INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

CENTRO DE FINANZAS

Seguimiento en Modelos de Regresión Logística Model Follow-Up In Logistic Regression Models

Cristián Bravo, Sebastián Maldonado y Richard Weber

El Centro de Finanzas cuenta con el significativo apoyo del Banco de Crédito e Inversiones BCI

fcfm

SEGUIMIENTO EN MODELOS DE REGRESIÓN LOGÍSTICA

MODEL FOLLOW-UP IN LOGISTIC REGRESSION MODELS[1,2]

CRISTIAN BRAVO ROMAN[1]
SEBASTIAN MALDONADO[1]
RICHARD WEBER[1]

[1] Dpto. de Ingeniería Industrial, Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.
Avda. República 701. Santiago. Chile.
cbravo@dii.uchile.cl, semaldon@ing.uchile.cl, rweber@dii.uchile.cl

RESUMEN

La gran mayoría de los proyectos de minería de datos que utilizan la metodología KDD en la vida real entregan solamente soluciones estáticas que con el paso del tiempo pierden la capacidad de explicar los fenómenos para los que fueron construidos inicialmente. Presentamos un marco teórico-práctico que permite realizar un seguimiento cercano a los modelos para determinar el momento donde éstos deben ser actualizados, manteniendo un estricto control sobre la evolución de los mismos, las variables presentes en ellos y los cambios relevantes que pueden ocurrir en la población desde que fueron inicialmente diseñados. Los tests estadísticos incluyen tests clásicos como las pruebas de Kolmogorov-Smirnov o la prueba de Chi-Cuadrado para medir los cambios en las medias de las variables en los modelos, más un test novedoso diseñado en base a la distribución de los coeficientes en los modelos y la desviación estándar observada de las variables que permite medir cuándo la población ha cambiado más allá de los intervalos de confianza definidos por los parámetros iniciales. La metodología fue puesta a prueba utilizando las bases de datos reales de dos proyectos de Credit Scoring a microempresarios realizados entre los años 2007 y 2008, con muy buenos resultados.

Palabras clave: Regresión Logística, Seguimiento, Credit Scoring.

ABSTRACT

Most data mining projects in real life applications give as a result only static solutions which, in time, lose their inherent capacity to explain the phenomena they were originally built for. We introduce an theoretical-practical framework that allows to closely follow up logistic regression models to determine the moment when they must be updated, maintaining an strict control over their evolution, the variables in them and relevant changes that can occur in the population since they were originally designed. The statistical test presented include classical tests such as Kolmogorov-Smirnov and Chi-Squared statistic to measure changes in means of the variables present in the models, plus a novel test designed from the distribution of the models coefficients

---

[1] Esta es la versión de archivo interno del paper publicado en la Revista de Ingeniería Industrial de la Universidad del Bío-Bío. La versión original de esta publicación se puede encontrar en http://www.ici.ubiobio.cl/revista/index.php?option=com_docman&task=doc_download&gid=90&&Itemid=15 .

[2] Por favor cite este paper como sigue: Cristián Bravo, Sebastián Maldonado, Richard Weber (2009). Seguimiento en Modelos de Regresión Logística. Revista de Ingeniería Industrial 8(2):31-44.

that allows to measure the moment when a population has changed more than the confidence intervals defined from the original parameters. The methodology was tested using the databases from two real world micro-entrepreneurs credit scoring projects developed between the years 2007 and 2008, with very good results.

Key Words: Logistic Regression, Change Detection, Credit Scoring.

1. INTRODUCCIÓN

Dentro de la aplicación práctica de modelos de minería de datos y en general de la aplicación del proceso de descubrimiento de conocimiento en bases de datos (Famili *et al.* 1997 y Fayyad *et al.* 1996), o KDD por sus siglas en inglés, se tiene un procedimiento altamente estructurado y probado como efectivo al intentar modelar fenómenos que se presentan en la práctica con fundamento estadístico. Sin embargo, dentro de este proceso no se incluye el problema que surge al considerar que los procesos operacionales que crean los registros en una entidad siguen funcionando, por lo que es relevante considerar qué cambios han sido registrados en estas variables a medida que pasa el tiempo.

A la rama de la minería de datos, la estadística y la econometría que se encarga de estudiar los procesos de evolución de las variables en los modelos se le conoce como "Detección del Cambio" (*Change Detection*), en los que está enmarcado este trabajo. El objetivo de este trabajo es presentar un marco general de trabajo para realizar el estudio de los cambios en los modelos basado en tests estadísticos clásicos y empíricos, de tal manera que permitan tener señales de alerta para seleccionar el mejor momento en que un modelo estático debe ser actualizado para el caso particular de los modelos de regresión logística.

La regresión logística (Hosmer & Lemeshow, 2000) corresponde a una técnica estadística clásica ampliamente utilizada en la práctica, cubriendo una amplia gama de aplicaciones, desde la bioestadística (Lachin, 2000) hasta el *credit scoring* (Thomas, 2006). Es por ello que la alerta temprana del momento donde se deben realizar cambios es muy importante, siendo el enfoque de este trabajo.

