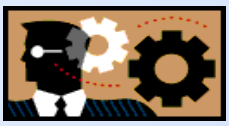
Colegio Nacional de Actuarios, A. C.

Actuarios Trabajando: Revista Mexicana de Investigación Actuarial Aplicada.

q_x μ_x d_x

$_tV_x$  A_x

l_x $\ddot{a}_x$ p_x

AÑO 4, NUM. 6

FEBRERO 2011

ACTUARIOS TRABAJANDO: REVISTA MEXICANA DE INVESTIGACION ACTUARIAL APLICADA

COORDINADOR Y EDITOR:

Dr. Gabriel Núñez Antonio
Comité de Investigación y Desarrollo Actuarial
Profesor Visitante del departamento de Estadística
de la Universidad Carlos III de Madrid.

gab.nunezantonio@gmail.com

REVISORES ASOCIADOS:

CONTENIDO

Estimados lectores y colegas:

En este documento ponemos a su disposición el sexto número de nuestra revista ***"Actuarios Trabajando: Revista Mexicana de Investigación Actuarial Aplicada".*** El objetivo sigue siendo contar con un medio que permita promover, ampliar y difundir el interés en las diversas áreas de desarrollo del actuario en México y en el mundo.

En esta ocasión contamos con un primer artículo el cual fue ganador de un premio en el "Concurso Francisco Aranda" en su edición 2010. Este concurso es organizado bianualmente por la Asociación Mexicana de Estadística. El artículo muestra una aplicación de la estadística a un aspecto del riesgo crediticio. Además, les presentamos un último artículo donde se aplican técnicas estadísticas modernas en el área de Mercadotecnia.

Continuamos exhortándolos a que nos envíen sus colaboraciones para futuros números, así como sus opiniones a los artículos ya publicados, además de sus sugerencias para mejorar la revista.

Quiero agradecer la preferencia por *Actuarios Trabajando* a todos los autores que enviaron su trabajo. Como siempre a la Dra. Ángeles Yáñez y a Christian Martello por su contribución a la realización de la revista y al continuo desarrollo de este proyecto.

Esperando que disfruten la lectura les mando un cordial saludo.

--
Dr. Gabriel Núñez Antonio.

Editor.
Comité de Investigación y Desarrollo Actuarial

Crédito al Consumo: La estadística aplicada a un problema de riesgo crediticio

Soraida Nieto Murillo
gussynm@hotmail.com
Blanca Rosa Pérez Salvador
psbr@xanum.uam.mx
Universidad Autónoma Metropolitana Iztapalapa
Av. San Rafael Atlixco No. 186, Col Vicentina, Iztapalapa D. F. CP 09340
Teléfono: (+52 55) 5804-4654
José Fernando Soriano Flores
BBVA-Bancomer / Banca Popular
Av. Canal de Miramontes No. 2600 local E-11 P.A.,
Col. Avante, Coyoacan DF, CP 04460
Teléfono: (+52 55) 55991553
josefernando.soriano@bbva.bancomer.com

Resumen

En este trabajo se presenta una visión general del scoring estadístico utilizado como una herramienta para discriminar los buenos de los malos prospectos al otorgar un crédito. Se trata de una metodología que pronostica el riesgo de incumplimiento de pagos durante una ventana de tiempo determinada. Se basa en el análisis de dos tipos de datos referente a los clientes, datos demográficos y datos de buró de crédito. Esta metodología da un marco para la decisión final en el otorgamiento o rechazo de la solicitud de crédito.

Palabras claves: Índice de Gini, Prueba Kolmogorov Smirnov, Matriz de transición, Regresión logística, Scorecard.

I. Introducción

Una de las necesidades más importantes de las instituciones crediticias es contar con criterios confiables para determinar a quien y de qué monto deben otorgarse un crédito; de ahí la razón por la que es importante tener un instrumento con el cual medir el riesgo que se corre al otorgar un crédito y poder reducir lo más posible este riesgo al aceptar a nuevos clientes.

El *scoring* experto o estadístico es una herramienta que sirve para discriminar a los buenos de los malos clientes. Se trata de una metodología que pronostica el riesgo que un cliente caiga en incumplimiento de pagos. Se basa en el análisis de dos tipos de datos: datos demográficos como pueden ser edad, sexo, ingresos, situación laboral, y datos de buró de crédito, como pueden ser su historial crediticio y su comportamiento en cuanto a la morosidad de pagos.

En los últimos años se ha incrementado el número de solicitantes de crédito al consumo, [ver Moreno (2008)] razón por la cual, el riesgo de otorgar créditos a clientes que dejarán de pagar aumenta, por esta razón el *credit scoring* es una herramienta indispensable en las instituciones de crédito.

En este artículo se revisa las diferentes etapas que forman el proceso del *credit scoring*. Se compone de cuatro sesiones, la primera sesión es esta introducción, la segunda sesión estudia a las tarjetas de crédito, ya que es el instrumento de crédito al consumo más popular, la tercera sesión revisa los diferentes pasos que dan origen a una *scorecard*. La última sesión corresponde a las conclusiones del trabajo.

II. Tarjetas de crédito

Las tarjetas de crédito son tarjetas de plástico personalizadas con las que se otorga créditos pequeños. La tarjeta de crédito sustituye al dinero en efectivo. Son utilizadas cuando las personas se ven en la necesidad de adquirir productos, servicios o cancelar deudas y no cuentan con efectivo a la mano. Son utilizadas en cajeros automáticos y medios electrónicos en los comercios. Las compras realizadas con tarjeta generalmente son acumuladas en un periodo de un mes (Fecha de Corte) y debe ser pagada toda la deuda o bien solo un porcentaje (Pago mínimo) de las deuda en la fecha límite de pago (Generalmente 20 días después de la fecha de corte). Se puede gastar hasta un límite concedido y el crédito se repone automáticamente una vez se ha pagado la deuda de la tarjeta (Créditos revolventes).

La tarjeta de crédito ha seguido un proceso evolutivo que se remonta hacia el año de 1914 en Estados Unidos, donde fue creada. Las instituciones financieras y las necesidades del mercado fueron dándole el formato que tienen el día de hoy. En México las tarjetas de crédito empezaron su desarrollo a partir de 1956, sin reglamentos específicos aplicables. Es hasta 1967 que la Secretaría de Hacienda y Crédito Público dictó un reglamento para tarjetas de crédito bancarias, aplicable exclusivamente a instituciones bancarias de depósito. Este reglamento no limitó a otras instituciones que sin tener carácter de institución de crédito promovieron su difusión en el país. Actualmente se intenta reglamentar su uso para proteger a los consumidores debido a la gran demanda de las tarjetas de crédito y al aumento en la cantidad de clientes que dejan de pagar (pasan a cartera vencida). Según informes de la Comisión Nacional Bancaria de Valores (CNBV) y del Banco de México (BdeM), la

cartera vencida en el crédito al consumo sufrió un incremento de 50.3 por ciento con respecto al monto registrado en diciembre de 2007 [ver González (2008)]. Al término del año 2009, el índice de morosidad, se elevó en 8.75 por ciento [ver Lino (2009)]. En los créditos personales el índice de morosidad aumentó de 5.93 a 6.53 por ciento de enero a diciembre de 2009 [ver Zúñiga et al (2009)].

II. 1 Posibles condiciones de un cliente

Cuando el cliente paga su adeudo total entre la fecha de corte y la fecha límite de pago, no se le cobran intereses de mora sobre su saldo ni intereses por financiamiento, (alrededor de 20 días), y se considera un cliente al corriente "*current*". Si el cliente paga el mínimo entre la fecha de corte y la fecha límite de pago, el monto no cubierto de la deuda se carga al siguiente periodo. El cliente está al corriente pero si se le cobran intereses por financiamiento.

Si el cliente no cubre su deuda, ni tampoco realiza el pago mínimo a la fecha límite de pago, se le empieza a localizar para recordarle su adeudo e incurre en gastos de cobranza, esto se realiza entre la fecha límite de pago y la próxima fecha de corte, aproximadamente una semana y se le conoce como tiempo "*prevent*". Los gastos de cobranza se cargarán al siguiente periodo de corte. Si el cliente paga la totalidad o el mínimo en *prevent* pasa a uno o dos estatus atrás. No se le aplicarán intereses moratorios, pero no evitará el cobro de "call center" por localizarlo. Si durante el periodo de corte se paga menos del mínimo, no se considera este pago como cumplimiento de la obligación, por lo que avanza a un pago vencido. Se le cargan intereses de moratoria sobre todo el monto y toma el estatus de cliente moroso. En la fecha de facturación se calcula el nuevo saldo y se empiezan a contar los días de retraso. Se envía al

grupo de clientes que tienen de 1 a 29 días de moratoria, llamado “Bucket 1" o canasta B1. Si se paga más del mínimo y menos del total del saldo facturado se cargan intereses. Así, la cartera de crédito se clasifica en las siguientes categorías:

- Bucket 0 (B0) 0 días de mora (al corriente)
- Bucket 1 (B1) 1 a 29 días de mora
- Bucket 2 (B2) 30 a 59 días de mora
- Bucket 3 (B3) 60 a 89 días de mora
- Bucket 4 (B4) 90 a 119 días de mora
- Bucket 5 (B5) 120 a 149 días de mora
- Bucket 6 (B6) 150 a 179 días de mora
- Bucket 7 (B7) Mayor o igual a 180 días de mora (Charge Off, Write Off, Perdida)

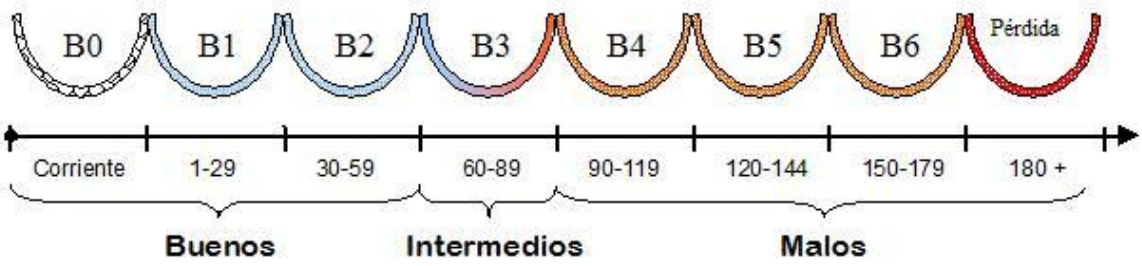


Figura 1: Posibles estados o canastas en los que puede caer un cliente.

Ejemplo. Supongamos que un individuo se le otorga una tarjeta de crédito departamental el 2 de enero (ver Figura 2). Decide que su fecha límite de pago sea el día 18 de cada mes, con facturación los días 25. El 5 de enero realiza compras por un monto de $5000. Su primera factura se emitirá el 25 de enero

con un saldo de $5000. Y tendrá como fecha límite de pago el 18 de febrero, sin que se le haga algún cargo adicional.

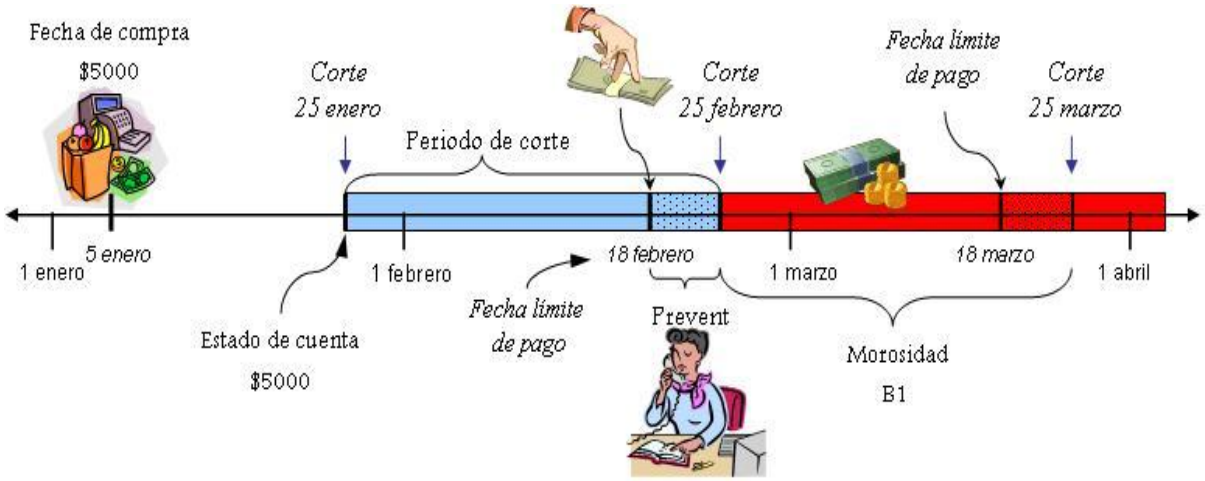


Figura 2: Estructura general en los eventos y tiempos asociados a una tarjeta de crédito.

En el tiempo preventivo del 18 al 25 de febrero se le aplicarán técnicas de cobranza. Los gastos de cobranza se facturaran hasta el 25 de marzo. Si el cliente no paga el mínimo, a partir del 26 de febrero se contará los días de moratoria. Del 26 de febrero al 25 de marzo el cliente se encuentra en la canasta B1, del 25 de marzo al 25 de abril, en la canasta B2 y así sucesivamente hasta B6. Es común considerar que después de B6 se cae en pérdida o *write off* esto es, cuando un cliente tiene más de 180 días de mora generalmente la institución crediticia los cataloga como perdida (NCL=Net Credit Losses) en sus estados financieros. Todo comportamiento de pago del cliente es enviado al buró de crédito, institución que guarda la información. La información se envía generalmente mensualmente y se registra hasta 24 meses.

III. Credit Scoring.

Hay diferentes maneras de obtener un *credit scoring*, dependiendo del país y leyes que lo rigen. Algunos bancos, por ejemplo, construyen sus propios *scorecards* (s*coring* estadístico), otros solo adaptan las construidas con información de otras instituciones (*scoring experto*). Aquí se presenta uno de estos métodos.

III. 1 ¿Qué es el credit scoring?

El *credit scoring* es una exitosa colección de técnicas estadísticas que se han utilizado para otorgar créditos en la industria del crédito [ver Simbaqueba (2004)]. Ha sido utilizado por más de 40 años permitiendo el crecimiento de consumidores de crédito, crecimiento que ha sido propiciado por el uso de la computadora y la aplicación la estadística para el manejo de grandes cantidades de datos. El *credit scoring* es una técnica bastante utilizada y rentable, dado que una pequeña mejora en el desempeño de la empresa puede significar un incremento considerable en las ganancias de los prestamistas debido al volumen de préstamos realizados.

III. 2 Tipos de score

La colección de técnicas que conforman el *scoring* tiene como propósito principal generar un puntaje de riesgo a las solicitudes de crédito o a cuentas ya existentes Lyn (2002). Hablando muy general del crédito al consumo, se puede describir el ciclo del riesgo, en el cual participan todos los clientes, en tres partes.

Figura 3: Ciclo de riesgo

Originación punto donde se otorga crédito por primera vez en la institución.

Administración. Consiste en premiar a los clientes que se “portan bien" (incrementos de límite de crédito) y castigar a los que se “portan mal"; Se busca detectar cuentas de alto riesgo y realizar acciones tempranas de corrección.

Recuperación. En esta parte del ciclo de riesgo se pretende recuperar a todos aquellos clientes que dejaron de pagar. Se aplican actividades de recaudación a clientes con un alto puntaje de score según la empresa y determinar los clientes no recuperables para hacer el traspaso a una empresa recaudadora y así recuperar parte del capital perdido. Dependiendo en qué parte del ciclo estemos trabajando se calcula uno de los siguientes puntajes de score:

Aquisition Score o Score de originación. En el departamento de Originación se utiliza éste puntaje para la aceptación o rechazo de las solicitudes de crédito. Su elaboración se basa en variables demográficas y de buró de crédito. Este puntaje estima la probabilidad de incumplimiento de pago de un posible cliente y de esta manera se decide si se acepta o rechaza como posible consumidor de crédito. Permite definir productos de crédito personalizados y realizar actividades de mercadeo para aumentar el número de clientes con características deseables y cumplir con las metas corporativas.

Behavior score o Score de comportamiento. Es utilizado en la etapa de administración del ciclo de riesgo. Predice la probabilidad de incumplimiento de

los clientes que ya son objeto de crédito en la institución. Se basa en variables de comportamiento de las cuentas dentro de la propia institución. Permite dar seguimiento al comportamiento de los clientes para que el departamento de cobranzas emplee técnicas para que un cliente siga siendo rentable.

Collection score. Puntaje que se calcula para estimar la probabilidad de recuperar a un cliente. Las variables utilizadas resultan de la combinación de variables de comportamiento y buró de crédito. Es posible determinar el valor preciso de la deuda antes de traspasarla a una empresa recaudadora.

En éste trabajo se estudia solamente el score de originación (Aquisition Score) aunque es importante mencionar que las técnicas usadas pueden ser emuladas en los demás tipos de score. Esta idea también puede ser utilizada en otros medios donde se necesite un tipo de clasificación binaria.

III. 3 Tipos de modelos

Modelo experto: Son modelos construidos con información de otras instituciones; está listo para usar. Se trata de modelos de crédito genéricos que se compran a consultores externos, Esto es común en instituciones sin historial de sus clientes.

Modelo estadístico: Son modelos que se construyen con información propia. Se pueden construir de manera específica para distintos segmentos de la población. Se adquiere conocimiento y experiencia sobre su población, y habilidad en el diseño e interpretación de los resultados. Se conserva la confidencialidad de la información.

