (II)

- Ahora podemos estimar los valores beta a partir de la ecuación de verosimilitud.
- Cox propone:

$$L_{\text{Cox}}(\beta) = \prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\sum_{R \in R(t_{(i)};\, d_i)} \exp(s'_R\beta)}.$$

- El denominador es MUY difícil de estimar.
- Existen otras propuestas:
  - Efron:

$$\prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\prod_{j=1}^{d_i}\left[\sum_{l \in R(t_{(i)})} \exp(x'_l\beta) - \frac{j-1}{d_i}\sum_{l \in D_i} \exp(x'_l\beta)\right]}$$

  - Breslow:

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\left[\sum_{l \in R(t_{(i)})} \exp(x'_l\beta)\right]^{d_i}}$$

Incorporando Características Dependientes del Tiempo

- Para calcular el score, **sólo se necesitan los odds**, que NO dependen de $h_0(t)$. Pero las variables PUEDEN depender del tiempo.
- Se necesitan dos aproximaciones:
  1. Chequear si las variables dependen del tiempo.
  2. Incorporar estas variables.
- La solución de (2) es directa: Incorporar variable $x_i \cdot t$ en el modelo.
- ¿Cómo saber cuáles variables incorporar?

Incorporando Características Dependientes del Tiempo (II)

- Para testear se utiliza el test de Harrel.
  - Hipótesis nula: ρ = 0. No hay correlación entre la variable y el tiempo.
  - Estadístico:

$$Z = \rho \sqrt{\frac{n_u - 2}{1 - \rho^2}}$$

    - n_u: Casos defaulters (no censados).
  - La correlación se calcula sobre los Residuos de Schoenfeld y el tiempo t:

$$r_{ik} = x_{ik} - E[x_{ik}|R_{t_i}]$$

Calculando Tabla h_t: Estimador de Kaplan-Meier

- Falta calcular la tabla de factores $h_0(t)$.
- Usamos en este caso los estimadores de Kaplan – Meier, que cuentan los casos donde existen "muertes" y los divide sobre los sobrevivientes.
- Supongamos que ya obtuvimos el vector β de parámetros asociado a cada una de las variables.

Calculando Tabla h_t: Estimador de Kaplan-Meier (II)

- Consideremos los siguientes conjuntos.
  - R_i: Conjunto de personas que aún no han caído en default previo a t_i.
  - D_i: Conjunto de personas que pertenecen a R_{t_i} tal que fallaron en t_i.
- Se cumple que:

$$\sum_{j \in R_i} e^{\beta \cdot x_j} = \sum_{j \in D_i} \frac{e^{\beta \cdot x_j}}{1 - \left(1 - h_0(t_i)\right)^{e^{\beta \cdot x_j}}}$$

- Los software especializados resuelven esta ecuación automáticamente.

Calculando Probabilidad de Default

- Ahora que se cuenta con $h_0(t)$ y β, ¿cómo calcular la probabilidad?
- Si se sigue la definición estándar (que una persona no haya caído en default hasta un tiempo t^*) entonces:

$$p(x_j, t^*) = e^{-\int_0^{t^*} h(u,x)du} = e^{\left(-e^{\beta \cdot x_j} \sum_{u \le t^*} h_0(u)\right)}$$

$$p\left(x_j, t^*\right) = \left(e^{-\sum_{u \le t^*} h_0(u)}\right)^{e^{\beta \cdot x_j}} = (S_0(t^*))^{e^{\beta \cdot x_j}}$$