La estructura del escrito es la siguiente: el siguiente capítulo introduce la regresión logística y el procedimiento KDD para la estimación de modelos estadísticos, el capítulo tres presenta el marco teórico para realizar seguimiento a los modelos, describiendo los problemas que surgen al realizar estas estimaciones, las experiencias previas y los test estadísticos que serán utilizados; el capítulo cuatro presenta los resultados experimentales de estos tests en proyectos de *credit scoring* reales realizados por los autores. Finalmente, se presentan las conclusiones de este trabajo.

2. REGRESIÓN LOGÍSTICA Y PROCESO KDD

2.1. Proceso *Knowledge Discovery in Databases*

El proceso KDD corresponde a una metodología ordenada para recolectar conocimiento a partir de un conjunto de datos presentes en bases de datos. El proceso se puede describir en los siguientes pasos:

- Selección de Datos: Corresponde a seleccionar los distintos orígenes de datos de los que se disponen. En general se clasifican en orígenes internos (bases de datos de la entidad), externos (fuentes de datos que provengan de otras organizaciones) y variables generadas, que corresponden a todos aquellos indicadores que sean generados a partir de los datos disponibles en la otras dos fuentes de datos.

- Preprocesamiento: La segunda fase del proceso KDD corresponde a analizar las variables que se han obtenido de tal forma de eliminar todas aquellas que no sean utilizables por un modelo. En este paso se realiza la eliminación de variables concentradas o con alta cantidad de datos nulos, la imputación de datos nulos, la eliminación o reemplazo de variables fuera de rango y la eliminación de atributos que no son discriminantes con respecto al fenómeno.
- Transformación: En esta etapa del proceso se realizan todas aquellas transformaciones que permiten que los datos sean utilizables por el modelo. Cada modelo estadístico posee distintos requerimientos acerca de qué tipo de datos acepta, como por ejemplo sólo variables entre $[0,1]$ o variables sin datos nulos. Los pasos de la transformación incluyen el escalamiento, la agregación de atributos, la creación de variables *dummy* (binarias) para representar atributos categóricos, etc.
- Data Mining: En esta etapa se realiza la estimación del modelo estadístico elegido. Corresponde a estimar los valores de los distintos parámetros utilizando distintos algoritmos que intentan, en general, minimizar alguna medida de eficiencia del ajuste del modelo a los datos. Los modelos en general se clasifican en supervisados (se conocen las clases a las que pueden pertenecer los objetos) o no supervisados (no se conocen estas clases), por lo que las medidas de eficiencia pueden corresponder a cuán bien el modelo se ajusta a los datos de entrada (modelos no supervisados) o cuán bien los datos se ajustan al fenómeno que se intenta modelar (modelos supervisados).
- Interpretación y evaluación: El paso final del proceso KDD corresponde a analizar las salidas obtenidas buscando revisar si el modelo es satisfactorio al momento de explicar el fenómeno o interpretar los datos, las medidas de ajuste que se posean y las interpretaciones que se puedan extraer de él.

Un resumen del proceso se puede observar en la figura 1, extraída de Mackinon y Glick (1998).

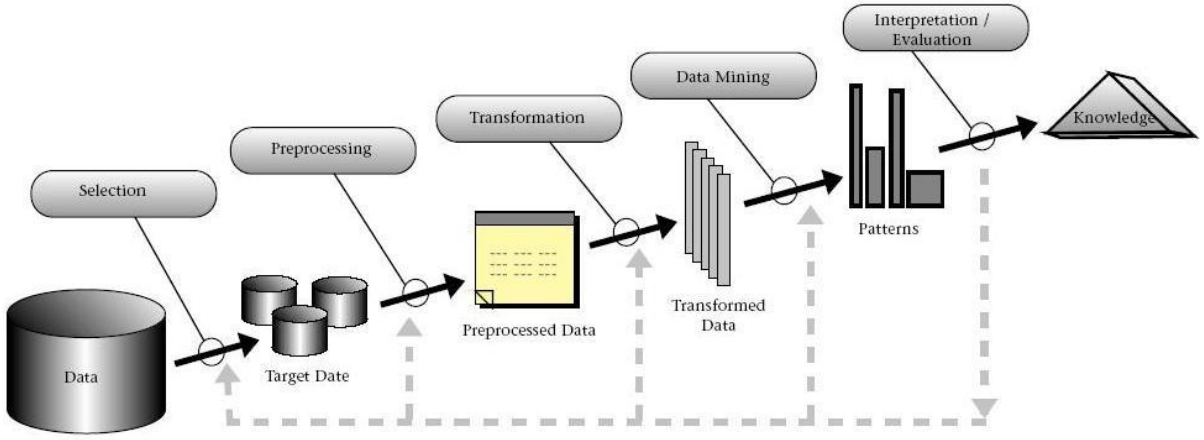

Figura N°1. Proceso KDD.