III. 4 La scorecard

Una *scorecard* es una tabla que contiene los puntajes asignados a cada atributo de cada una de las variables usadas [ver Thomas et al. (2002)]. El

puntaje determina, por ejemplo, la probabilidad de pago de la deuda para un cliente cuando se le otorgue una tarjeta de crédito. Así que a mayores puntajes corresponden a una mayor probabilidad de pago. La compañía prestamista es la que define finalmente la probabilidad mínima de pago para determinar cuando un cliente es considerado bueno. Este puntaje llamado *cut off*, es indicado en principio por los analistas de *credit score* pero se verá influido por las decisiones gerenciales o se basará en las metas corporativas de la propia institución.

Ejemplo. Consideremos una *scorecard* simple con cinco variables o atributos: edad, estado civil, antigüedad en el empleo, sexo y nivel de estudios como se muestra en la tabla 4. Un individuo con 37 años de edad, soltero, con 5 años de *antigüedad* en su empleo, de sexo masculino y profesionista tendrá un puntaje (score) en base a esta tabla de $41 = 10 + 0 + 4 - 10 + 37$. Son varios los pasos a seguir y las metodologías a usar para obtener la *scorecard* [ver Barberena (2002)].

Característica	**Atributos**	**Score**
Edad	Menor a 24 años	– 40
	4 – 30 años	– 28
	31 – 40 años	10
	Mayor a 40 años	30
Estado Civil	Casado	12
	Soltero	0
	Otros	– 60
Antigüedad Empleo	0 – 1 años	– 5
	2 – 5 años	4
	6 – 10 años	10
	Mayor a 10 años	15
Sexo	Masculino	– 10
	Femenino	8
Nivel Estudios	Superior	– 15
	Medio	3
	Básica	20
	Profesionista	37

Tabla 4. *Scorecard*

A continuación listamos de forma general los pasos seguidos para encontrar la *scorecard*:

1. Conformar la base de datos. Se inicia con el vaciado en un archivo electrónico de la información contenida en las solicitudes de los clientes, el buró de crédito y el comportamiento de pago del cliente.
2. Depurar la base de datos. Consiste en excluir las variables con exceso de campos sin respuesta o respuestas múltiples.
3. Agrupar los datos contenidos en la base. Con la base depurada se forman intervalos de clase o grupos de clase (atributos) para cada característica (variable).
4. Determinar los clientes buenos y malos. Mediante métodos estadísticos se forman tres grupos: los clientes buenos, los clientes malos y los clientes indeterminados. Se forma una base solamente con los clientes buenos y malos.
5. Seleccionar las características con mayor valor de predicción. Mientras mas diferentes son las proporciones de buenos y malos en los diferentes atributos, la característica tiene mayor valor de predicción.
6. Determinar una función de clasificación. La función de clasificación se estima mediante la regresión logística de los datos en la base.
7. Elaborar la *scorecard*. Se calculan los valores de la *scorecard* mediante una translación y un cambio de escala de los estimadores de la regresión logística.
8. Medir la eficiencia de la *scorecard*. Este proceso se realiza con métodos estadísticos como el índice de Gini y la prueba de Kolmogorov-Smirnov.
9. Establecer el punto de corte. El punto de corte es el puntaje mínimo requerido para que un cliente sea aceptado.

III. 5 Determinación de clientes buenos y malos

Para los efectos de la *scorecard* los clientes se clasifican en buenos y malos. Los clientes buenos son los que pagan sus mensualidades a tiempo o permanecen en mora poco tiempo. Para determinar su estatus de bueno, malo o indeterminado se considera:

- El comportamiento de los clientes.
- El proceso de cobranza.
- Las metas corporativas de la institución de crédito.

Existen varias técnicas estadísticas para determinar el estatus de los clientes, aquí se utiliza una matriz de transición con la que se estima la probabilidad que un cliente caiga en moratoria. Los datos utilizados consisten en el seguimiento de clientes durante un periodo de tiempo que se le conoce como ventana de muestreo.

III. 5.1 Ventana de muestreo

La ventana de muestreo es un periodo de tiempo en el que se observa el comportamiento de las cuentas a partir de su apertura. Conforme avanza la edad de las cuentas, la tasa de moratoria va variando y se espera que en un momento determinado se estabilice, esto significa que a partir de ese momento ya podemos clasificar con una variación mínima a un cliente como bueno o malo.

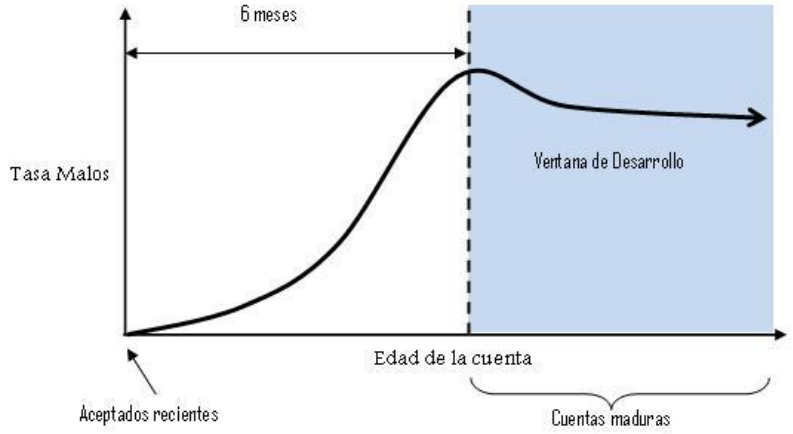


Figura 5. Ventana de muestreo. Periodo de observación del comportamiento de la tasa de mora en las cuentas a partir de su alta

Cuando la tasa de morosidad se ha estabilizado se dice que las cuentas llegan a la madurez de su comportamiento. La muestra debe representar a la población actual. La curva empieza a decaer cuando las cuentas se estabilizan y se quitan los clientes considerados indeterminados. Usualmente se considera un intervalo de tiempo de entre 12 y 18 meses anteriores a la fecha en que se hace el estudio.

III. 5.2 Proceso para determinar a los clientes buenos y a los clientes malos

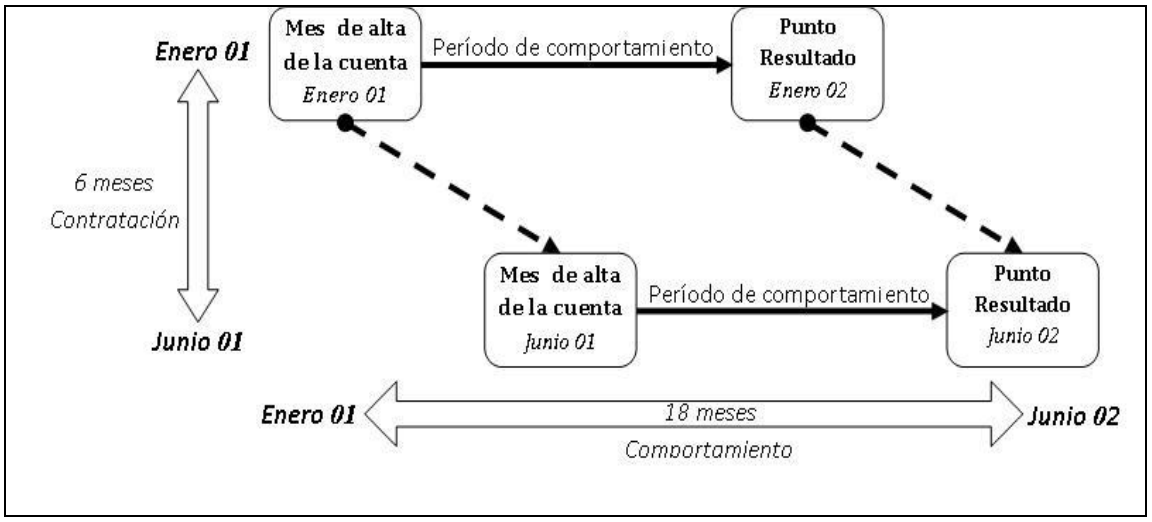


Figura 6. Cuentas consideradas en un periodo de contratación y un periodo de comportamiento

Para analizar la moratoria de los clientes se construye una matriz de transición, considerando el comportamiento de las cuentas de la institución en periodos de tiempo, generalmente de seis meses.

El periodo de observación o ventana de muestreo se contabiliza como primer mes, segundo mes, tercer mes, etc. a partir del momento de apertura, esto es, las cuentas se alinean en un punto inicial igual a cero (figura 6).

Los estados de la matriz de transición se definen en función del número de pagos vencidos durante los seis meses y su estado al final de esos seis meses (Figura 7). Los estados forman una partición de las cuentas; esto es, una cuenta debe pertenecer exclusivamente a un estado. Los estados se muestran en el tabla 8.

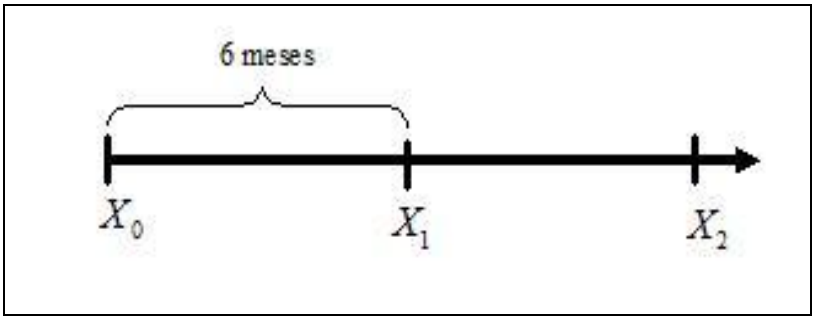


Figura 7. Estados de la matriz de transición en periodos de seis meses.

Al incorporar la información de las cuentas durante los seis meses se obtiene un refinamiento de estos estados, por ejemplo, el estado PV0 (*current*) se divide en cinco nuevos estados que se reportan en la tabla 9. El estado PV1 se divide en cuatro nuevos estados que se encuentran en la tabla 10. Nótese que en la tabla 10 no aparece el estado PV10 dado que no tiene sentido hablar de clientes con un pago vencido al final de los seis meses y cero

Estados	Descripción al final de los 6 meses
PV0	Al corriente
PV1	1 pago vencidos
PV2	2 pagos vencidos
PV3	3 pagos vencidos
PV4	4 o más pagos vencidos

Tabla 8: Posibles estados al final de 6 meses

pagos vencidos como máximo en esos seis meses. Esta misma idea se utiliza para dividir los demás estados. Así, el último estado tendría un único elemento, el PV44, para 4 o más pagos vencidos en el sexto mes y como máximo 4 o más pagos vencidos durante los seis meses. Así se obtiene una partición de 15 estados. El estado PV44 es un comportamiento indeseado por las instituciones de crédito, los clientes que están en este estado se pueden identificar como malos graves. El malo grave es el que cae en mora mayor a 90 días.

Estados	**Descripción al final de los 6 meses**
PV00	al corriente y máximo 0 pagos vencidos durante los 6 meses
PV01	al corriente y máximo 1 pagos vencidos durante los 6 meses
PV02	al corriente y máximo 2 pagos vencidos durante los 6 meses
PV03	al corriente y máximo 3 pagos vencidos durante los 6 meses
PV04	al corriente y 4 o más pagos vencidos durante los 6 meses

Tabla 9: Posibles estados con 0 pagos vencidos en los primeros 6 meses

Estados	**Descripción al final de los 6 meses**
PV11	1 pago vencido y máximo 1 pagos vencidos durante los 6 meses
PV12	1 pago vencido y máximo 2 pagos vencidos durante los 6 meses
PV13	1 pago vencido y máximo 3 pagos vencidos durante los 6 meses
PV14	1 pago vencido y 4 o más pagos vencidos durante los 6 meses

Cuadro 10: Posibles estados con 1 pago vencido en los primeros 6 meses

Ahora, buscamos encontrar que estados, llegan con probabilidad alta al estado no deseado, PV44, en los siguientes seis meses. Los elementos de la matriz de transición (roll rate) son las estimaciones de las probabilidades de que estando en el estado j se pase al estado i,

$$\hat{P}(X_2 = i \mid X_1 = j) = \frac{No.\ de\ cuentas\ en\ el\ estado\ j\ que\ pasan\ al\ estado\ i}{No.\ de\ cuentas\ en\ el\ estado\ j}$$

Dada la matriz de transición se determina que estados tienen una alta probabilidad de pasar al estado PV44 en el siguiente periodo. La idea es estimar la probabilidad de caer en PV44, esto es:

$$P(X_2 = PV44 \mid X_1 = i) \quad \text{donde} \quad i = PV00, PV01, \ldots, PV44\,.$$

Luego se seleccionan dos valores a y b tal que $0 < a < b < 1$,

- Si la cuenta *i*, satisface la relación $\hat{P}(X_2 = PV44 | X_1 = i) < a$, se considera que la cuenta es de un cliente bueno.
- Si la cuenta *i*, satisface la relación $\hat{P}(X_2 = PV44 | X_1 = i) > b$, se considera que la cuenta es de un cliente malo.
- Si la cuenta *i*, satisface la relación $a \leq \hat{P}(X_2 = PV44 | X_1 = i) \leq b$, se considera que la cuenta es de un cliente indeterminado.

III. 6 Obtención de la función de clasificación

Una vez que se tiene identificados a los clientes buenos y a los clientes malos se procede a estimar una función para clasificar a los nuevos clientes y así poder determinar si se les otorga o no un crédito. Debido a la naturaleza de las variables en la base, se elige a la regresión logística para estimar la función clasificadora [ver Thomas (1997)].

III. 6.1 Construcción de las variables explicativas

Los términos b_i y m_i corresponden al número de cuentas buenas y cuentas malas en la característica i. A su vez, los términos, b_{ij} y m_{ij} corresponden al número de cuentas buenas y cuentas malas en el atributo (valor) j de la característica i.

$$b_j = b_{i1} + b_{i2} + b_{i3} + \ldots + b_{in_i} \quad \text{y} \quad m_i = m_{i1} + m_{i2} + m_{i3} + \ldots + m_{in_i}$$

donde n_i es el número de atributos para la característica i. Con estos valores se calculan los siguientes términos:

- La distribución de buenos en la característica i

$$Pb_{ij} = \frac{No.\ de\ buenos\ en\ el\ atributo\ j\ de\ la\ característica\ i}{No.\ de\ buenos\ en\ la\ característica\ i} = \frac{b_{ij}}{b_i}$$

- La distribución de malos en la característica i

$$Pb_{ij} = \frac{No.\ de\ malos\ en\ el\ atributo\ j\ de\ la\ característica\ i}{No.\ de\ malos\ en\ la\ característica\ i} = \frac{m_{ij}}{m_i}$$

- La distribución de atributos de la característica i, $Pc_{ij} = \frac{b_{ij} + m_{ij}}{b_i + m_i}$:

- El *bad rate* que corresponde a la proporción de malos respecto del total de casos en el atributo *j* de la característica i, $M_{ij} = \frac{m_{ij}}{b_i + m_i}$

- El *good-:bad odds* que corresponde a la proporción de buenos respecto de los malos, término que también se conoce como la probabilidad de Buenos/Malos, $O_{ij} = \frac{b_{ij}}{m_{ij}}$:

Para ejemplificar esta etapa en el procedimiento presentamos una tabla con los cálculos correspondientes a la variable o característica "Tipo de Escuela".

Atributo	m_{ii}	b_{ii}	$m_{ii} + b_{ii}$	Pm_{ii}	Pb_{ii}	Pc_{ii}	M_{ii}	O_{ii}
Null	1	25	26	0.0002	0.0003	0.0003	0.0385	25:1
Pública	2498	47835	50333	0.4991	0.6685	0.6575	0.0496	19.1:1
Privada	273	3710	3983	0.0545	0.0519	0.0520	0.0685	13.6:1
Privada	353	4030	4383	0.0705	0.0563	0.0573	0.0805	11.4:1
Privada	784	7427	8211	0.1566	0.1038	0.1073	0.0955	9.5:1
Privada	1096	8525	9621	0.2190	0.1191	0.1257	0.1139	7.8:1
Total	**5005**	**71552**	**76557**	**1.0000**	**1.0000**	**1.0000**	**0.0654**	**14.3:1**

Cuadro 11: Atributos de la característica "Tipo de Escuela"

Estos datos corresponden a una muestra de 76,557 cuentas de solicitantes de una institución de crédito. En el primer renglón de la tabla 11 se encuentra el atributo "NULL" con 26 datos y sin presencia significativa. En el segundo renglón aparece la información del atributo "Pública" con un 65.7% del total de la muestra (ver columna Pc_{ij}), este mismo atributo tiene un 49.9% en la distribución de malos y un 66.9% en la distribución de buenos. Es interesante

observar que los clientes que estudian en una universidad privada son más proclives a caer en mora que los clientes que estudian en universidades públicas, ver columna M_{ij}, donde los datos crecen. Los clientes de las universidades “Privada Tipo IV" tienen mayor proporción de malos, esto es 343/4383 = 0.114 que corresponde al 11.4%. En consecuencia la columna O_{ij} muestra una probabilidad decreciente. Si dos o más atributos tienen semejante O_{ij} o M_{ij} su valor predictivo es semejante, por lo que todos ellos deben formar un único grupo o atributo. A esta agrupación se le conoce como “Clasificación dura" la cual consiste en juntar atributos con proporción de buenos y malos semejante.