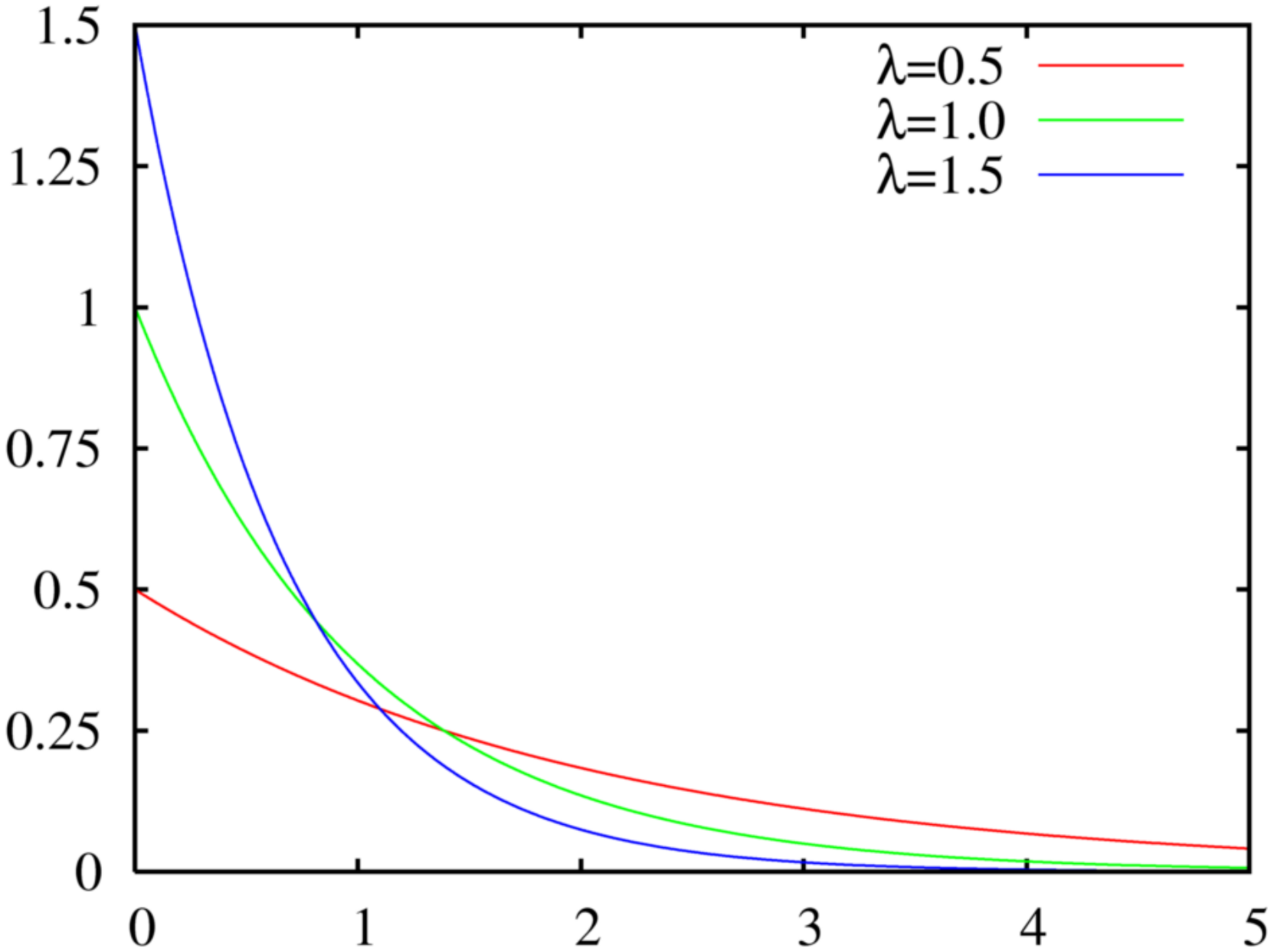
- Donde $S_0(t)$ se le conoce como la función de supervivencia de t.

Análisis de Especificación

- El modelo debe ser validado para revisar si existe algún problema en el ajuste.
- Se utilizan los Residuos de Cox-Snell.

$$r_{C_k} = \exp(\beta x_k) \sum_{t \le t_i} h_0(t) = -\log(S_k(t_i))$$

- Donde a S se le conoce como la función de sobrevivencia.
- Estos residuos deben tener **distribución exponencial con media uno**.


1.5
1.25
1
0.75
0.5
0.25
0
0
1
2
3
4
5
λ=0.5
λ=1.0
λ=1.5

Validando el Modelo

- Se pueden usar las técnicas clásicas para comparar los modelos.
  - Accuracy: Definir como cliente malo a aquel que posee una esperanza menor a 12 meses.
  - Curvas ROC: Por corte de plazo de espera.