2.2. Regresión Logística

La regresión logística corresponde a un modelo estadístico que busca encontrar la probabilidad continua $y \in [0,1]$ de ocurrencia de un evento en base a utilizar regresores $x \in \mathbb{R}^n$, los que son extraídos a partir de los datos en base al proceso KDD. El modelo calcula una función

logística (1) que estima la probabilidad de ocurrencia del fenómeno utilizando como datos de entrada los vectores x_i, con $i \in \{1, \dots, N\}$ la muestra de N objetos que se poseen de muestra, y las etiquetas (salidas) reales del modelo $y_i \in \{0,1\}$.

$$p(x_i) = \frac{1}{1 + e^{\beta_0 + \beta^t \cdot x}} \tag{1}$$

El proceso de calibración de los parámetros se realiza utilizando el método de máxima verosimilitud (Eliason, 1993) el cual busca maximizar la probabilidad estimada de obtener los resultados categorizados según y_i. En particular, se maximiza la siguiente función de verosimilitud:

$$\mathcal{L}(\beta) = \prod_{i\ /\ y_i=1} \frac{1}{1+e^{\beta_0+\beta^t \cdot x}} \prod_{i\ /\ y_i=0} \left(1 - \frac{1}{1+e^{\beta_0+\beta^t \cdot x}}\right) \tag{2}$$

A partir de esta función se alcanzan estimadores que son asintóticamente eficientes, insesgados y distribuidos normalmente. Los parámetros de la regresión logística se distribuyen asintóticamente entonces siguiendo una distribución normal de media $\hat{\beta}$ (el parámetro obtenido) y desviación estándar σ_β, lo que es relevante para el desarrollo de los test estadísticos del presente trabajo.

El uso que se le da comúnmente al modelo pasa por seleccionar un punto de corte p_c (valor para la probabilidad *p*) tal que para valores mayores a ese punto de corte se selecciona que la etiqueta esperada para la observación es uno ($p(x_i) \geq p_c \Rightarrow y_i = 1$) y en caso contrario se le asigna el valor cero, produciéndose así la discriminación activa.

3. MODELO DE SEGUIMIENTO PARA REGRESIÓN LOGÍSTICA

3.1. El Problema del Seguimiento a Modelos

La definición del problema corresponde a determinar cuándo ha ocurrido un movimiento significativo de las variables presentes en el modelo que amerite realizar una actualización de los parámetros calibrados en el modelo. La pregunta que surge es entonces ¿qué significa un "movimiento significativo"? Se definen para ello tres posibles cambios que pueden suceder en el modelo que inducen un cambio en los parámetros de las variables.

- Capacidad discriminante de las variables: Para que una variable sea incluida en un modelo de regresión logística es necesario que esta discrimine entre las dos clases en estudio. Por discriminar se entiende el hecho que la distribución (media, desviación, etc.) de la variable sea distinta para cada uno de las clases, de tal forma que a distintos valores de ella se obtengan distintas capacidades discriminantes. Este entonces corresponde a la primera condición que debe ser chequeada al momento de revisar cambios en el modelo.
- Distribución de las variables: Los supuestos básicos del modelo indican que cada una de las observaciones (x_i) es extraída de un conjunto X tal que se distribuye en base a una función $f(x)$ desconocida, pero idéntica para cada elemento. Este supuesto trae como consecuencia que los parámetros extraídos tengan aplicabilidad sólo mientras se tienen variables extraídas a partir de esta distribución, sin embargo, las distribuciones de las variables tienden a cambiar en el tiempo, pues la población modifica sus patrones de comportamiento. Este patrón se observa sobre todo en fenómenos sociales, como el riesgo

crediticio, dónde empíricamente cada dos años se observan cambios en la población suficientes para impactar en el modelo (Thomas, 2000).

- Capacidad discriminante del modelo en su conjunto: El cambio más drástico que puede tener una población puede volver el modelo en su conjunto no discriminante, si bien cada variable por separado puede mantener capacidad discriminante. Esto puede suceder por ejemplo si los cambios en la población inducen correlaciones destructivas entre las variables, lo que provoca que la probabilidad predicha se concentre en valores donde ninguna decisión puede ser tomada.

En este trabajo nos centraremos en la manera de detectar de forma temprana los dos primeros problemas. Por temprana, se refiere a intentar detectar estos cambios de forma previa a conocer las salidas reales de una muestra, conociendo solo los datos de la predicción y las variables de entrada. El tercer problema involucra necesariamente conocer las salidas reales de la muestra, pues así es posible determinar si las probabilidades obtenidas ya no presentan diferencias notables entre un caso y otro.

3.2. Experiencias Previas

El problema de seguimiento a modelos ha sido abordado previamente por investigadores, pero no siguiendo los mismos enfoques presentados en este trabajo. En particular, el problema de los cambios en las variables ha sido enfrentado de forma amplia al tratar con series de tiempo, donde por ejemplo se ha buscado identificar cambios en las variables cuando se está en presencia de datos correlacionados en serie (Jones *et al.*, 1970) o se han desarrollado modelos que permiten generar una nueva serie de tiempo fraccionada en términos del cambio en un parámetro (Keohg, 2001).

En el caso de modelos de clasificación (como la regresión logística), las experiencias han seguido el camino de buscar métodos que se actualicen automáticamente a medida que se incorporen datos a la muestra, siguiendo procesos como actualizar los parámetros en modelos basados en métodos que utilicen kernels (Yang, 2007) o métodos semi-incrementales de aprendizaje (Shen, 1997), pero estos modelos no enfrentan en realidad el problema de detectar cuándo ocurren cambios significativos en la población, sino que enfrentan los cambios en el ambiente en base a actualizar en línea los parámetros asociados.