III. 6.2 Pesos de Evidencia (WOE)

El poder de predicción en cada atributo o grupo de atributos se calcula con los Pesos de Evidencia (*Weigth of Evidence*) o WOE, cuya definición es:

$$WOE_{ij} = 100 \cdot \ln\left(\frac{Distribución\ \ de\ \ buenos\ \ en\ \ el\ \ atributo\ \ j\ \ de\ \ la\ \ característica\ \ i}{Distribución\ \ de\ \ malos\ \ en\ \ el\ \ atributo\ \ j\ \ de\ \ la\ \ característica\ \ i}\right)$$
$$= 100 \cdot \ln\left(\frac{Pb_{ij}}{Pm_{ij}}\right) = 100 \cdot \ln\left(\frac{b_{ij} \cdot m_i}{m_{ij} \cdot b_i}\right)$$

Obsérvese que

- El WOE_{ij} varía dependiendo de la forma en que se forman los atributos.
- Para que el WOE_{ij} esté definido, ninguna de las clases debe estar formada únicamente por buenos o por malos.
- Se sugiere no tener más de 8 clases y para que cada clase sea significativa debe contener al menos un 5% de los datos analizados.

Esto permite identificar datos *outliers* y clases raras, además de identificar comportamientos y adquirir conocimiento acerca del portafolio. Para

ejemplificar como se forman los grupos se considera los datos de la característica "Tipo de Universidad" (tabla 11). El atributo "NULL" se agrupa con el atributo "Pública" ya que sus *odds* están contiguos en el ordenamiento, y aunque no son semejantes "NULL" tiene únicamente 26 casos (menos del 5% del total) y no puede ser un grupo, los atributos Privada Tipo I y II los agrupamos en una clase dado que tienen *odss* cercanos y contiguos, así mismo con los atributos Privada Tipo III y IV se forma un nuevo grupo al final esta característica queda con tres grupos o atributos.

Con esta agrupación el *good:bad odds* de los diferentes atributos de una misma característica son significativamente diferentes. Para medir esta diferencia se utilizan algunos estadísticos como la χ^2 y el Valor de Información (IV).

Atributo	Malo	Bueno	Total	Pm_{ij}	Pb_{ij}	woe_{ij}
Pública, NULL	2499	47860	50359	0.4993	0.6689	29.240
Privada Tipo I y II	626	7740	8366	0.1251	0.1082	-14.518
Privada Tipo III y IV	1880	15952	17832	0.3756	0.2229	-52.167
Total	5005	71552	76557	1.0	1.0	

Cuadro 12: WOE de los atributos para "Tipo de Universidad"

Ejemplo. Consideremos los datos del cuadro 12 que muestra el WOE para cada atributo de la característica "Tipo de Universidad". Obsérvese que en las columnas Pm_{ij} y Pb_{ij} se muestran las distribuciones respectivas de malos y buenos. El WOE del atributo "Pública-NULL" es:

$$100 \cdot \ln\left(\frac{0.6689}{0.4993}\right) = 29.240$$

III. 6.3 Pruebas sobre la clasificación de atributos

Estadístico χ^2

Suponiendo que la proporción de cuentas buenas y cuentas malas en el atributo j de la característica i coincide con la proporción de cuentas buenas y cuentas malas en la característica i, entonces el estimador del valor esperado de las cuentas buenas y malas en el atributo j es igual a:

$$\hat{b}_{ij} = \frac{(b_{ij} + m_{ij})b_i}{b_i + m_i} \quad \text{y} \quad \hat{m}_{ij} = \frac{(b_{ij} + m_{ij})m_i}{b_i + m_i}$$

y el estadístico χ_c^2 para la característica i esta dado por

$$\chi_c^2 = \sum_{j=1}^{n_i} \left(\frac{(b_{ij} - \hat{b}_{ij})^2}{\hat{b}_{ij}} + \frac{(m_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \right) \sim \chi_{k-1}^2$$

Este valor será “pequeño" si $b_{ij} \approx \hat{b}_{ij}$ y $m_{ij} \approx \hat{m}_{ij}$, y será “grande" en caso contrario. De esta manera, χ_c^2 se usa para determinar cuan diferentes son los *odds* en cada clase. Entre más grande resulte el estadístico χ_c^2 refleja mayores diferencias en los *odds*, entonces al comparar dos agrupaciones se prefiere la de mayor valor de χ_c^2, ya que de esta manera la probabilidad de equivocarse al rechazar que los atributos de la característica i predicen mal es pequeña $P(\chi^2 > \chi_c^2)$.

Ejemplo. Con los datos de la característica “Tipo de Universidad" de la tabla 11 se establecerán tres formas diferentes de formar los atributos. La primera clasificación (Clasificación A) corresponde a la del Cuadro 12, los valores para el cálculo del estadístico χ_c^2 están en el cuadro 13, y el calculo es

$$\chi_c^2 = \frac{(47860 - 47067)^2}{47067} + \frac{(2499 - 3292)^2}{3292} + \frac{(7740 - 7819)^2}{7819}$$

$$+\frac{(626-547)^2}{547}+\frac{(15952-16666)^2}{16666}+\frac{(1880-1166)^2}{1166}$$

$$=204.5+49.7+519.1=773.3$$

Atributo	m_{ii}	b_{ii}	$m_{ii}+b_{ii}$	$\hat{b}_{ii}$	$\hat{m}_{ii}$
Pública, NULL	2499	47860	50359	47067	3292
Privada Tipo I y II	626	7740	8366	7819	547
Privada Tipo III y IV	1880	15952	17832	16666	1166
Total	5005	71552	76557	71552	5005

Cuadro 13: Datos para obtener χ_c^2 en la clasificación A de los atributos de La característica "Tipo de Universidad"

La segunda forma de clasificar (Clasificación B) es "NULL" con "Pública" (47860 buenos y 2499 malos), Privada tipo I, II y III (15167 buenos y 1410 malos), y Privada Tipo IV (8525 buenos y 1096 malos) se obtiene para χ_c^2.

$$\chi_c^2=\frac{(47860-47067)^2}{47067}+\frac{(2499-3292)^2}{3292}+\frac{(15167-15493)^2}{15493}$$

$$+\frac{(1410-1084)^2}{1084}+\frac{(8525-8992)^2}{8992}+\frac{(1096-629)^2}{629}=680.6$$

Los cálculos anteriores indican que la clasificación A es mejor que la B.

Valor de Información (IV)

El Valor de Información (IV) es una medida de entropía que aparece en la teoría de información [ver Siddiqi (2006)] y es una medida del poder de predicción global de la característica, y se define como

$$IV=\sum_j\left(\frac{b_{ij}}{b_i}-\frac{m_{ij}}{m_i}\right)\ln\left(\frac{b_{ij}m_i}{m_{ij}b_i}\right)$$

Los valores que puede tomar el estadístico IV son no negativos, y es cero cuando $\frac{b_{ij}}{b_i} = \frac{m_{ij}}{m_i}$ lo que equivale, directamente de la definición, que $b_{ij} = \hat{b}_{ij}$ y $m_{ij} = \hat{m}_{ij}$.

Siddiqi (2006) considera que una característica con un IV

- menor a 0.02 es tiene nulo valor predictivo
- entre 0.02 y 0.1 es de predicción débil
- entre 0.1 y 0.3 es de predicción media
- más de 0.3 es de predicción fuerte. (IV mayor a 0.5 sobre predice)

Siddiqi (2006) aconseja que las características con IV < 2% deben excluirse del modelo.

Ejemplo. Con los datos de la característica “Tipo de Universidad" se calcula el IV para la Clasificación A y la Clasificación B.

Clasificación A:

$$IV = (0.669 - 0.499)\ln\left(\frac{0.669}{0.499}\right) + (0.108 - 0.125)\ln\left(\frac{0.108}{0.125}\right) + (0.222 - 0.376)\ln\left(\frac{0.222}{0.376}\right)$$
$$= 0.132$$

Clasificación B:

$$IV = (0.669 - 0.499)\ln\left(\frac{0.669}{0.499}\right) + (0.212 - 0.282)\ln\left(\frac{0.212}{0.282}\right) + (0.119 - 0.219)\ln\left(\frac{0.119}{0.219}\right)$$
$$= 0.130$$

Se obtiene una mayor diferencia en el primera clasificación por lo que nuevamente, los datos indican que es la mejor opción para agrupar las clases.

III. 7 El modelo logístico

El modelo de regresión logística esta dado por

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Donde, β_0, β_1, β_2, …, β_{k0} son parámetros desconocidos y $x_1, x_2, \ldots, x_k$ son variables explicativas cuyos valores numéricos son los WOEs de la característica i, esto es:

$$x_i = woe_{i1}, woe_{i2}, \ldots, woe_{in_{ii}}$$

III. 8 Construcción de la Scorecard

Los puntajes del score son resultado de un re-escalamiento y una traslación del modelo logístico, dado por la ecuación

$$Score = Offset + Factor \cdot \ln(odds) = Offset + Factor \cdot \left(\hat{\beta}_0 + \sum \hat{\beta}_i woe_{ij}\right)$$

donde Offset es un término de traslación (o compensación) y Factor es un término de re-escalamiento. Offset y Factor deben satisfacer condiciones impuestas por la empresa de crédito. Este procedimiento permite la estandarización del score para que diferentes *scorecard* puedan ser comparadas. Se acostumbra a calibrar la *scorecard* de tal manera que cada cierto incremento en el puntaje P_0, se obtenga el doble de la relación *good/bad*.

Para ello se resuelve el sistema de ecuaciones

$$\begin{aligned} Score &= Offset + Factor \cdot \ln(odds) \\ Score + P_0 &= Offset + Factor \cdot \ln(2 \cdot odds) \end{aligned}$$

Cuya solución es: $Factor = \frac{P_0}{\ln(2)}$ y $Offset = Score - Factor \cdot \ln(Odds)$

Por ejemplo considere que un *odds* de 1:1 equivale a 600 puntos en la *scorecard* y que los *odds* se duplican cada $P_0 = 80$ puntos en la *scorecard*; es decir, 680 puntos equivalen a un *odds* de 2:1 y 760 puntos equivalen a 4:1, entonces.

$$Factor = \frac{80}{\ln(2)} = 115.4156 \quad y \quad Offset = 600 - 115.4156 \cdot \ln(1) = 600$$

Con esto se obtiene la función de score dada por

$Score = 600 + 115.416 \cdot \ln(odds)$

La relación del modelo de regresión logística con los WOE está dada por

$$Score = Offset + Factor \cdot \ln(odds) = Offset + Factor \cdot \hat{\beta}_0 + Factor \cdot \left(\sum \hat{\beta}_i woe_{ij} \right)$$

Los puntajes en la *scorecard* se descomponen en un puntaje inicial y un puntaje asociado al atributo j de la característica i, respectivamente

$$Offset + Factor \cdot \hat{\beta}_0 \quad \text{y} \quad Factor \cdot \hat{\beta}_i woe_{ij}$$

Es claro que los puntajes de una *scorecard* dependen de los parámetros de traslación y re-escalamiento que se utilicen

III. 9 Índice de Gini

En 1960 se pensó medir la desigualdad en los índices de salud con la curva de Lorenz, el índice de Gini se deriva de ésta [ver Medina (2001)]. El índice de Gini es uno de los instrumentos más utilizados para medir la desigualdad entre dos poblaciones. En este caso se utiliza para medir la desigualdad de las

poblaciones de buenos y malos clientes. Sea F(x) y G(x) las funciones de distribución de los buenos y malos clientes, con x el puntaje de la *scorecard*, la curva de Lorenz es el subconjunto

$$L(F,G) = \{(u,v) | u = F(x), v = G(x) \quad con \quad x \in \Re\}$$

Cuando la *scorecard* discrimina bien a los clientes buenos de los malos, el puntaje de score para buenos es mayor que el puntaje score para malos, entonces la curva de Lorenz de F y G es cóncava hacia arriba como en la figura 14. Cuando F(x) = G(x) la curva de Lorenz L(F;G) describe la recta u = v con $u \in (0,1)$. Por lo tanto mientras L se separe más de la recta v = u, mayor será la diferencia entre F(x) y G(x). Por esta razón, el área A que se encuentra entre la identidad y la curva de Lorenz es una medida de desigualdad entre las distribuciones F y G. El índice de Gini resulta de la razón entre el área A y el área del triangulo delimitado por la identidad, el eje horizontal u y la recta u = 1. Mientras mayor sea su valor, la *scorecard* discrimina mejor.

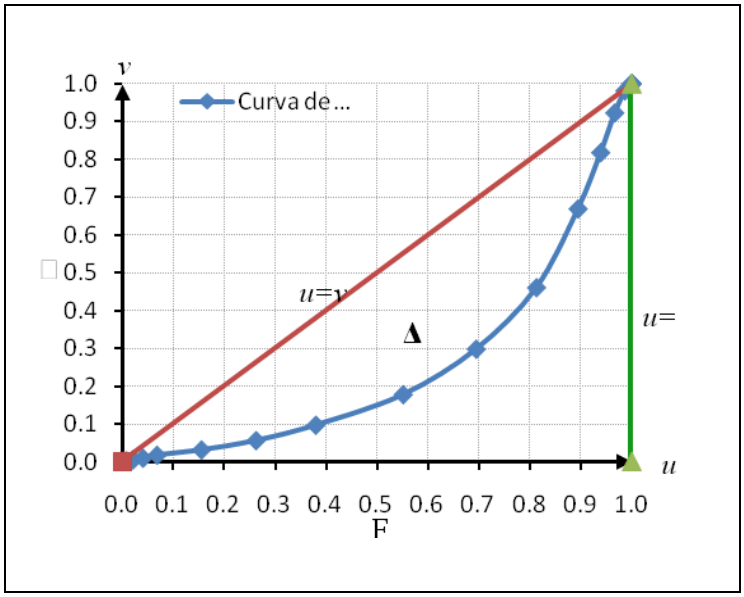


Figura 14. La curva de Lorenz

III. 10 Determinación del punto de corte o Cut Off

Cuando se tiene los datos de un nuevo solicitante, se calcula su score y con el resultado se decide si se le otorga o no el crédito. Se elije un punto "a" llamado punto de corte o Cut Off tal que si $Score > a$ se otorga el crédito, en caso contrario si $Score \leq a$ se rechaza la solicitud, es importante determinar el valor que optimiza la decisión. En esta sección presentamos dos maneras de estimar el punto de corte. La primera forma es suponer que si la probabilidad de ser buen cliente está por arriba de un medio, se aprueba el crédito y si está por debajo, se rechaza; esto es, para aprobar un crédito se utiliza la desigualdad

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1^T x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1^T x}} = \hat{p} > \frac{1}{2}$$

de aquí se sigue que

$$e^{\hat{\beta}_0+\hat{\beta}_1^T x} > \frac{1}{2}\left(1+e^{\hat{\beta}_0+\hat{\beta}_1^T x}\right) \Rightarrow \frac{1}{2}e^{\hat{\beta}_0+\hat{\beta}_1^T x} > \frac{1}{2} \Rightarrow e^{\hat{\beta}_0+\hat{\beta}_1^T x} > 1 \Rightarrow \hat{\beta}_0 + \hat{\beta}_1^T x > 0 \Rightarrow a = Offset$$

La segunda forma de obtener el punto de corte es calcular el score para todas las cuentas de la base, estos puntajes se ordenan y el valor de a satisface la ecuación

$$\frac{\#\{Score \mid Score > a\}}{\#\{Score\}} = \text{ Una proporción}$$

con el valor de la proporción seleccionado por la empresa.

IV. Conclusiones

Existen diversas compañías a nivel internacional que ofrecen sus servicios en el análisis de riesgo, sus servicios son caros y los productos entregados son cajas negras para las empresas adquirientes; por esta razón es importante desarrollar en nuestro país formas novedosas de obtener el *credit scoring*.

Generar una *scorecard* es una combinación de ciencia y arte [ver Simbaqueba (2004)]; su elaboración es un trabajo que debe ser realizado en equipo por todos los implicados en el proceso de su generación y su implementación. Los procedimientos presentados en este artículo fueron aplicados a datos reales de una empresa crediticia y la evaluación sobre los resultados fue aceptable, la falta de espacio en este artículo impide presentar los resultados obtenidos. Consideramos que algunos de los procedimientos del *credit scoring* pueden ser modificados y las modificaciones deberían ser probadas para determinar si ofrecen mejoras en las estimaciones. Por ejemplo, usando toda la teoría que respalda el estudio de los procesos de Markov se puede buscar una potencia de la matriz de transición que estabilice las probabilidades de transición y con estos valores definir a los buenos y malos clientes. De esta manera queda pendiente el desarrollar nuevas formas de conformar cada uno de los pasos que conforman el proceso del *credit scoring*.