Dentro del conocimiento de los autores, el enfoque más próximo al aquí detallado corresponde al trabajo realizado por Zeira *et al.* (2005), el cual desarrolla test estadísticos para el caso general de modelos en el cual el error de validación se distribuye normal y las variables poseen un comportamiento tal que se puedan construir estadísticos tal que distribuyan según una distribución χ^2. La diferencia con este enfoque es que en este caso los modelos se generarán centrados en los movimientos de los parámetros inducidos por las variables.

3.3. Tests Estadísticos Propuestos

Se desarrollan tres test estadísticos para probar los cambios en la población, en primer lugar se estudian los cambios a nivel de capacidad discriminante de la variable. Para efectos notación, se considerará que existe una muestra original $X_0 \subset \mathbb{R}^J$ de tamaño N_0 asociada a etiquetas conocidas $y(x), x \in X_0$, que estimaron un modelo de regresión logística con parámetros $\hat{\beta}$ y una nueva muestra $X' \subset \mathbb{R}^J$ de tamaño N que inducen probabilidades calculadas $p(x'), x' \in X'$ y etiquetas (estimadas) $y'(x')$.

3.3.1. Estadísticos para Cambio en la Capacidad Discriminante

Para estimar la capacidad discriminantes de las variables se utilizarán test estadísticos clásicos para determinar si la distribución es idéntica entre los casos originales y los casos nuevos. En particular, se utilizará el test χ^2 de independencia y el test de Kolmogorov – Smirnov (Hines & Montgomery, 1990) para probar si una variable ha dejado de tener capacidad discriminante.

Sea una variable $x_i^{'j} \in x_i'$, $x'_i \in X'$ tal que sus valores son discretos, entonces, si dividimos la muestra X' completa en base a la predicción $y'(x')$ se crean dos grupos de observaciones. En este caso se utiliza la salida estimada como la manera de dividir los casos, lo que tiene sentido, ya que con esto se estudia si las clases predichas poseen diferencias para la variable en estudio.

Para probar independencia, se calcula una tabla que posee como filas los distintos valores que puede tomar la variable x_j, mientras en las columnas posee las clases 0 o 1. En cada celda $O_{r,c}$ se localiza la cantidad de casos que poseen el valor dado por fila y columna, para cada una de las muestras. Finalmente, suponiendo que la variable x_j tiene r valores distintos, se construye el estadístico dado por:

$$X^2 = \sum_{l=1}^{r}\sum_{c=1}^{2}\frac{(O_{l,c}-E_{l,c})^2}{E_{l,c}}, E_{r,c} = \frac{(\sum_{c=1}^{2} O_{r,c}\sum_{l}^{r} O_{l,c})}{N} \quad (3)$$

Este estadístico posee distribución χ^2 con $(r-1)$ grados de libertad. La hipótesis nula corresponde a que las muestras son independientes (no poseen efectos en la discriminación entre las clases), por lo que se espera rechazar esta hipótesis al 95% de confianza.

Para las variables x_j que sean continuas se utiliza el test de Kolmogorov-Smirnov (K-S) el cual permite estudiar cuando existen cambios en la distribución de la variable. La ventaja de este test es que no realiza suposición alguna sobre la distribución de la variable x_j. En este caso se necesita nuevamente dividir la muestra según las etiquetas $y'(x_i')$ definidas y se construye un estadístico que en base a la máxima diferencia entre los histogramas acumulados en la muestra, según la figura 1.

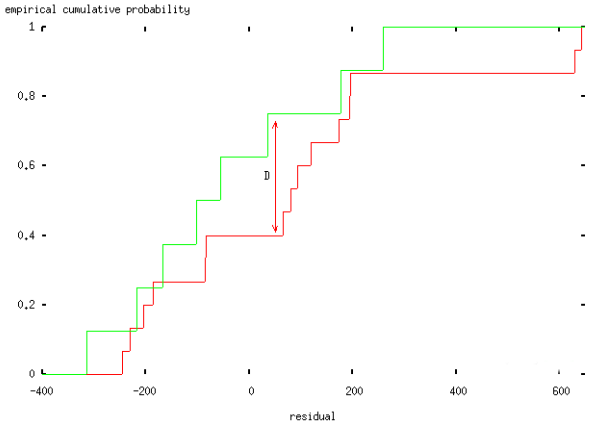

Figura N°1: Ejemplo de Distribución del Estadístico K-S.

El estadístico D (la máxima distancia) se distribuye en forma asintótica en base a la distribución de Kolmogorov, la cual se utiliza para estimar la significancia del parámetro. Esta distribución corresponde a la expresión de la ecuación 4.

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{2i^2x^2} \quad (4)$$

Actualmente la mayoría de los *softwares* especializados permiten calcular asintóticamente este estadístico.

La aplicación de este método permite realizar un seguimiento cercano a cuándo la capacidad discriminante se ha perdido, pues se está simultáneamente buscando si la división realizada presenta una segmentación adecuada (al utilizar las salidas $y'(x')$) y adicionalmente si la variable presenta una discriminación adecuada para esta segmentación.

3.3.2. Estadísticos para Cambio en la Distribución de las Variables

Se busca ahora detectar si ha existido un cambio en la distribución de las variables suficiente para inducir un cambio en los valores de los parámetros, sin embargo, un nuevo problema surge de esto pues los modelos si son bien construidos, son robustos frente a cambios pequeños en las variables. La pregunta que surge es entonces ¿cuándo un cambio en la distribución de las variables es suficiente para inducir un cambio en los parámetros mayor a lo que permite el modelo? A manera de ejemplo se presenta la figura 2, dónde se observan dos cambios distintos en las distribuciones.

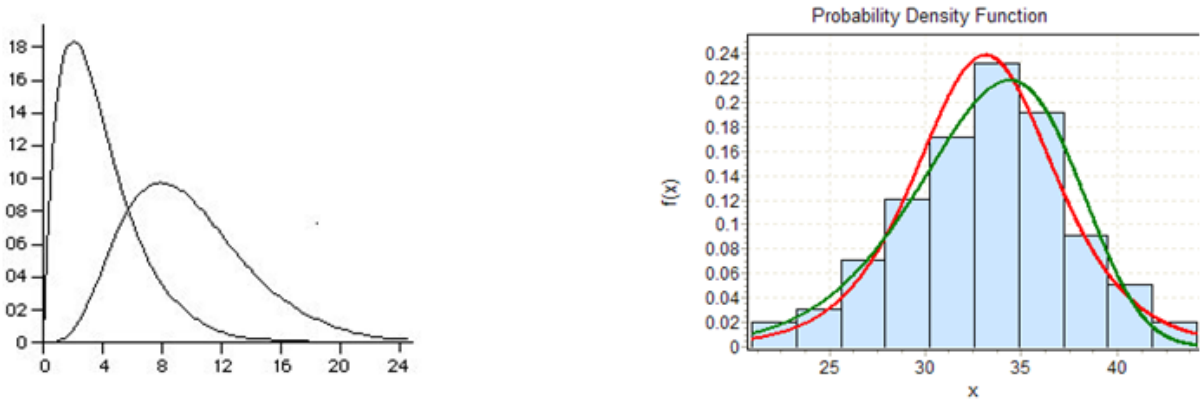

Figura Nº2: Cambios posibles en una distribución. A la izquierda un cambio radical, a la derecha un cambio pequeño.

A partir de estas figuras se podría pensar que en la figura de la izquierda, lógicamente el cambio fue suficientemente drástico para inducir un cambio en los parámetros, mientras a la derecha, posiblemente no. Para formalizar este fenómeno, se utilizará que cada parámetro β_j asociado a una variable x_j posee un intervalo de confianza dado por $\beta_j \in [\beta_j^{inf}, \beta_j^{sup}]$ donde se cree que puede yacer el valor real (poblacional) del parámetro. En este caso entonces, el supuesto que se realiza es que si la población cambia de tal manera que el nuevo parámetro estimado está fuera de este intervalo de confianza, entonces el modelo no es válido para esa muestra nueva.

Dos posible enfoques para tratar esto se desarrollaron, ambos basados en este concepto. En primer lugar, si se considera la variación máxima permitida para el parámetro, esta corresponde a:

$$C_j \in \left[\frac{\beta_j^{inf}}{\widehat{\beta_j}}, \frac{\beta_j^{sup}}{\widehat{\beta_j}}\right] \quad (5)$$

Entonces una posible medida empírica del cambio máximo de la variable puede expresarse en términos de la media de las variables nuevas observadas. Si la media inicial de x_j es $\overline{x_j}$ y de la nueva muestra es $\overline{x_j'}$ entonces:

$$\frac{\overline{x_j'}}{\overline{x_j}} \in \left[\frac{\beta_j^{inf}}{\widehat{\beta_j}}, \frac{\beta_j^{sup}}{\widehat{\beta_j}}\right] \quad (6)$$

La expresión (6) determina un umbral para considerar que la variable aún no varía lo suficiente. Este test, empírico y sencillo, entrega muy buenos resultados con respecto a los cambios en la variable, según se muestra en el capítulo 4.

El otro enfoque seguido, mucho más riguroso, pero al mismo tiempo de mayor costo, corresponde a generar un nuevo parámetro $\widehat{\beta_j'}$ tal que se pueda testear de forma rigurosa si la variable posee un valor dado en el intervalo de confianza. Para estimar estos valores se requiere seguir los siguientes pasos para la muestra X':

1. Estimar nuevos parámetros $\widehat{\beta'}$ asociados a la muestra X': Esto se realiza siguiendo el algoritmo de máxima verosimilitud y utilizar como etiquetas los valores $y'(x')$ estimados a partir de la función inicial.

2. Para cada parámetro, se estima además una desviación estándar $\widehat{\sigma_{\beta'}}$ la que indica el error cometido a partir de esta variable. Esta estimación es directa a partir de la regresión.

Si la muestra es suficientemente grande, entonces se tiene que:

$$\frac{\beta'_j - \beta_{ref}}{\sigma_{\beta'}} \sim t \quad (7)$$

Ahora es posible testear directamente si el valor de los parámetros pertenece al intervalo. Se realizan dos test independientes:

$$\begin{matrix} H_0: \beta'_j = \beta_j^{inf} \\ H_a: \beta'_j < \beta_j^{inf} \end{matrix} \quad y \quad \begin{matrix} H_0: \beta'_j = \beta_j^{sup} \\ H_a: \beta'_j > \beta_j^{sup} \end{matrix} \quad (8)$$