V. Bibliografía

1. Barberena, Manuel y Barberena, Viterboo. (Febrero de 2002) La minería de Datos en la Industria Financiera: Un nuevo enfoque de Investigación de Mercados. AMAI, No. 31, Año 9, México.
2. González Roberto, (11 de diciembre de 2008) Cada día caen en cartera vencida unos 3 mil 305 préstamos al consumo, La jornada, Finanzas, http://www.jornada.unam.mx/2008/12/11/index.php?section-=economia&article=029n1eco
3. Lino Arturo. (27 de enero de 2009) Baja 2.68% el crédito al consumo". El Sol de México, D.F. http://www.oem.com.mx/elsoldemexico/notas/n1022683.htm
4. Medina Fernando, (marzo 2001) Consideraciones sobre el índice de Gini para medir la concentración del ingreso. Estudios estadísticos y prospectivos, serie 9. Publicación de las Naciones Unidas. Santiago de Chile.
5. Moreno, Tania M. (28 de marzo de 2008) “Crédito al consumo conquista a México". cnnexpansion, México. http://www.cnnexpansion.com/midinero/2008/03/28/credito-al-consumo-2018conquista2019-a-mexico-1
6. Ocejo Rojo, Iñigo (9 de febrero 2009) Usuarios de tarjetas de crédito, sólo 8% paga puntual. El Economista, Diario de circulación nacional, México. http://www.astromante.com/nota.php?NOT_ID=7401

7. Siddiqi, Naeem. (2006) Credit Risk Scorecards: developing and implementing intelligent credit scoring. John Wiley & Sons, New Jersey, 2006.
8. Simbaqueba Lilian. (2004) ¿Que es el scoring? Una visión práctica de la gestión del riesgo de crédito. Instituto del Riesgo Financiero, Bogotá.
9. Thomas Lyn, (2002) Edelman David and Crook Jonathan. Credit Scoring and its applications. SIAM, Philadelphia.
10. Thomas P. Ryan. (1997) Modern Regression Methods. John Wiley and Sons, Wiley Series in Probability and Statistics, New York.
11. Zúñiga Antonio y Rodríguez Israel. (27 de enero de 2009) Creció más de 50% la cartera vencida del crédito al consumo. La Jornada, http://www.jornada.unam.mx/2009/01/27/index.php?section=economia-&article=020n1eco

Construcción de una tabla de mortalidad con un enfoque Bayesiano

Elizabeth Aquino Pérez

Subgerencia de Pensiones. Banco de México.

eaquino@banxico.org.mx

Teléfono: (+52 55) 52372000 ext. 6088

Resumen

El objetivo de este artículo es difundir un método estadístico Bayesiano para la construcción de una tabla de mortalidad que establezca la precisión de las estimaciones y márgenes de seguridad en términos probabilísticos que garanticen que el pago de las obligaciones de la institución esté debidamente financiado con un alto grado de confiabilidad. Se presentará el trabajo de Mendoza et al. (1999) desarrollado para la Comisión Nacional de Seguros y Fianzas cuya metodología consiste en un análisis Bayesiano de modelos de regresión lineal con transformación logística. Asimismo se propone el modelo de regresión logística. Finalmente, se presenta el ajuste de la tabla de mortalidad utilizando ambos modelos con datos reales de una institución.

Palabras clave: tabla de mortalidad, tabla de mortalidad CNSF, estadística Bayesiana.

I. Introducción

La predicción de tasas de mortalidad es una herramienta que se aplica en una amplia variedad de campos académicos, investigaciones médicas, farmacéuticas, seguridad social y planes de retiro. En este último caso, a mayor aproximación en la predicción de las tasas de mortalidad, menor será la desviación en el riesgo de los recursos destinados al pago de rentas vitalicias.

En los últimos años, el desarrollo de modelos estadísticos ha permitido a los investigadores incorporar nuevas variables a los modelos existentes, lo que ha generado resultados con mayor precisión.

En vista de lo anterior, el objetivo del presente documento consiste en difundir el ajuste a los datos de mortalidad utilizando un enfoque Bayesiano, en particular se presentará el trabajo desarrollado por Mendoza et al. (1999) para la Comisión Nacional de Seguros y Fianzas, cuya metodología estadística consiste en un análisis Bayesiano de modelos de regresión lineal con transformación logística. Asimismo, en el presente documento se propone el modelo de regresión logística. Finalmente, se presentará la aplicación práctica de los dos modelos mencionados anteriormente y se compararán los resultados obtenidos.

II. Modelos Bayesianos

La estimación de las tasas de mortalidad mediante modelos paramétricos ha sido foco de atención de actuarios, demógrafos y estadísticos, ya que dicha

estimación no es sólo una graduación con aplicaciones actuariales, sino técnicas de regresión que suelen utilizarse en estadística y que ahora aplicamos con un objetivo determinado, la descripción de los aspectos más relevantes de la mortalidad humana.

En este contexto, los métodos Bayesianos representan una herramienta bastante útil ya que permiten incorporar información a priori sobre la mortalidad y complementarla con nueva información proveniente de los datos.

II. 1 Modelo de regresión lineal con transformación logística

El análisis de regresión trata del estudio de la relación entre dos o más variables observables "y" y "x" de tal forma que una de ellas pueda ser descrita y predicha a partir de la otra(s). En general, estamos interesados en la distribución condicional de y dado x, parametrizada por $f(y|x)$, en donde se tienen acceso a n observaciones (x_i, y_i) condicionalmente independientes.

Un modelo de regresión normal queda definido como sigue:

Sean Y_i, $x_i' = (1, x_{i1}, x_{i2}, \dots, x_{i\,p-1})$, $i = 1,2,\dots,n$ un conjunto de variables aleatorias tal que

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i\,p-1} + \varepsilon_i$$

i.e.,

$$Y_i = x_i'\beta + \varepsilon_i \ ,$$

donde,

$\beta' = (\beta_0, \beta_1, \dots, \beta_{p-1})$ es un vector de p parámetros, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ son v.a. i.i.d. tal que $\varepsilon_i \sim N(0,\tau)$ para $i = 1,2,..,n$, con $Var(\varepsilon_i) = 1/\tau = \sigma^2$

Lo anterior implica que $E(Y_i|\beta,\sigma^2,x_i) = x_i'\beta$ y que $Var(Y_i|\beta,\sigma^2,x_i) = 1/\tau = \sigma^2$

Para hacer referencia a la variabilidad de los datos alrededor de la media se puede utilizar la varianza o la precisión, siendo esta última el inverso multiplicativo de la varianza.

Alternativamente, el modelo se puede escribir como:

$$Y_i|\beta,\tau,x_i \sim N(x_i'\beta,\tau), i = 1,\dots,n$$

Ahora bien, el procedimiento propuesto por Mendoza et al. (1999) es el que se describe a continuación.

1. **Transformación logística de los datos**

   La distribución Normal tiene como soporte toda recta real y dado que las tasas de mortalidad se encuentran en el intervalo (0,1) y por otro lado, en efecto las tasas guardan una relación esperada creciente respecto a la edad pero no necesariamente exhiben una tendencia lineal, es necesario aplicar una transformación a las tasas para poder tener un ajuste razonable. Utilizaremos una transformación logística de las tasas de mortalidad observadas. Por lo tanto, la variable respuesta resultante es:

   $$Y = ln\left(\frac{q_x}{1-q_x}\right)$$

   Donde q_x es la probabilidad de fallecer entre las edades exactas x y $x+1$.

2. **Análisis Bayesiano de un modelo de regresión lineal simple con los datos transformados**

   Con los datos transformados, se ajusta un modelo de regresión lineal Bayesiano (Congdon, P. (2001)), lo que significa que se determina la distribución predictiva conjunta para el vector y dado el vector de edades x.

   El análisis del modelo desde un punto de vista Bayesiano requiere de especificar una distribución inicial sobre los parámetros (β,τ).

Consideremos la función de verosimilitud,

$$f(y|\beta,\tau,X) \propto \tau^{n/2} exp\left\{-\frac{\tau}{2}(y-X\beta)'(y-X\beta)\right\}$$

Con la intención de fundamentar las inferencias en la información provista por los datos, se utilizan distribuciones de referencia o mínimo informativas. Por lo tanto, tenemos las siguientes distribuciones:

Inicial:

$$\pi(\beta,\tau) \propto \tau^{-1}$$

Final:

$$\pi(\beta,\tau|y,X) = N(\beta|\hat{\beta},(X'X)\tau)Ga(\tau|(n-p)/2,(n-p)s^2/2),$$

Marginalmente:

$$\pi(\beta|y,X) = St(\beta|\hat{\beta},(X'X)(s^2)^{-1},n-p), \text{ y}$$

$$\pi(\tau|y,X) = Ga(\tau|(n-p)/2,(n-p)s^2/2),$$

donde,

$$\hat{\beta} = (X'X)^{-1}X'y,$$

$$s^2 = \frac{1}{n-p}(y-X\hat{\beta})'(y-X\hat{\beta})$$

3. Distribución predictiva en el modelo

Después de que conseguimos la distribución posterior de los parámetros y que podemos obtener la información que dan los datos de la muestra aleatoria bajo los supuestos del modelo, en ocasiones se requiere obtener la información para saber cómo se comportará una observación futura. Para esto se tiene la función predictiva final, la cual en este análisis de referencia, dado un vector de covariables x_F queda determinada por:

$$\pi(y_F|y) = St\left(y_F|x_F'\hat{\beta},(1+x_F'(X'X)^{-1}x_F)^{-1}(s^2)^{-1},n-p\right)$$

Recordemos que la inferencia se realiza utilizando la distribución final de los parámetros y la predictiva final.

4. **Distribución predictiva para las tasas de mortalidad con la transformación inversa**

Finalmente, obtenemos la distribución predictiva para las tasas de mortalidad con la transformación inversa, es decir regresamos los datos a la escala “original”:

$$q_x = \frac{exp\ (Y)}{1 + exp\ (Y)}$$

En este punto, es conveniente mencionar que la transformación inversa que aplicamos para regresar a las tasas de mortalidad observadas no preserva el valor esperado, sin embargo, la mediana y, en general, cualquier cuantil, sí se preserva. Es decir, el cuantil a nivel α de la distribución predictiva para las tasas observadas se obtiene de manera directa al transformar el correspondiente cuantil de la distribución $\pi(y_F|y)$.

Este hecho es sumamente importante porque de aquí se desprende la ventaja de esta metodología con respecto a los modelos actuariales convencionales, en el sentido de que la curva definida por los cuantiles del mismo orden, digamos α de las distribuciones predictivas marginales de las tasas observadas tiene la propiedad de que, sea cual sea la edad de interés, la probabilidad de observar, para ese grupo de edad, una tasa de mortalidad superior al indicado por la curva es $1 - \alpha$.

Como podemos observar, esta metodología evalúa la incertidumbre involucrada en la estimación de las tasas de mortalidad y establece un criterio para sobrecargar las tablas que valoren, técnicamente, los efectos en términos de riesgo.

II.2 Modelo de regresión logística

El modelo de regresión lineal es una forma de describir la relación entre una variable respuesta "y" y un conjunto de variables explicativas $x' = (x_1, x_2, \dots, x_{p-1})$. Una forma más general de describir la distribución condicional de Y dado X, $f(y|\beta, x)$, es mediante la clase de modelos lineales generalizados; al respecto debemos mencionar que el modelo de regresión logística es un caso particular de esta clase de modelos.

La idea general de los modelos lineales generalizados consiste en modelar el valor esperado de Y, digamos $\mu(x) = E(Y|x)$, a través de una función paramétrica simple de las variables explicativas, digamos $\phi(\beta, x)$, es decir $\mu(x) = \phi(\beta, x)$.

Al considerar distintas distribuciones para la variable respuesta Y y distintas formas para la función $\phi(\cdot)$, este modelo produce una clase muy amplia de modelos de regresión generalizados.

Un modelo lineal generalizado es un modelo lineal para la media transformada de una variable que tiene una distribución en la familia exponencial natural. En particular, cuando la función de la familia exponencial que se elige es la distribución Binomial y la función liga es la transformación logística, entonces se habla de un modelo de regresión logística.

Es oportuno mencionar un supuesto básico en el contexto de tablas de mortalidad. Sea E_x el número de personas iniciales expuestos al riesgo, esto significa que E_x representa a las personas que entran en observación a edad exacta x y continúan hasta edad $x + 1$.

Supongamos que q_x representa la probabilidad de muerte dentro de un año y d_x la variable aleatoria que representa el número de muertes durante un año. Una conclusión inmediata, es que d_x tiene una distribución Binomial con parámetros E_x y probabilidad q_x, es decir:

$$d_x|E_x, q_x \sim Bin(E_x, q_x)$$

De ahí que el modelo de regresión logística sea adecuado para la construcción de una tabla de mortalidad. El procedimiento para llevar a cabo el ajuste es análogo al caso de la regresión lineal con transformación logística que se mencionó en la sección II.1. De igual forma, se pueden obtener los intervalos de pronóstico de interés a través de los cuantiles de la distribución predictiva.

III. Construcción de tablas de mortalidad

III.1 Descripción de los datos

En la notación que se usará, se omiten literales que se refieren al sexo de la población, ya que los conceptos se aplican de igual forma a hombres y mujeres. Se realizarán dos tablas porque la mortalidad es distinta en hombres y mujeres (las mujeres viven, generalmente, más tiempo que los hombres).

La información empleada en la construcción de la tabla de mortalidad es la que proviene de la experiencia empírica de cierta institución (cuyo nombre mantendremos en el anonimato por motivos de confidencialidad) resultante de los registros demográficos de la población objetivo.

Debido a que el grupo poblacional de trabajadores y pensionados de la institución es numéricamente bajo, y con el fin de tener mayor representatividad estadística, se manejan las series de datos del periodo 2000-2009 agrupadas para distintos intervalos temporales es decir, como si personas y eventos de esos intervalos hubieran ocurrido simultáneamente.

Denotemos a las poblaciones como:

$T_x(t)$ Trabajadores de edad cumplida x al inicio del año t.

$J_x(t)$ Jubilados (pensionados por antigüedad) de edad cumplida x al inicio del año t.

$d_x(t)$ Defunciones, de edad cumplida x al inicio del año t.

El sobre índice derecho t, j en las defunciones, se refiere a los decesos de trabajadores y a los de pensionados, respectivamente.

Si bien la referencia al tiempo en los eventos debiera ser $(t, t+1)$, se abrevia con (t) para simplificar la nomenclatura y dado que en todo momento se trabaja con duraciones anuales.

La población total expuesta al riesgo es, de acuerdo a la nomenclatura anterior:

$$E_x(t) = T_x(t) + J_x(t).$$

Las defunciones:

$$d_x(t) = d_x^t(t) + d_x^j(t).$$

Los agregados para el cociente:

$$d_x = \sum_{t=2000}^{2009} d_x(t) \quad y \quad E_x = \sum_{t=2000}^{2009} E_x(t).$$

La tasa observada de mortalidad q_x se obtiene a partir de la relación:

$$q_x = \frac{d_x}{E_x}$$

III.2 Ajuste con los métodos Bayesianos

Para la graduación de las tablas de mortalidad, la estimación se realizó con base en el enfoque Bayesiano de los modelos de regresión descritos en la sección II. Estos modelos se ajustaron con el software WinBUGS, el cual está creado para resolver problemas de inferencia estadística Bayesiana haciendo uso de métodos MCMC (Monte Carlo vía Cadenas de Markov). Los códigos utilizados se encuentran en el apéndice.

El modelo de regresión lineal con transformación logística fue el siguiente:

$$\eta_i = logit\left(\frac{d_i}{E_i}\right)$$

$$\eta_i|\mu_i,\tau \sim N(\mu_i,\tau)$$

$$\mu_i = \beta_1 + \beta_2 x_i$$

$$con\ \beta_1 \sim N(0,0.001), \beta_2 \sim N(0,0.001)\ y\ \tau \sim Ga(0.001,0.001)$$

Mientras que el modelo de regresión logística fue:

$$d_i|E_i,q_i \sim Bin(E_i,q_i)$$

$$logit(q_i) = log\left(\frac{q_i}{1-q_i}\right) = \beta_1 + \beta_2 x_i$$

$$con\ \beta_1 \sim N(0,0.001) y\ \beta_2 \sim N(0,0.001)\ ,$$

donde para ambos modelos:

$i = 1, \ldots, 17$ es el grupo de edad.

d_i es el número de muertes en el grupo de edad i.

E_i es el número de expuestos al riesgo en el grupo de edad i.

q_i es la probabilidad de muerte del grupo de edad i.

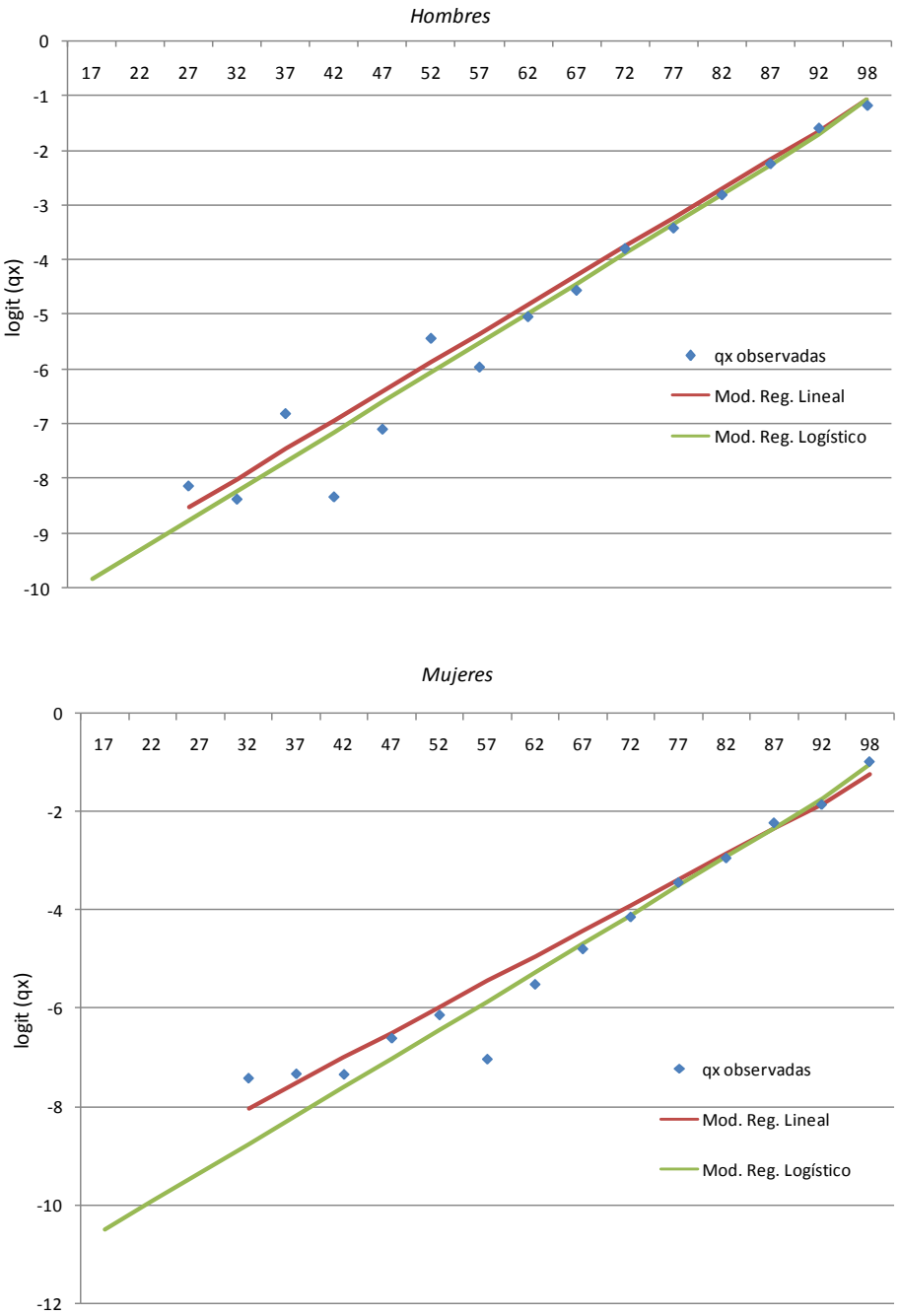
x es la edad y es la única variable independiente del modelo.

Derivado del análisis de los datos de la población, se observó que existen grupos de edad en los cuales no hubo defunciones. Este hecho causará inconveniente en el modelo de regresión lineal con transformación logística (no así en el modelo de regresión logística), ya que si $q_i = 0$, entonces al calcular el logit de q_i, éste queda indeterminado.

Por este motivo, es necesario realizar un ajuste a los datos. Se tienen dos opciones. La primera consiste en eliminar aquellos grupos de edad en los cuales no hubo defunciones y con el modelo de regresión lineal con transformación logística estimar las tasas de mortalidad de los grupos en los cuales $d_i \neq 0$ y por lo tanto $q_i > 0$. Ahora bien, la otra opción que se propone es modificar el número de defunciones d_i, es decir cambiar las $d_i = 0$ por un número muy pequeño, por ejemplo sería conveniente hacer $d_i = 0.1$ ó bien, sustituir por $0.01, 0.001, 0.0001$ para fines comparativos, sin embargo debemos estar conscientes de que se obtendrán distintos ajustes de acuerdo al valor que seleccionemos. En el apartado III.3 presentaremos las medidas de comparación de modelos con ambas vertientes y para fines de las gráficas que se mostrarán a continuación, el ajuste de las tablas de mortalidad para el caso del modelo de regresión lineal con transformación logística se realizó eliminando aquellos grupos de edad donde el número de defunciones es igual a cero, mientras que en el caso del modelo de regresión logística se utilizaron todas las observaciones.

En la Gráfica 3.1 se muestra la recta ajustada que describe la relación de la edad con el valor medio (esperado) del logit de la tasa de mortalidad observada.

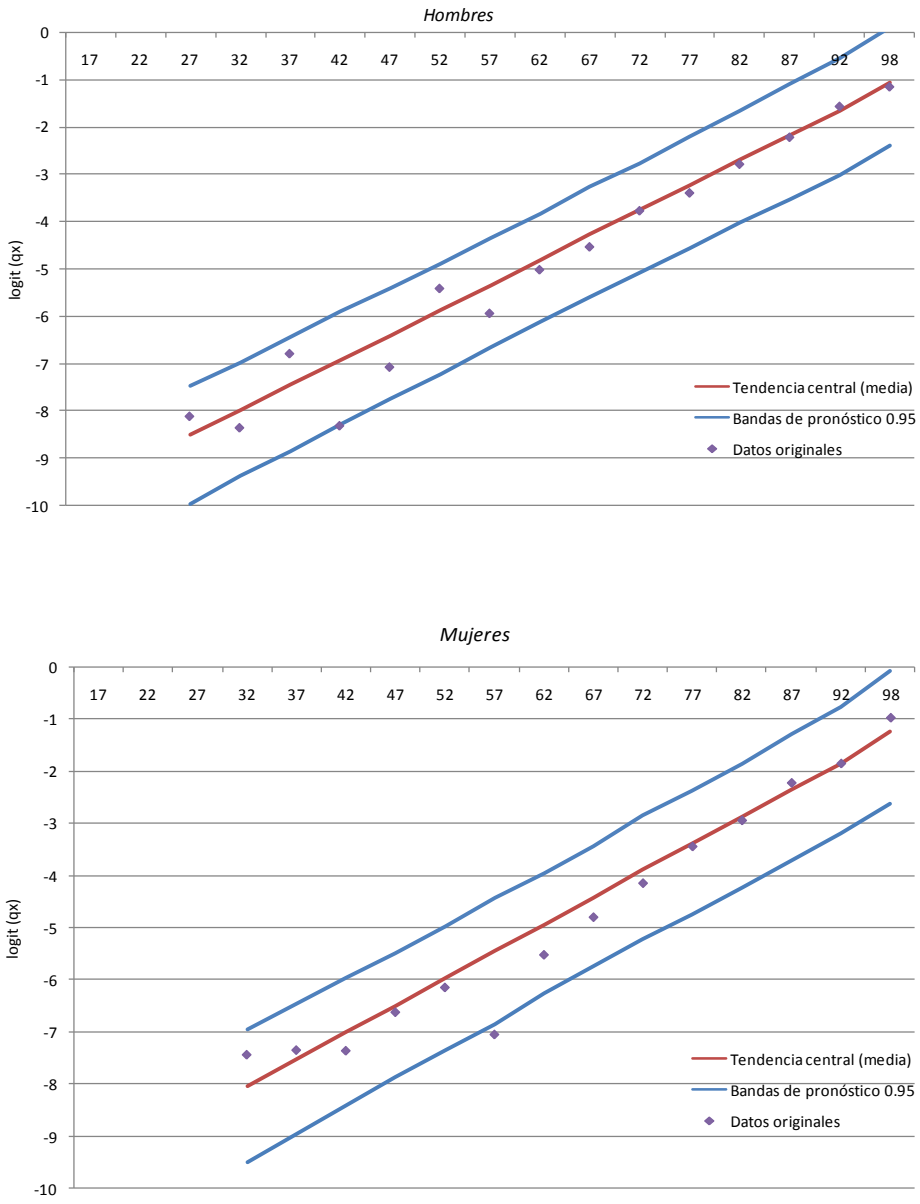
Gráfica 3.1. Modelos de regresión con un enfoque bayesiano para datos transformados por edad y género, 2000-2009



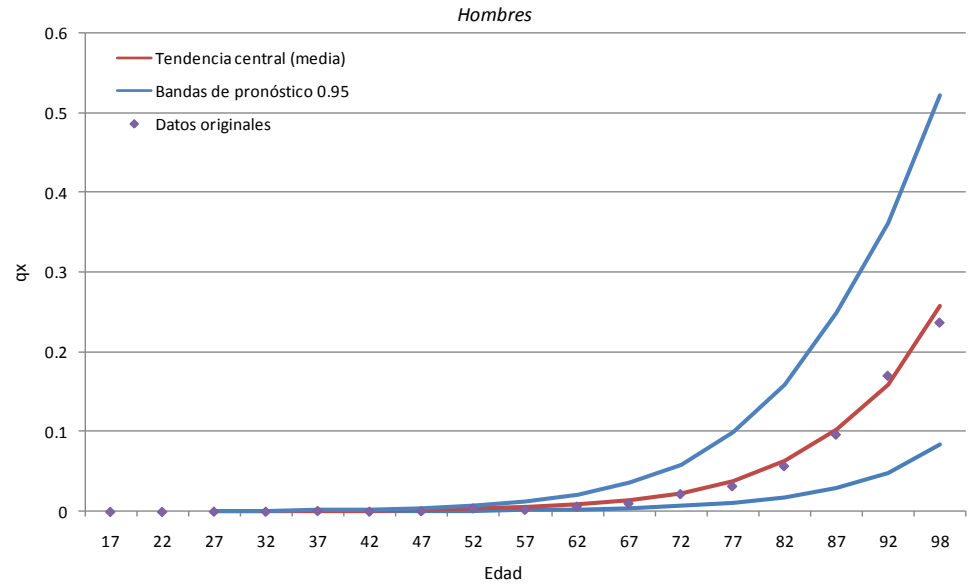
Dentro de las ventajas de los modelos Bayesianos que hemos mencionado está la de recargar las tablas haciendo uso de los cuantiles. La banda al nivel $1-\alpha$ se obtiene calculando, para cada edad, los cuantiles de orden $\alpha/2$ y

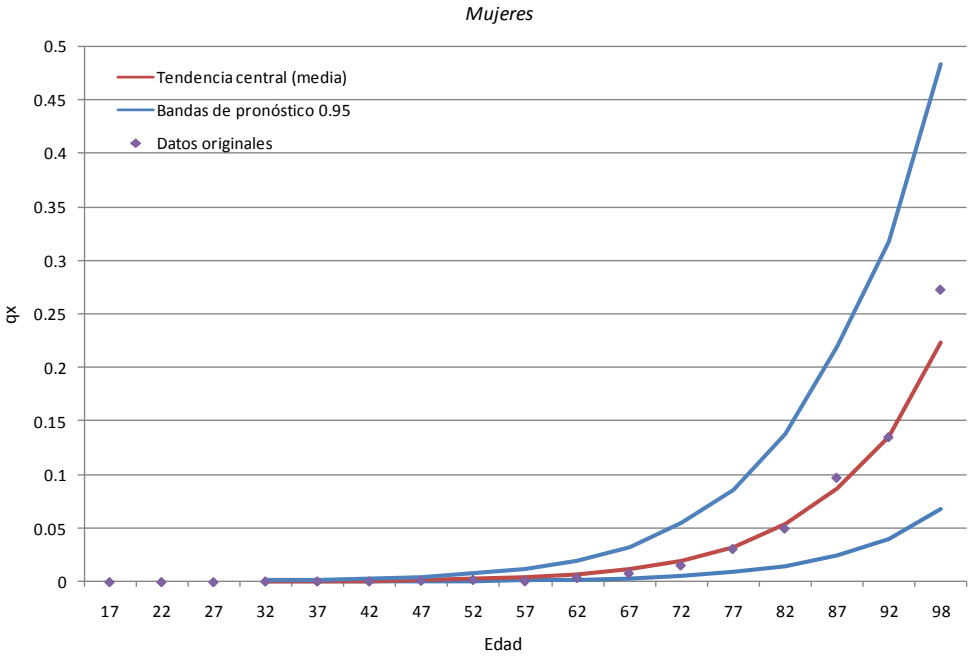
$1 - \alpha/2$. En la Gráfica 3.2 se muestran las bandas de credibilidad al 0.95 para los dos modelos de regresión.

Gráfica 3.2. Modelo bayesiano de regresión lineal con transformación logística para datos transformados por edad y género, 2000-2009

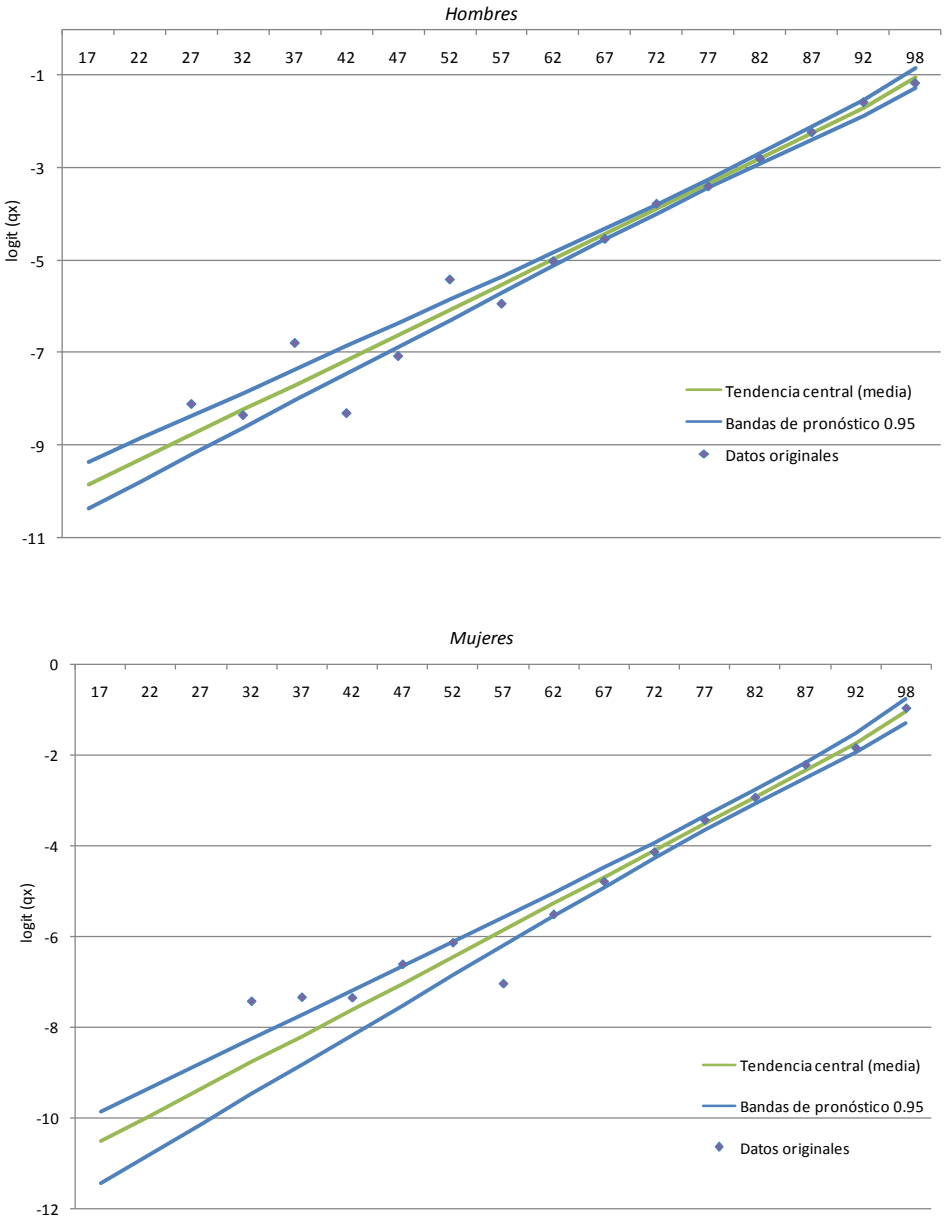


Gráfica 3.3. Modelo bayesiano de regresión lineal con transformación logística en términos de las tasas observadas por edad y género, 2000-2009

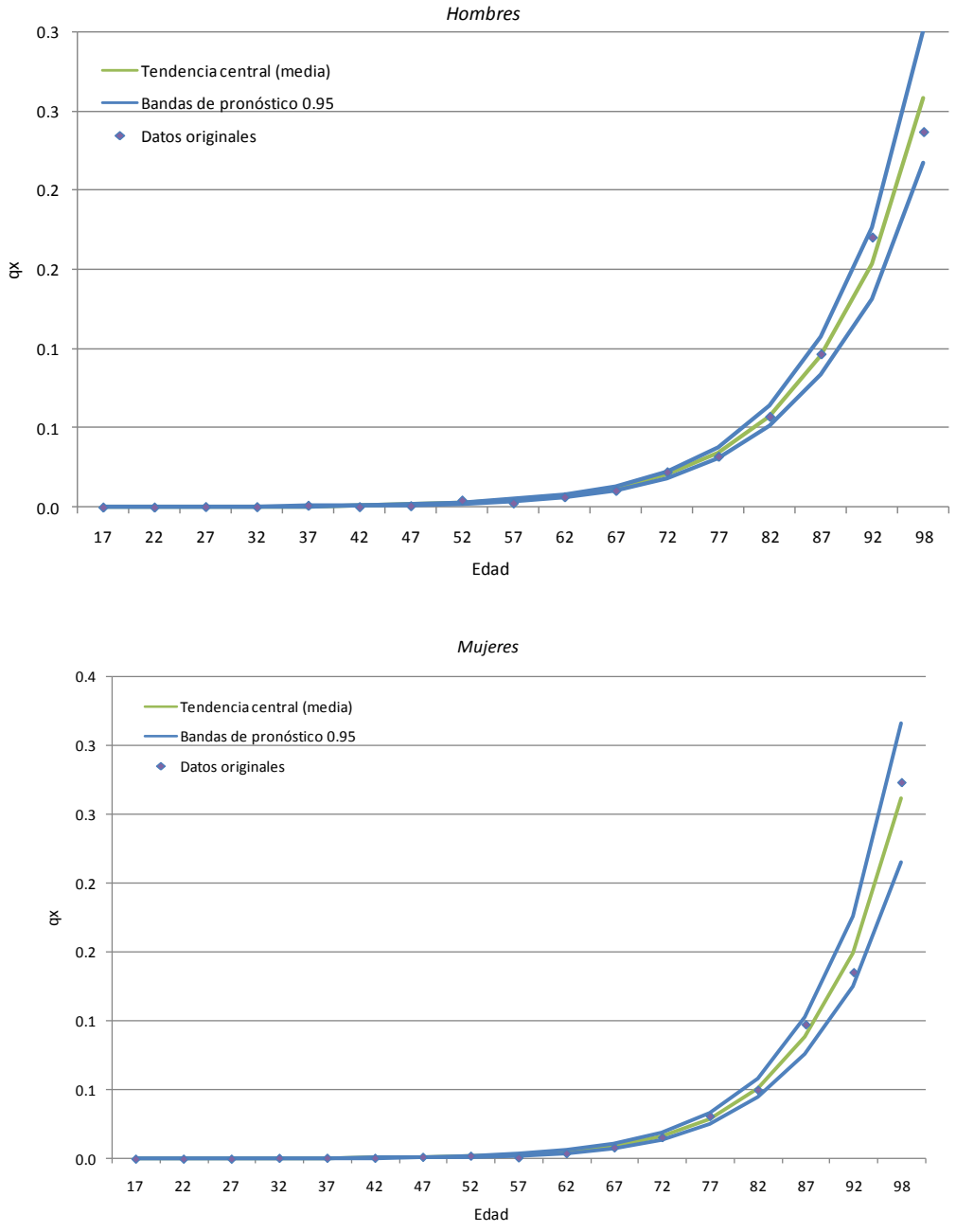




Gráfica 3.4. Modelo bayesiano de regresión logística para datos transformados por edad y género, 2000-2009