Esta aplicación permite revisar si los nuevos parámetros se encuentran al interior del intervalo de confianza determinado por los parámetros antiguos, utilizando para ello la nueva estimación realizada. Se espera no rechazar las hipótesis nulas para ambos casos, dónde el valor crítico para el estadístico t con infinitos grados de libertad está dado por 1,645 para el test unilateral para el límite superior y de -1,645 para el test de unilateral asociado al límite inferior. Esta aplicación debe cumplir con los siguientes requisitos:

- Se debe contar con suficientes casos en la muestra. Esto es importante por dos razones, en primer lugar, el número debe ser lo suficientemente grande para poder estimar parámetros, y en segundo lugar, la expresión (7) sólo se cumple si existe una cantidad alta de casos en muestra, es decir, el estimador t efectivamente presenta infinitos grados de

libertad. En general, estos test de seguimiento se recomienda realizarlos cada tres o seis meses, de tal forma de acumular suficientes casos en muestra.

- Se deben almacenar los datos de cada caso de forma metódica. Esta es una recomendación obligatoria para cualquier aplicación real, los nuevos casos deben ser almacenados manteniendo sus variables, la probabilidad predicha y la clase seleccionada o el punto de corte utilizado para estimarla.

4. RESULTADOS EXPERIMENTALES

La base de datos utilizada para medir los cambios corresponde a la de una institución real que otorga créditos la cual conocía *a priori* que un cambio en las variables había entorpecido la operación del modelo de *credit scoring* original. Se cuentan con los datos para distintas variables, 13 en total, con más de 300.000 observaciones, lo que permite que se cumplan a cabalidad los supuestos para la aplicación de las pruebas estadísticas.

Un ejemplo de cambio corresponde al observado en la figura 3, donde se dividió la base de datos para dos periodos, 2000-04 y 2005-07, según se requería por la entidad que encargó el estudio.

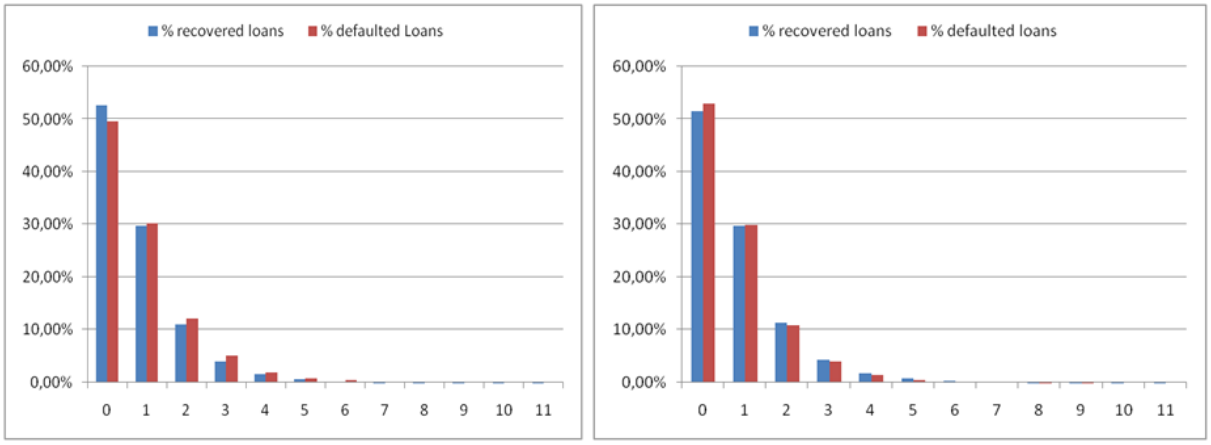

Figura Nº 3: Comparación de las distribuciones para una variable en los períodos 2000-04 (izq.) y 2005-07 (der.).

En la imagen se observa como la distribución de la variable cambió a lo largo de los periodos, pues en un comienzo la proporción entre barras azules (créditos pagados) y rojas (créditos no pagados) varía significativamente entre los casos. Para probar la solución propuesta se estimaron los intervalos de confianza para las variables según el mejor modelo encontrado, para luego comparar los intervalos obtenidos con la variable original siguiendo el modelo definido por la ecuación (6), los resultados y estadísticos se pueden observar en la tabla 1.

Se observa en la tabla que las variables 10 y 11 presentan cambios, mientras el resto no se observan valores demasiado alejados. El test es muy poco sensible a cambios en las medias, como era de esperar, pero presenta una propiedad interesante: si la variable tiene originalmente mucha desviación estándar, entonces el test es menos sensible a variaciones, mientras que si no la tiene los intervalos serán más acotados. Una segunda propiedad, negativa en este caso, son las variables que presentan un movimiento conjunto (variables discretas cuya distribución es el porcentaje acumulado) se tiene que la prueba, al no considerar estas relaciones, no dimensiona los cambios de manera correcta. Esto sí es considerado por el segundo test.

El siguiente paso es probar el test t de la ecuación (8) para los intervalos. Los resultados se observan en la tabla 2, en color marcados los valores que son significativos (hipótesis nula rechazada).

En este caso, se observa como el test permite reproducir los cambios de muy buena manera, considerando también las variables categóricas expresadas en porcentajes (variables 3 a 5 y las variables 6 y 7), ambas con cambios drásticos y notorios en sus coeficientes. Ambos tests coinciden en que existe un cambio significativo para la variable 10 y difieren en la variable 11, aunque la primera prueba está muy cerca del límite inferior y la segunda está apenas dentro de él. Se observan estos cambios en la distribución en los histogramas de la figura 4.