Gráfica 3.5. Modelo bayesiano de regresión logística en términos de las tasas observadas por edad y género, 2000-2009

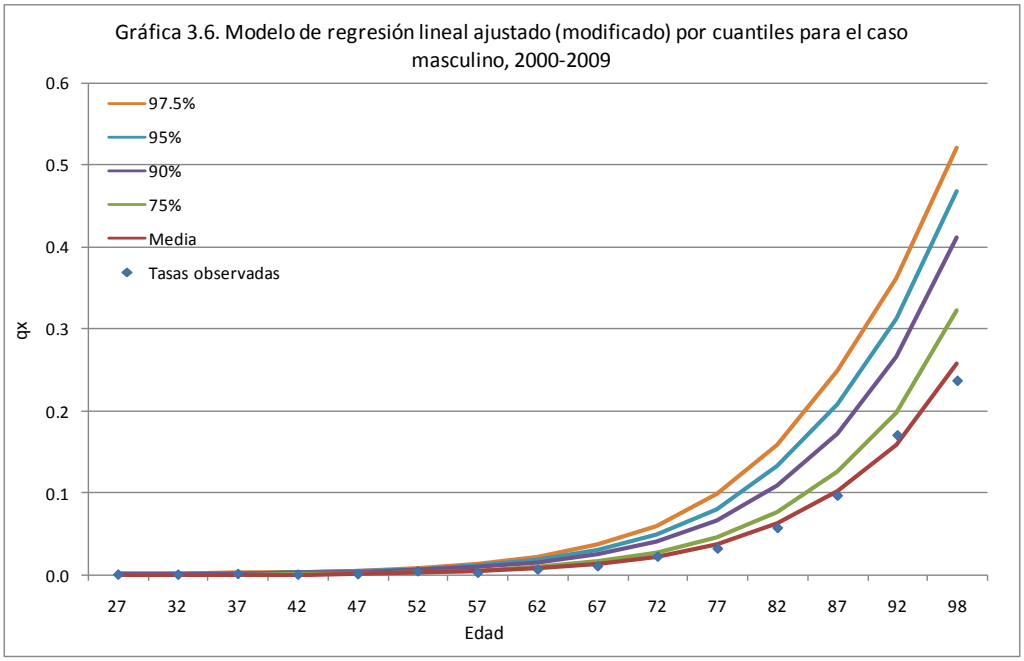


Claramente se observa que las bandas de credibilidad en el modelo de regresión lineal son más anchas que las que se obtienen con el modelo de regresión logística. Esto se debe a que el modelo de regresión lineal con transformación logística tiene la ventaja de tener un parámetro extra: la precisión (τ). Es importante destacar que, a diferencia del modelo de regresión logística, la varianza en el modelo lineal no depende de la edad.

Recordemos que si $d_i|E_i,q_i \sim Bin(E_i,q_i)$, la varianza está dada por $E_i q_i(1-q_i)$ donde i es el grupo de edad, por lo que al calcular los cuantiles con este modelo y compararlos con los obtenidos por el modelo de regresión lineal, es de esperarse que el modelo de regresión lineal produzca bandas de credibilidad más anchas que el modelo de regresión logística, ya que al no depender τ de la edad, el modelo de regresión lineal permite involucrar mayor incertidumbre.

Ambos modelos tienen sus ventajas y desventajas, pero independientemente del modelo que se seleccione, hay que tener presente que la transformación preserva cualquier cuantil de la distribución predictiva condicional (de las tasas observadas dada la edad). Para ilustrar este beneficio de los métodos Bayesianos, en la Gráfica 3.6 presentamos el modelo ajustado (modificado) por cuantiles en el caso del ajuste con regresión lineal con transformación logística para el caso masculino.

Gráfica 3.6. Modelo de regresión lineal ajustado (modificado) por cuantiles para el caso masculino, 2000-2009



En este caso, se puede elegir el cuantil que nos interese como tabla de mortalidad, es decir esta es una manera de recargar una tabla de mortalidad

con un fundamento estadístico que valora, técnicamente, los efectos en términos de riesgo.

III.3 Comparación de modelos

Recordemos que en el caso de la regresión lineal con transformación logística se eliminaron los grupos de edad donde no hubo defunciones, mientras que para el caso de la regresión logística se consideraron todos los datos disponibles. A continuación se muestran los ajustes finales que resultaron de aplicar dichos modelos.

Cuadro 1. Cálculo de las probabilidades de fallecer por edad y género, 2000-2009

					Probabilidades de fallecer					
	Población expuesta		Defunciones		Observadas		Regresión lineal		Regresión logística	
Edad	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres
Total	43,098	18,022	426	212						
15-19	59	15	0	0	0.000000	0.000000	0.000000	0.000000	0.000053	0.000027
20-24	1,160	492	0	0	0.000000	0.000000	0.000000	0.000000	0.000091	0.000049
25-29	3,362	1,554	1	0	0.000297	0.000000	0.000200	0.000000	0.000157	0.000087
30-34	4,278	1,689	1	1	0.000234	0.000592	0.000337	0.000322	0.000269	0.000156
35-39	4,486	1,546	5	1	0.001115	0.000647	0.000579	0.000545	0.000461	0.000278
40-44	4,102	1,565	1	1	0.000244	0.000639	0.000980	0.000900	0.000792	0.000497
45-49	3,574	1,498	3	2	0.000839	0.001335	0.001644	0.001487	0.001360	0.000889
50-54	3,414	1,400	15	3	0.004394	0.002143	0.002797	0.002532	0.002335	0.001592
55-59	3,457	1,147	9	1	0.002603	0.000872	0.004733	0.004268	0.004009	0.002850
60-64	3,535	1,252	23	5	0.006506	0.003994	0.008018	0.007065	0.006878	0.005105
65-69	3,713	1,468	39	12	0.010504	0.008174	0.013680	0.011790	0.011780	0.009134
70-74	3,420	1,606	76	25	0.022222	0.015567	0.022880	0.019900	0.020120	0.016310
75-79	2,341	1,358	75	42	0.032038	0.030928	0.037800	0.032650	0.034180	0.028970
80-84	1,346	781	77	39	0.057207	0.049936	0.062860	0.053880	0.057490	0.051020
85-89	631	411	61	40	0.096672	0.097324	0.101600	0.086850	0.095170	0.088360
90-94	182	185	31	25	0.170330	0.135135	0.159200	0.135900	0.153500	0.148800
95-100	38	55	9	15	0.236842	0.272727	0.257400	0.223900	0.258500	0.261700

Para comparar los modelos se utilizó el criterio DIC (Deviance Information Criterion), el cual resume tanto el ajuste como la complejidad del modelo y además es calculado directamente por WinBUGS. Por otro lado, se obtuvo el error cuadrático medio (ECM) y la cantidad conocida como Ji-cuadrada (X^2):

$$ECM = \frac{1}{17}\sum_{i=1}^{17}(q_{obs_i} - \hat{q}_i)^2$$

y,

$$X^2 = \sum_{i=1}^{17} \frac{(q_{obs_i} - \hat{q}_i)^2}{\hat{q}_i}$$

Resulta conveniente diferenciar la interpretación de las medidas de comparación de modelos. El criterio DIC está relacionado con el ajuste del modelo (valores pequeños de DIC indican mejor ajuste), sin embargo el ECM y la Ji-cuadrada hacen referencia a la capacidad predictiva del modelo, es decir si las tasas de mortalidad observadas son muy similares a las tasas estimadas, el valor de $(q_{obs_i} - \hat{q}_i)^2$ será pequeño y consecuentemente, el valor del ECM o de la Ji-cuadrada también lo será.

El DIC, el ECM y la Ji-cuadrada son estadísticos que hacen referencia a criterios diferentes, en ocasiones se busca un mejor ajuste y en otras, se le da primacía a la capacidad predictiva del modelo. La decisión de elegir un modelo u otro depende del contexto del problema a tratar y de los objetivos que se persigan.

Debido a que la desventaja del modelo de regresión lineal con transformación logística radica en que dicho modelo no puede manejar valores iguales a cero, para fines comparativos, en el Cuadro 2 se muestran los resultados obtenidos de acuerdo a los distintos valores que se probaron para el número de defunciones d_x en vez de cero.

Finalmente, después de un análisis derivado de los resultados obtenidos en el Cuadro 2 y del comportamiento histórico de las tablas de mortalidad elaboradas en los últimos tres quinquenios, el departamento actuarial de la institución en

cuestión decidió seleccionar, para ambos géneros, el modelo de regresión logística utilizando todas las observaciones disponibles.

Cuadro 2. Medidas de comparación para los modelos bayesianos de regresión, por género

	Hombres		
	DIC	ECM	X^2
	Modelo lineal con transformación logística		
Quitando los ceros	27.093	0.000039	0.008478
dx=0 por dx=0.1	49.757	0.000073	0.030465
dx=0 por dx=0.01	46.117	0.000254	0.033815
dx=0 por dx=0.001	58.074	0.002195	0.174938
dx=0 por dx=0.0001	70.911	0.006738	0.453534
	Modelo de regresión logística		
Quitando los ceros	86.057	0.000043	0.007989
Con todos los datos	85.892	0.000045	0.008285

	Mujeres		
	DIC	ECM	X^2
	Modelo lineal con transformación logística		
Quitando los ceros	25.885	0.000151	0.018782
dx=0 por dx=0.1	59.772	0.000416	0.099608
dx=0 por dx=0.01	58.677	0.000319	0.075155
dx=0 por dx=0.001	66.93	0.002465	0.244223
dx=0 por dx=0.0001	77.007	0.007769	0.554798
	Modelo de regresión logística		
Quitando los ceros	63.196	0.000025	0.006705
Con todos los datos	63.171	0.000024	0.006868

IV. Conclusiones

La construcción de una tabla de mortalidad debe estar en función al uso que pretende dársele y debe tomar cuenta márgenes para posibles desviaciones. Independientemente de su relevancia actuarial, el problema de producir una tabla de mortalidad es en realidad de carácter estadístico. En este sentido, es

necesario reconocer que las técnicas actuariales convencionales por lo general no toman en consideración la incertidumbre involucrada en el proceso de estimación. De ahí que la ventaja de los modelos Bayesianos sea que éstos desarrollan criterios para sobrecargar las tablas de mortalidad de forma que valoran técnicamente los efectos en términos del riesgo.

V. Bibliografía

[1] Bernardo, J. M. (1981). *Bioestadística: Una perspectiva Bayesiana*. Vicens Vives.

[2] Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons.

[3] Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley & Sons.

[4] Debon, A. (2003). *Graduación de tablas de mortalidad. Aplicaciones actuariales.* Universitat De Valencia.

[5] Mendoza, M., Madrigal, A. M. y Martínez, E. (1999). "Modelos Estadísticos de Mortalidad. Análisis de datos 1991-1998", *Documento de Trabajo No.77, Comisión Nacional de Seguros y Fianzas*, México.

Apéndice

Modelo de regresión lineal con transformación logística (caso masculino)

```
model {
for (i in 1:15) {
#verosimilitud
        q[i] <-y[i]/n[i]
        yt[i]<-logit(q[i])
        yt[i] ~ dnorm(mu[i],tau);
#media
        mu[i] <- beta[1]+beta[2]*x[i];
```

```
#predicción
        mu1[i] <- beta[1]+beta[2]*x[i];
        yt1[i] ~ dnorm(mu1[i],tau)
        q1[i]<- exp(yt1[i])/(1+exp(yt1[i]))
}
#dist. iniciales
        beta[1] ~ dnorm(0.0,0.001);
        beta[2] ~ dnorm(0.0,0.001);
        tau ~ dgamma(0.001,0.001);
}
list(
        x=c(27., 32., 37., 42., 47., 52., 57., 62., 67., 72., 77., 82., 87., 92., 98.),
        y=c(1., 1., 5., 1., 3., 15., 9., 23., 39., 76., 75., 77., 61., 31., 9.),
        n=c(3362., 4278., 4486., 4102., 3574., 3414., 3457., 3535., 3713., 3420.,
        2341., 1346., 631., 182., 38.),
        )
list(beta=c(0,0),tau=1)
```

Modelo de regresión logística (caso masculino)

```
model {
for (i in 1:17) {
#verosimilitud
        d[i] ~ dbin(qi[i],e[i]);
#liga
        logit(qi[i]) <- beta[1]+beta[2]*x[i];
#prediccion
        logit(qihat[i]) <- beta[1]+beta[2]*x[i];
        dhat[i] ~ dbin(qihat[i],e[i]);
        }



#dist. iniciales
        beta[1] ~ dnorm(0.0,1.0E-3);
        beta[2] ~ dnorm(0.0,1.0E-3);
}
list(
        x=c(17., 22., 27., 32., 37., 42., 47., 52., 57., 62., 67., 72., 77., 82., 87.,
92., 98.),
        d=c(0., 0., 1., 1., 5., 1., 3., 15., 9., 23., 39., 76., 75., 77., 61., 31., 9.),
        e=c(59., 1160., 3362., 4278., 4486., 4102., 3574., 3414., 3457., 3535.,
3713., 3420., 2341., 1346., 631., 182., 38.),
        )
list(beta=c(0,0))
```

Árboles de Regresión

Daniel Iván Ugalde Gutiérrez

Metlife, México

Rinc. Pintores, Edif. Coronel 104, Col. Pedregal de Carrasco, C.P. 04700.

diug85@yahoo.com.mx

Teléfono: (+52 55) 56652229

I. Introducción

El propósito de este documento es mostrar una alternativa para abordar el problema de pronosticar los valores futuros de alguna variable con una técnica distinta a los modelos paramétricos más comúnmente utilizados como son los modelos lineales, en particular los de regresión.

Los modelos mencionados suelen tener problemas al incorporar variables explicativas categóricas, tienen problemas en la ausencia de relación lineal entre las variables explicativas con la variable de respuesta, presentan dificultades cuando no se cumplen algunos supuestos (normalidad, no correlación, etc.), las estimaciones se ven altamente afectadas por la presencia de datos atípicos, entre otros. Adicionalmente, estos modelos tienen la desventaja de presentar dificultad al interpretar los resultados, pues estos se resumen en una ecuación y su representación gráfica queda limitada cuando el número de variables explicativas es mayor a 2.

La técnica de árboles de regresión es una metodología no paramétrica que cumple con algunas características favorables: su metodología es sencilla, su aplicación (gracias a las computadoras) es fácil y rápida y resuelve los problemas mencionados: implementan las variables categóricas sin problemas, funcionan bien cuando no hay relación lineal entre las variables, son robustos

ante datos atípicos, etc., además de que la interpretación de sus resultados es sencilla, pues se pueden representar gráficamente sin problemas.

Los árboles de regresión pertenecen a la familia de los árboles binarios, los cuales tienen una estructura formada por un conjunto de nodos, de los cuales el nodo principal es dividido en dos nodos, posteriormente cada uno de los nuevos nodos es vuelto a dividir. Se sigue dividiendo cada uno de los nuevos nodos hasta obtener los nodos terminales, que son definidos a partir de una regla de paro.

En los árboles de regresión cada nodo representa un subconjunto de la población para la cual se proporciona una predicción. La forma para decidir cómo se divide cada nodo para crear dos nuevos, así como la regla para dejar de dividir y así obtener los nodos terminales son cuestiones que se resolverán a lo largo de este artículo.

Al tratarse de una técnica no paramétrica los árboles de regresión no hacen ningún supuesto distribucional sobre las variables explicativas, lo que en principio evita todo el trabajo que implica hacer estimaciones y pruebas estadísticas para comprobar supuestos, etc. Sin embargo, más allá de proporcionar una predicción puntual y una medida del error cometido por el árbol, no es posible realizar inferencia.

Los únicos supuestos que hacen estos árboles son independencia entre las observaciones y que las observaciones de la muestra provienen de la misma distribución (sólo para el cálculo de la desviación estándar del estimador del error cometido).

II. Árboles de regresión

Sea $\mathcal{L}$ la muestra que contiene las mediciones observadas, a partir de las cuales se construirá el árbol. Los árboles de regresión se construyen al tomar la muestra $\mathcal{L}$ y crear biparticiones sucesivas de ella, separándola cada vez en

dos subconjuntos mutuamente excluyentes (nodos). Enseguida se toma cada uno de estos nuevos subconjuntos y se repite sucesivamente el proceso de

hacer la bipartición hasta que una regla de paro (previamente definida) indique que se debe dejar de dividir cada subconjunto, entonces se llega a los nodos terminales, los cuales proporcionan un valor de predicción.