Tabla 1: Resultados test empírico de cambio de variable.

Variable	Media Nueva	Media Orig.	Cambio	Cambio Inf.	Cambio Sup.
Var_1	9,56	8,40	1,14	0,75	1,25
Var_2	51,64	51,35	1,01	0,53	1,47
Var_3	53,94	57,75	0,93	0,11	1,89
Var_4	20,90	16,13	1,30	0,13	1,87
Var_5	3,30	3,42	0,96	0,36	1,64
Var_6	29,15	37,15	0,78	0,03	1,97
Var_7	41,58	36,05	1,15	0,66	1,34
Var_8	3,04	2,79	1,09	0,87	1,13
Var_9	-0,01	-0,01	0,69	-0,07	2,07
Var_10	7,49	5,89	1,27	0,82	1,18
Var_11	0,08	0,12	0,64	0,83	1,17
Var_12	0,46	0,41	1,12	0,75	1,25
Var_13	0,14	0,15	0,95	0,84	1,16

Claramente la variable 10 presenta un cambio en la distribución que es detectado correctamente por ambas pruebas. El caso de la variable 11 posee algunos cambios, mucho menos notorios, lo que se observa en la figura 5.

En la figura 5 se observa un cambio en la distribución de casos que no pagan el crédito (barras de color rojo), pero este cambio es pequeño y, acorde al test dos, no lo suficientemente significativo para inutilizar el parámetro, lo que parece razonable. Sin embargo, este cambio ya es suficientemente grande para ser recogido por el test uno, por lo que potencialmente esta variable, si cambia más, afectará de forma notoria el modelo.

Las pruebas, trabajando en conjunto, permiten detectar cambios relevantes en los modelos, cumpliendo su cometido.

Tabla 2: Prueba de cambios en parámetros ajuste nuevo.

Variable	Beta Estimado	Error Estandar	Beta Original	Lim. Inf.	Lim Sup.	Estadístico Inf.	Estadístico Sup
Var_1	,696	,100	,794	,592	,995	1,03	-2,98
Var_2	,340	,089	,394	,208	,580	1,48	-2,71
Var_3	-,785	,077	-,177	-,335	-,019	-5,83	-9,92
Var_4	-1,136	,138	-,364	-,682	-,047	-3,30	-7,92
Var_5	-,728	,080	-,264	-,432	-,096	-3,71	-7,93
Var_6	1,150	,069	,151	,005	,298	16,56	12,32
Var_7	1,845	,089	,495	,325	,665	17,01	13,21
Var_8	-,576	,037	-,618	-,697	-,540	3,25	-0,98
Var_9	,467	,085	,057	-,004	,117	5,54	4,11
Var_10	,132	,011	,085	,069	,100	5,75	2,97
Var_11	-1,115	,091	-1,353	-1,583	-1,123	5,14	0,08
Var_12	-,513	,068	-,553	-,694	-,412	2,67	-1,48
Var_13	-1,506	,103	-1,548	-1,795	-1,302	2,81	-1,98

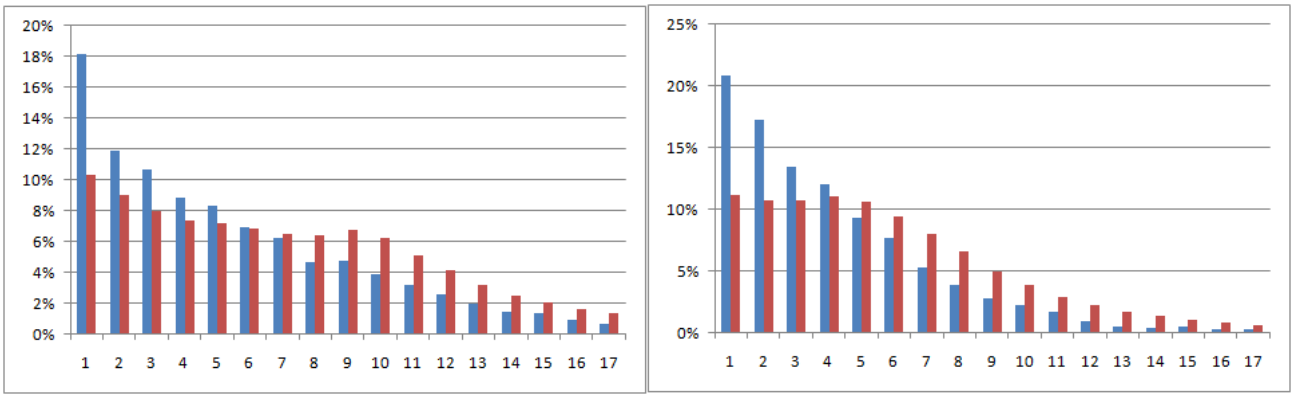

Figura Nº 4: Distribución para variable 10 para años 2000-04 (izq.) y 2005-07 (der.)

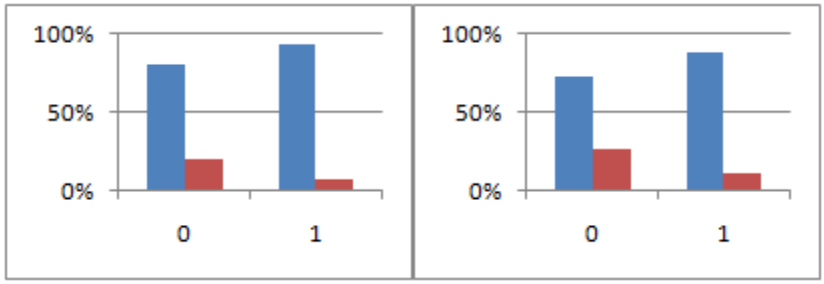

Figura Nº 5: Cambio en la variable 11 para años 2000-04 (izq.) y 2005-07 (izq.)

CONCLUSIONES

En el presente trabajo se presentaron una serie de pruebas estadísticas que permiten medir de forma anticipada los problemas que surgen cuando las variables en el modelo cambian lo suficiente para producir una pérdida en la capacidad discriminante de los modelos de regresión logística. Esta preocupación, obviada comúnmente en la literatura, es relevante para todas las áreas donde es necesario aplicar estos modelos, pues en caso de no medirla se utilizarán los modelos hasta que existan pruebas patentes del fallo en su aplicación, con una pérdida (ya sea monetaria o asociada a la aplicación particular) que puede ser evitada.

Los cambios a los que hay que prestar atención corresponden a aquellos que tienen impacto en los parámetros de los modelos, por lo que los supuestos asociados a la estimación de los parámetros son los que definirán qué se debe medir. En este caso, la capacidad discriminante de las variables y la distribución subyacente a ellas son las dos aristas que se deben enfrentar.

En el caso de la capacidad discriminante, simples test de χ^2 y de Kolmogorov-Smirnov permiten detectar de forma temprana los cambios que suceden con las capacidades discriminantes, pues rápida y eficientemente detectan diferencias notables entre las distribuciones asociadas a la variable que están en una u otra clase predicha.

Para la distribución de las variables se presentan dos pruebas, una empírica y otra con fundamento estadístico, para probar cuando estos cambios han provocado una modificación relevante en los modelos. El test empírico, si bien simple, se comporta bien detectando cambios extremos en las variables. Por otro lado, el test estadístico presenta una mucha mayor sensibilidad a los cambios, pero viene asociado a un costo computacional mucho mayor, al requerir estimar un nuevo modelo de tal manera de contrastar los parámetros. Ambas pruebas son no paramétricas y aplicables a cualquier distribución original, lo que las hace una potente herramienta.

La prueba en un caso real de los tests anteriores arroja que es necesaria una medición conjunta, pues el test empírico no logra capturar de buena manera los cambios que se dan en variables que por su naturaleza presentan correlación en la distribución (variables categóricas). De todos modos, las pruebas permiten estudiar de forma efectiva los cambios en las variables y cuando estos cambios afectan el modelo, objetivo principal de este trabajo.

Como trabajo futuro, están en estudio los efectos de la división en base a las etiquetas predichas y los potenciales sesgos que ellas poseen, más sus correcciones en estas pruebas, de tal manera de alcanzar pruebas estadísticas mucho más robustas.

AGRADECIMIENTOS

Se agradece al Instituto Chileno de Investigación de Operaciones (www.ichio.cl) por su apoyo en el desarrollo de este trabajo. Bravo y Maldonado además agradecen especialmente al Doctorado en Sistemas de Ingeniería (www.sistemasdeingenieria.cl/doctorado) por su apoyo, a CONICYT por las becas otorgadas que permiten esta investigación.

REFERENCIAS

Eliason, S. (1993). Maximum Likelihood Estimation: Logic and Practice. Sage Publishing Inc.

Famili, A., Shen, W.-M., Weber, R., Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis, 1, 3-23.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Hines, W. H. y Montgomery, D. C. (1990). Probability and Statistics in Engineering and Management Sciences. 3ª Edición. Wiley.

Hosmer, D. y Lemeshow, H. (2000). Applied Logistic Regression. John Wiley & Sons.

Jones, R. H., Crowell, D. H. y Kapuniai, L. E. (1970). Change Detection Model for Serially Correlated Data. Biometrica, 26, 269-280.

Keogh, E., Chu, S., Hart, D. y Pazzanini, M. (2001). An Online Algorithm for Segmenting Time Series. In: Proceedings of IEEE International Conference on Data Mining (pp. 289-296).

Lachin, J. (2000). Biostatistical Methods: The Assessment of Relative Risks. Wiley-Interscience.

Mackinon, M. y Glick, N. Data Mining and Knowledge Discovery in Databases – An Overview.

Shen, W. M. (1997). An Active and Semi-Incremental Algorithm for Learning Decision Lists. Technical Report, USC-ISI-97. Information Science Institute, University of Southern California.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149-172.

Thomas, L. C. (2006). Credit Scoring: The State of the Art. FORESIGHT: International Journal of Applied Forecasting, 3, 33-37.

Yang, Y. (2007). Adaptative Credit Scoring with Kernel Methods. European Journal of Operations Research, 183(3), 1521-1536.

Zeira, G., Last, M. y Maimon, O. (2005). Segmentation on Continuous Data Streams Based on a Change Detection Methodology. In: Advanced Techniques in Knowledge Discovery and Data Mining (pp. 103-126). Springer.