El procedimiento para crear un árbol de regresión es el siguiente:

1. Se toma como nodo $i = 1$ a $\mathcal{L}$.
2. Se buscan todas las posibles particiones del nodo que lo dividan en dos subconjuntos (nodos $2i$ y $2i + 1$)
   2.1. Con la información proporcionada por los registros contenidos en el nodo $2i$ se obtiene un pronóstico puntual para la variable de respuesta.
   2.2. Se hace lo mismo con el nodo $2i + 1$.
   2.3. Se mide el error cometido por los pronósticos de los nodos $2i$ y $2i + 1$
3. De las posibles particiones encontradas en el punto 2 se escoge aquella con el menor error.
4. Si existe otro nodo que no sea terminal se toma dicho nodo y se continua en el punto 5, en otro caso se salta al paso 6.
5. Si se cumple la regla de paro se considera nodo terminal y se regresa al paso 4, en otro caso se regresa al paso 2.
6. Ya que se tiene un árbol (con tantos nodos como lo permita la regla de paro) se podan algunas de sus ramas (en ocasiones ninguna) para obtener un árbol con buenos resultados pero que no sea demasiado grande (parsimonia).
7. Fin

El procedimiento descrito en los 7 pasos anteriores es el que se puede ver en el Anexo 1.

Lo primero que se necesita para hacer las particiones de $\mathcal{L}$ es contar con un conjunto de preguntas (cuya respuesta sea “verdadero” o “falso”) para conocer las posibles particiones.

Tipos de preguntas

Sea $x = (x_1, x_2, \cdots, x_M)$ el vector de observaciones, en donde M es la dimensión (fija) del vector y las variables $x_1, x_2, \cdots, x_M$ son variables numéricas, categóricas o una mezcla de ambas.

Las preguntas para buscar la mejor partición deben tener como respuesta “Verdadero” o “Falso” y debe quedar en función de una sola variable x_m. Entonces la serie de preguntas que proporcionan dichas particiones pueden ser de dos tipos, dependiendo de la naturaleza de la variable con la que se esté tratando:

- Si la variable es numérica las preguntas son del tipo ¿$x_m \leq c$?, donde c es un valor observado dentro de $\mathcal{L}$.
- Si la variable es categórica con posibles valores $\{b_1, b_2, \cdots, b_L\}$, las preguntas son del tipo ¿$x_m \in S$?, donde S es algún elemento del conjunto $P(\{b_1, b_2, \cdots, b_L\}) - \emptyset - U$, es decir, del conjunto potencia de $\{b_1, b_2, \cdots, b_L\}$ sin contar el conjunto vacío ($\emptyset$) y el universo (U).

Se debe notar que el número de preguntas es finito, pues estas son obtenidas a través de los valores observados en la muestra $\mathcal{L}$. En el caso de variables:

- numéricas: el número de posibles preguntas es $N' \leq N$, en donde N es el número de observaciones en la muestra;
- categóricas: considerando que las preguntas ¿$x_m \in S$? y ¿$x_m \in S^c$? generan la misma partición y que no se cuenta la pregunta ¿$x_m \in \mathrm{U}$?, en donde U es el conjunto universo, pues no generaría ninguna partición, el número total de posibles subconjuntos S es $2^{L-1} - 1$.

Para poder realizar los pasos 2.1, 2.2 y 2.3 del procedimiento descrito es necesario conocer una regla de predicción y una forma de medir los errores de dichas predicciones.

Regla de predicción y estimación de los errores

En regresión lineal múltiple un cambio pequeño en cualquiera de las variables explicativas implica también un cambio pequeño en la variable de respuesta. En los árboles de regresión un cambio pequeño en alguna (o varias) de las variables explicativas no necesariamente implica que también exista un cambio en la variable de respuesta estimada. Incluso cuando llega a haber un cambio en la variable de respuesta no siempre es “pequeño”.

Por otra parte, si se piensa en la superficie de predicción ajustada mediante regresión lineal múltiple como una sabana, entonces las estimaciones proporcionadas por el árbol de regresión se podrían pensar como un histograma n-dimensional que aproxima a la sabana, el cual proporciona el mismo valor de predicción para una región dada de posibles valores de las variables explicativas.

La variable de respuesta de estos árboles es una variable numérica continua, por lo que los valores predichos por estos árboles pueden ser comparados casi de manera inmediata con las predicciones obtenidas de un modelo de regresión lineal.

En los modelos de regresión lineal, la forma de medir qué tan bueno es un ajuste a un conjunto de datos es con el coeficiente de determinación $R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$, en palabras: la proporción de la varianza total explicada por el modelo de regresión ajustado.

De manera similar, el error de los árboles de regresión se mide con el error cuadrático medio (ECM)[1]:

$$R^*(d) = E\left[\left(Y - d(x)\right)^2\right]$$

en donde $d(x)$ es la regla de predicción para el árbol de regresión. Este ECM se calcula dentro de cada nodo terminal y la suma de estos es el ECM de todo el árbol.

Una regla de predicción es una función $d(x)$ que asigna un valor, que como su nombre lo dice, busca predecir o pronosticar el valor de la variable de respuesta para cada observación x. Partiendo de la definición del ECM y utilizando el concepto de esperanza condicionada, se puede demostrar que la regla de predicción que minimiza el ECM es:

$$d(x) = E(Y|X = x)$$

(Breiman *et al*, 1984), de manera que el mejor estimador de la variable de respuesta para cualquier nodo es el promedio de todas las y_i contenidas dentro de dicho nodo, es decir que el valor asignado al nodo t es[2]:

$$\bar{y}(t) = \frac{1}{N_t}\sum_{y_i \in t} y_i$$

A continuación se presentan tres metodologías para hacer estimaciones de estos errores.

Estimadores de resustitución

Este tipo de estimadores utilizan todos los registros contenidos en $\mathcal{L}$ tanto para hacer crecer el árbol como para hacer la estimación del ECM, por lo que tienen un sesgo grande y tienden a subestimar el error.

[1] También se puede definir como el valor absoluto de las diferencias, pero algunos de los resultados que se presentan serían distintos (Breiman *et al*, 1984).

[2] Si el error se estuviera midiendo con el valor absoluto de las diferencias, entonces el mejor estimador para el valor de predicción sería la mediana (Leo Breiman *et al*, 1984).

La forma en la que se estima el ECM con este tipo de estimadores es:

$$R(d) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - d(x_i)\right)^2$$

Estimadores con muestra de prueba

Estos estimadores se calculan dividiendo la muestra $\mathcal{L}$ con cardinalidad N en dos partes: una para crear el árbol y otra para estimar el error cometido con el árbol. Se recomienda utilizar estos estimadores cuando N es grande.

Se seleccionan aleatoriamente $N^{(2)}$ registros de $\mathcal{L}$ para formar la muestra de prueba $\mathcal{L}_2$. El resto $\mathcal{L}_1 = \mathcal{L} - \mathcal{L}_2$ es la muestra de aprendizaje con cardinalidad $N^{(1)}$. Entonces $N = N^{(1)} + N^{(2)}$.

Se hace crecer el árbol usando únicamente los casos contenidos dentro de $\mathcal{L}_1$. Una vez que se tienen los nodos terminales del árbol, se usa dicho árbol para pronosticar los casos contenidos en $\mathcal{L}_2$.

Como se conoce a priori la el valor de la variable de respuesta de cada registro de $\mathcal{L}_2$ se puede estimar el ECM ($R^{ts}(T)$) con los valores observados y pronosticados por el árbol para los registros de $\mathcal{L}_2$.

La forma en la que se calculan estos estimadores es la siguiente:

$$R^{ts}(d) = \frac{1}{N_2}\sum_{(x_i,y_i)\in\mathcal{L}_2}\left(y_i - d(x_i)\right)^2$$

Estimadores de validación cruzada

Este estimador hace un uso exhaustivo de la muestra, pues utiliza cada registro más de una vez para hacer crecer varios árboles, pero utiliza cada registro solo una vez para calcular el error. Es una buena manera de estimar el error, pero al

ser computacionalmente muy demandante, se sugiere utilizar únicamente cuando se tienen pocos registros como para usar los estimadores de muestra de prueba.

En esta forma de calcular el error se divide aleatoriamente la muestra $\mathcal{L}$ en V subconjuntos $\mathcal{L}_v$, cada uno con el mismo número de elementos (o lo más cercano posible).

Las muestras de aprendizaje son entonces

$$\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v, \quad v = 1, \ldots, V.$$

Por lo que cada muestra $\mathcal{L}^{(v)}$ contiene solamente $\left(\frac{V-1}{V}100\right)\%$ de los datos para crear el árbol. Se suele usar v=10 para que cada muestra de aprendizaje tenga 90% del total de datos.

Ahora con cada muestra de aprendizaje $\mathcal{L}^{(v)}$, v=1, … , V, se hace crecer un árbol $T^{(v)}$ y este se usa para pronosticar los registros contenidos en cada muestra de prueba $\mathcal{L}_v$. Se debe observar, como se menciono anteriormente, que cada observación aparece sólo una vez en alguna muestra de prueba $\mathcal{L}_v$, por lo que el error se puede calcular de la siguiente manera:

$$R^{cv}(d) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(x_i,y_i)\in\mathcal{L}_v}\left(y_i - d^{(v)}(x_i)\right)^2$$

Una vez que se hace la estimación del ECM para cada par de nodos creados por una partición dada ya se tiene la información necesaria para poder saber cuál de esos nodos es el que minimiza el error. Ahora se tiene que responde la pregunta de cuándo dejar de hacer particiones de los nodos, cuya respuesta es la regla de paro.

La regla de paro es un criterio que indica el momento o las condiciones que se deben cumplir para que se deje de hacer particiones de los nodos. Esta regla se puede definir de varias maneras:

- cuando el número de observaciones contenidas dentro del nodo alcanza un umbral mínimo,
- cuando todas las observaciones son idénticas, es decir, cuando su variable de respuesta sea la misma o
- cuando el ECM alcance un valor mínimo predefinido con anterioridad.

Ahora ya podemos hacer crecer un árbol de regresión tan grande como lo permita la regla de paro definida, por lo que se puede pasar al punto 6 del procedimiento, es decir, a la poda del árbol con la finalidad de obtener un modelo parsimonioso, en el sentido de que tenga un ECM aceptable pero que no sea demasiado grande.

Poda del árbol

La idea que se ha planteado es escoger aquel árbol que proporcione el menor error a la hora de pronosticar datos futuros. En este sentido, se podría pensar que el mejor árbol debería ser uno muy grande, sin embargo, la experiencia ha demostrado que cuando se usan árboles con demasiados nodos terminales el error tiende a ser mayor a partir de cierto punto (Breiman *et al*, 1984). Entonces surge la necesidad de podar el árbol hasta obtener uno de tamaño adecuado pero que no pierda eficiencia a la hora de pronosticar.

En regresión lineal múltiple, cuando se sobreparametriza un modelo el coeficiente de determinación incrementa. De la misma manera, cuando se hace crecer un árbol demasiado grande el ECM del árbol incrementa, mientras que cuando un árbol es demasiado pequeño no aprovecha toda la información de la muestra, por lo que resulta indispensable encontrar un árbol de tamaño correcto, con el cual no se desaproveche la muestra pero tampoco se haga crecer demasiado como para que pierda precisión en sus estimaciones.

Adicionalmente, si se supone el caso extremo en el que los nodos terminales de un árbol de regresión contengan solamente una observación, el estimador de resustitución del ECM sería cero, lo cual es difícil de creer que sea cierto,

por lo que se usan los estimadores con muestra de aprendizaje o de validación cruzada, los cuales proporcionan estimaciones más realistas del error.

Antes de desarrollar la poda para encontrar el árbol optimo, es necesario presentar algunos conceptos preliminares.

Antes de podar

Antes de presentar la metodología para escoger el árbol de tamaño correcto es necesario conocer las siguientes definiciones y conceptos:

i. Una rama $T_{\{t\}}$ con un nodo raíz $t \in T$ consiste del nodo {t} y todos sus descendientes. Ver Figura 1

ii. Podar la subrama $T_{\{t\}}$ del árbol T consiste en borrar todos los descendientes de {t} excepto su nodo raíz. Se denota como $T - T_{\{t\}}$. Ver Figura 2

iii. Si un árbol T' es obtenido al realizar podas sucesivas del árbol T se dice que el árbol T' es un subárbol podado de T y se denota como $T' \preccurlyeq T$

iv. $|\tilde{T}|$ es la complejidad de un árbol T y se define como el número de sus nodos terminales.

v. T_{max} es el árbol con el mayor número de nodos terminales posibles

Enseguida se desarrolla la manera en la que se debe podar un árbol de regresión.

Figura 1: Nodos y ramas

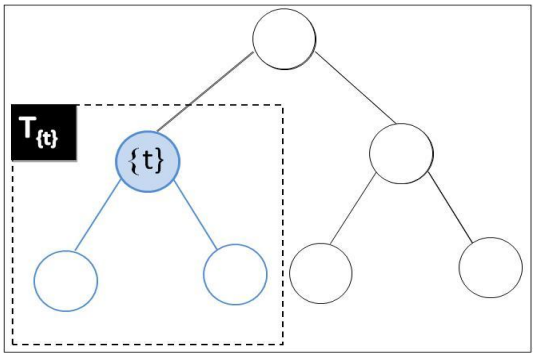
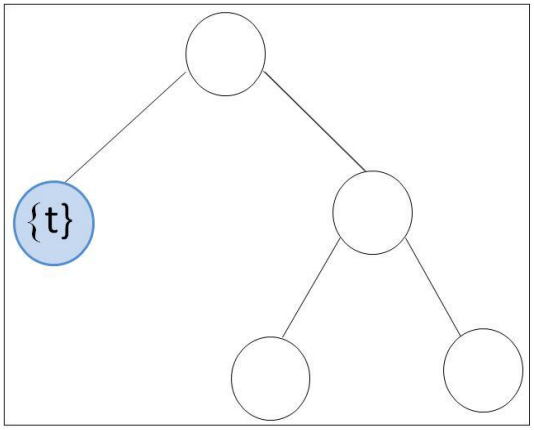


Figura 2: Rama podada



Poda de mínimo costo-complejidad

Bajo el principio de parsimonia es necesario hallar un árbol que tenga poca complejidad pero que a la vez tenga un error tan pequeño como sea posible, para lo cual se usa una medida basada en el estimador del error, pero que además penaliza a los árboles con mayor complejidad. Esta medida se define a continuación.

Sea $\alpha \geq 0$ el parámetro de complejidad, cuya finalidad es penalizar a los árboles con mayor número de nodos terminales. Se define la medida de costo-complejidad como:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|,$$

es decir, se trata del error penalizado por su complejidad.

La idea del procedimiento que esta por describirse es encontrar una serie de árboles
$T_1 \succ T_2 \succ \cdots \succ T_K$ cada vez de menor complejidad, con la característica de que si se comparan con cualquier otro árbol con la misma complejidad estos sean los que tienen el error más pequeño. Una vez que se conocen cuales son estos árboles lo que resta es decidir qué tamaño es el más adecuado.

El paso fundamental para encontrar cada uno de los árboles es buscar para cada nodo $\{t\}$ y cada subrama T_t el valor de α que hace preferible podar la subrama T_t, pues a partir de ese valor es cuando se encuentra el siguiente árbol de la serie.

Lo anterior se hace resolviendo la desigualdad

$$R_\alpha(\boldsymbol{T}_{\{t\}}) < R_\alpha(\{\boldsymbol{t}\})$$

con lo que se obtiene

$$\alpha < \frac{R(\{\boldsymbol{t}\}) - R(\boldsymbol{T}_{\{t\}})}{|\widetilde{\boldsymbol{T}_{\{t\}}}| - 1}$$

Ahora, para cualquier subrama T_t y cualquier nodo no terminal $\{t\}$ del árbol T_1 se cumple que:
$R(\{t\}) > R(T_{\{t\}})$

(Breiman *et al.*, 1984), por lo que los valores de α para los cuales se prefiere la subrama T_t sobre su nodo raíz $\{t\}$ son todos positivos.

El procedimiento para encontrar la serie de árboles $T_1 \succ T_2 \succ \cdots \succ T_K$ así como su serie de parámetros de complejidad $\alpha_1 < \alpha_2 < \cdots < \alpha_K$, es el siguiente:

i. Sean el árbol $T_1 = T_{max}$ y la función

$$g_1(\{t\}) = \begin{cases} \dfrac{R(\{t\}) - R(T_{1_{\{t\}}})}{\left|\tilde{T}_{1_{\{t\}}}\right| - 1} & t \notin \tilde{T}_1 \\ \infty & t \in \tilde{T}_1 \end{cases}$$

Se evalúa esta función en cada nodo, se determina el nodo $\{t_1^*\}$ que minimice la función $g_1(\{t\})$ y se define $\alpha_2 = g_1(\{t_1^*\})$.

La interpretación es la siguiente: conforme el parámetro de complejidad α (penalización por complejidad) aumenta, el nodo $\{t_1^*\}$ es el primero que hace que al podar sus ramas la medida de costo-complejidad del árbol sea menor que sin podar, es decir que $R_\alpha\left(T_{1_{\{t_1^*\}}}\right) > R_\alpha(\{t_1^*\})$; y el valor del parámetro α en el que se cumple esto es α_2.

ii. Se define el árbol $T_2 = T_1 - T_{1_{\{t_1^*\}}}$ (a T_1 se le podan las subramas de $\{t_1^*\}$), se define la función

$$g_2(\{t\}) = \begin{cases} \dfrac{R(\{t\}) - R(T_{2_{\{t\}}})}{\left|\tilde{T}_{2_{\{t\}}}\right| - 1} & t \notin \tilde{T}_2 \\ \infty & t \in \tilde{T}_2 \end{cases}$$

y se sigue el mismo procedimiento[3] para encontrar su respectiva α_3.

iii. Se repite el mismo procedimiento hasta llegar al subárbol $T_K = \{t_1\}$, en donde $\{t_1\}$ es el nodo raíz del árbol original.

Al seguir estos pasos se encuentran la serie de árboles $\{T_k\}$ y la serie de parámetros de complejidad $\{\alpha_k\}$. Lo que resta ahora es escoger cuál de estos árboles es el mejor.

Nótese que para $\alpha_k \leq \alpha < \alpha_{k+1}$ el árbol de mínimo costo-complejidad $T(\alpha)$ es el mismo.

[3] Si para el subárbol k en algún punto existen varias subramas que proporcionan el mismo valor (mínimo) de $g_k(\{t\})$, el árbol subárbol k+1 se define podando todas estas subramas del árbol k.

Por la forma en la que se define el costo-complejidad $\mathrm{R}_{\alpha}()$ y la forma en la que se escoge la serie de árboles y parámetros de complejidad se asegura que dado el árbol $\mathrm{T}(\alpha)$ no existe ningún otro subárbol con la misma complejidad pero con menor error $\mathrm{R}()$.

A continuación se muestran algunas consideraciones que se deben tomar en cuenta cuando se quiere usar este algoritmo con los estimadores de validación cruzada.

Implementación de la poda con errores de validación cruzada

La implementación de la poda requiere que se calculen todos los árboles de costo-complejidad para cada valor de α usando tanto $\mathcal{L}$ como $\mathcal{L}^{(v)},\ v = 1, 2, \ldots, V$.

Tomando la serie $\{\alpha_k\}$ calculada a partir del árbol ajustado a la muestra $\mathcal{L}$, se debe recordar que árbol T_k es el de mínimo costo-complejidad para $\alpha_k \leq \alpha < \alpha_{k+1}$, entonces se define α'_k como la media geométrica de α_k y α_{k+1}:

$$\alpha'_k = \sqrt{\alpha_k\ \alpha_{k+1}}$$

y nótese que $T(\alpha'_k) = T_k$.

La estimación del error con estimadores de validación cruzada es:

$$R^{CV}(T_k) = R^{CV}(T(\alpha'_k))$$

en donde $R^{CV}(T(\alpha'_k))$ es el error estimado, con la única diferencia de que no se calcula con T_{max} sino con $T(\alpha'_k)$.

Teniendo las herramientas y la información necesaria para realizar la poda, ahora se verán los argumentos que se toman en cuenta para decidir cuál de los

árboles de la serie que se acaba de generar es el adecuado para su uso en el futuro.

Selección del mejor árbol

La medida de costo-complejidad permitió encontrar el árbol que minimiza la complejidad (parsimonia) dado un nivel de error cometido, con lo que se obtuvo la serie de árboles $T_1 \succ T_2 \succ \cdots \succ T_K$ y la serie de parámetros $\alpha_1 < \alpha_2 < \cdots < \alpha_K$, pero aun queda la tarea de escoger cuál de estos árboles es el óptimo para clasificar en el futuro.

Una idea inicial para escoger dicho árbol sería escoger aquel con el menor error, la cual tiene un inconveniente: si se supone que el árbol con menor error es T_k y se vuelven a estimar sus errores pero usando distintas semillas (de números aleatorios) para generar ya sea la muestra de prueba (estimadores con muestra de prueba) o las particiones aleatorias (estimadores de validación cruzada), podría darse el caso de que los nuevos errores no coincidieran con que T_k sea el de menor error.

Para solucionar esto y estabilizar la variabilidad existente en las estimaciones del error se utiliza la regla de una desviación estándar.

Esta regla establece lo siguiente:

i. Se escoge T_{k_0} como el subárbol con el menor error

ii. Se escoge T_{k_1}como el subárbol más pequeño que cumpla que su error sea menor al error de T_{k_0} más una desviación estándar. Esto es:

$$\hat{R}(T_{k_1}) \leq \hat{R}(T_{k_0}) + \sigma\left(\hat{R}(T_{k_0})\right)$$

en donde:

- k_1 es el máximo valor posible

- $\sigma\left(\widetilde{\mathrm{R}}\left(\mathrm{T}_{k_0}\right)\right)$ es la desviación estándar de la estimación del error de T_{k_0}

iii. Si existe tal T_{k_1}, entonces dicho árbol es el que se debe usar, en otro caso T_{k_0} es el árbol que se debe usar.

Estimando la desviación estándar

Nótese que como cada elemento de la muestra es independiente entre sí, se tiene que la varianza del estimador R^{ts} es la suma de las varianzas de cada uno de sus términos. Además, como todos los términos tienen la misma distribución, se puede decir sin pérdida de generalidad que la varianza de cada uno de los términos es igual a la varianza del primero de ellos, por lo que la varianza del estimador R^{ts} es:

$$Var(R^{ts}) = \frac{1}{N_2}\left(E\left[\left(Y_1 - d(Y_1)\right)^4\right] - E\left[\left(Y_1 - d(Y_1)\right)^2\right]^2\right)$$

Usando los estimadores de momentos muestrales se tiene que

$$E\left[\left(Y_1 - d(Y_1)\right)^4\right] \cong \frac{1}{N_2}\sum_{i=1}^{N_2}\left(y_i - d(x_i)\right)^4$$

y

$$E\left[\left(Y_1 - d(Y_1)\right)^2\right] \cong \frac{1}{N_2}\sum_{i=1}^{N_2}\left(y_i - d(x_i)\right)^2 = R^{ts}$$

Entonces la estimación del error estándar del estimador del ECM para el caso de muestra de aprendizaje[4] es:

$$\hat{\sigma}(R^{ts}) = \frac{1}{\sqrt{N_2}}\sqrt{\frac{1}{N_2}\sum_{i=1}^{N_2}\left(y_i - d(x_i)\right)^4 - (R^{ts})^2}$$

[4] La estimación de la desviación estándar para el caso de estimadores de validación cruzada se puede consultar en Leo Breiman *et al*, 1984

Debido a la variabilidad del cuarto momento para calcular la desviación estándar de los árboles de regresión suele ocurrir que la regla de una desviación estándar sugiera que el árbol óptimo sea uno con un ECM muy grande o con muy pocos nodos terminales, por lo cual se sugiere usar el conocimiento previo del experimento que se está estudiando y poner a prueba los resultados de T_{k_0} y T_{k_1} antes de decidir cuál de los dos será que se utilizará para la predicción futura de datos.

III. Ejemplo

Como se puede deducir de la metodología, estos árboles funcionan mejor cuando el dominio de las variables explicativas puede ser replicado mediante rectángulos cuya base no presenta rotación con respecto a los ejes, situación que en el caso de los modelos lineales puede provocar que sus resultados sean pésimos.

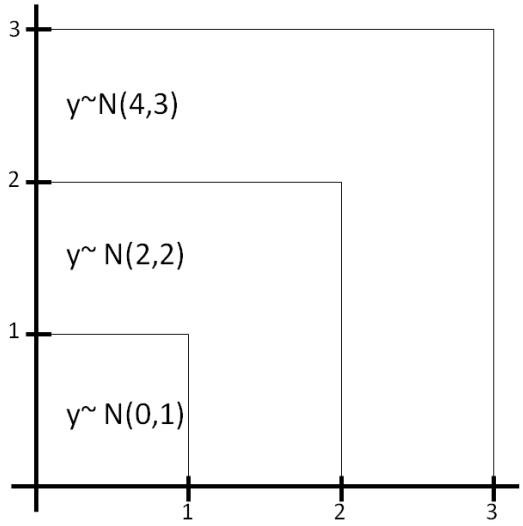
A continuación se presenta un ejemplo con datos simulados (sin variables categóricas), creado para mostrar una situación ideal en la que estos árboles podrían ser adecuados y más recomendables que algún modelo lineal.

Se simularon 1500 observaciones de dimensión 3 $x = (x_1, x_2, x_3)$ con $x_i \sim U(0,3),\ i = 1,2,3$ y su respectiva variable de respuesta, en donde

$$y \sim \begin{cases} N(\mu = 0, \sigma^2 = 1) & si \quad x_1 < 1 \ y \ x_2 < 1 \\ N(\mu = 2, \sigma^2 = 2) & si \quad x_1 < 2 \ y \ x_2 < 2 \\ N(\mu = 4, \sigma^2 = 3) & si \qquad e.o.c. \end{cases}$$

El uso de la variable x_3 es únicamente para mostrar que el algoritmo con el que se ajustan estos árboles automáticamente selecciona las variables que sirven para describir a la variable de respuesta. La representación gráfica de la relación entre las variables x_1, x_2 y y es la que se muestra en la Figura 3.

Figura 3:Relación entre x1, x2, y



Usando el software R y el paquete tree del mismo se ajustó un árbol de regresión en el cual se definió la regla de paro de manera que el número mínimo de observaciones dentro de los nodos fuera 25. El árbol obtenido es el que se muestra en el Anexo 2, en donde la expresión que se encuentra **sobre** cada partición es la pregunta que envía a las observaciones al nuevo nodo del lado izquierdo en caso de que su respuesta sea “Verdadero” y al nodo derecho en caso contrario. El número que aparece **debajo** de los nodos terminales es el valor con el que dichos nodos pronosticarán a las observaciones que caigan dentro de ellos.

En la gráfica del Anexo 2 se observa que el algoritmo de estos árboles sigue buscando una partición tras otra hasta que la regla de paro se cumpla, sin importar si estas particiones son de ayuda o no. Encontrar cual de las particiones son de ayuda y cuales se deben eliminar es trabajo del proceso de poda del árbol.

Para poder podar el árbol se simuló la muestra de prueba con 500 registros provenientes de la misma distribución que la muestra de aprendizaje, entonces generó la serie de árboles de mínimo costo-complejidad. El resumen de dicha serie de árboles se encuentra en la Tabla 1.

Tabla 1: Árboles de mínimo costo-complejidad

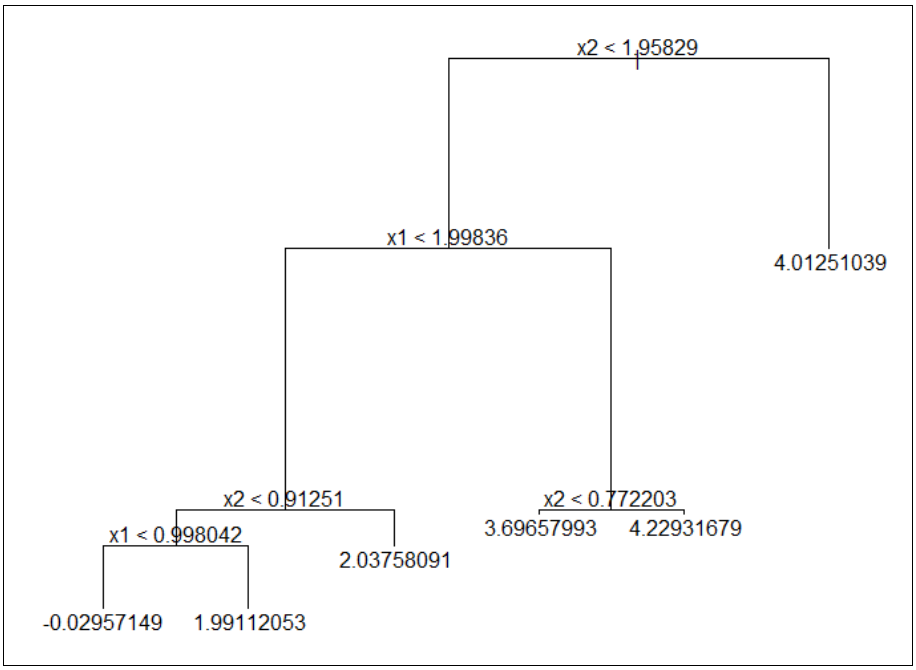
$\lvert\widetilde{T}\rvert$	α	ECM	σ	ECM+σ
36	0	2.76	0.209	2.97
35	4.42	2.76	0.209	2.97
32	5.83	2.74	0.208	2.94
31	6.98	2.72	0.207	2.93
28	8.05	2.69	0.205	2.90
26	8.58	2.68	0.204	2.88
25	8.82	2.67	0.203	2.88
24	9.08	2.66	0.202	2.87
23	10.02	2.65	0.202	2.85
21	10.19	2.66	0.202	2.86

$\lvert\widetilde{T}\rvert$	α	ECM	σ	ECM+σ
19	12.56	2.65	0.202	2.85
13	13.07	2.59	0.198	2.79
12	14.05	2.59	0.197	2.78
9	14.15	2.55	0.194	2.75
7	14.38	2.53	0.190	2.72
6	16.69	2.50	0.189	2.69
5	22.70	2.51	0.187	2.70
3	249.75	2.73	0.195	2.92
1	1143.53	4.00	0.269	4.27

En la Tabla 1 se aprecia que conforme el tamaño del árbol incrementa decrementa el error cometido, pero llega un punto en el que agregar más nodos terminales en lugar de ayudar provoca que el error incremente de nuevo (equivalente a sobreparametrizar un modelo lineal). Según la Tabla 1 el árbol de error mínimo es el de 6 nodos, pero la regla de una desviación estándar sugiere utilizar el de 5 nodos, pues el ECM del árbol de 3 nodos ya es mayor a 2.69.

El árbol de 6 nodos terminales es el que se puede ver en la
Figura **4**. Las particiones descritas por este árbol son parecidas a las de la Figura 3, excepto por la partición que se encuentra debajo de la condición $x_2 < 0.772203$ que difiere de la Figura 3, además de que sus valores de predicción son lejanos a 4 (que es el valor de la media que le corresponde).

Figura 4: Árbol podado con 6 nodos terminales



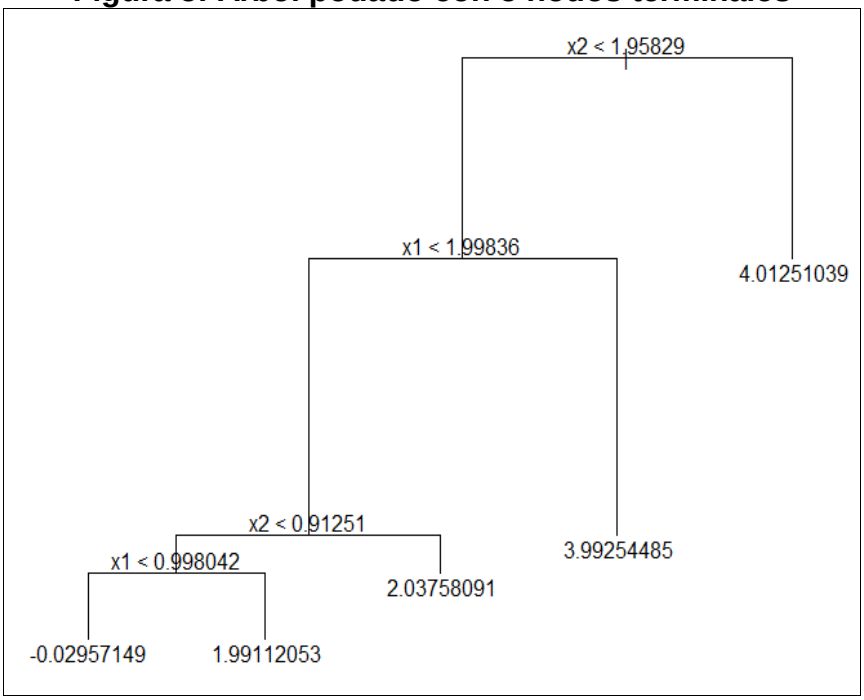
En la Figura 5 se encuentra el árbol sugerido por la regla de una desviación estándar, en donde se ve que las particiones (comparadas con las de la Figura 3) mejoran, además de que el valor de predicción del nodo que causaba problemas ahora es cercano a 4, por lo que (basado en la metodología y el conocimiento del experimento) se decide que el árbol óptimo es el de 5 nodos terminales.

IV. Conclusiones

La metodología de árboles de regresión se presenta como una alternativa más a las tradicionalmente utilizadas para el ajuste y pronóstico de datos, pues son una técnica no paramétrica de fácil implementación que suelen funcionar adecuadamente en situaciones en las que los modelos más comúnmente usados normalmente presentan problemas.

En el ejemplo que se presentó en este artículo teníamos el conocimiento previo de que no todas las observaciones provenían de la misma distribución, pues los parámetros de la variable y estaban condicionados en los valores que tomaran x_1 y x_2. La única consecuencia de lo anterior es que para calcular la desviación estándar se debió haber calculado la varianza de cada nodo y calcular la raíz cuadrada de la suma de todas ellas, lo cual es válido pues también se supone independencia de las observaciones.

Figura 5: Árbol podado con 5 nodos terminales



Retomando lo mencionado en el párrafo anterior y los resultados obtenidos en el ejemplo es relevante mencionar que estos árboles podrían ser usados no sólo para el pronóstico de datos, sino también como análisis exploratorio de ellos, pues su algoritmo busca agrupaciones naturales de las observaciones, de manera que si se tiene la sospecha de que dentro de la muestra no todas provienen de la misma distribución o fenómeno, los nodos terminales podrían estar identificándolas y agrupándolas con aquellas que provengan de la misma distribución.

Otro uso que se suele dar a estos árboles es únicamente para explicar o describir la relación entre variables. Por ejemplo, si se piensa en una base de datos de bienes raíces, se podría utilizar estos árboles para averiguar qué variables (distancia a centros comerciales o de trabajo, ubicación, número de recamaras, colonia, delegación, antigüedad, etc.) son las que hacen más o menos valiosa a las propiedades, en donde el tamaño del error queda de lado.

Finalmente, se resume a continuación las principales ventajas y desventajas de utilizar estos árboles para la predicción de datos:

Ventajas

- Al tratarse de una metodología no paramétrica no es necesario hacer supuestos distribucionales.
- Hacen automáticamente la selección de variables, equivalente a la selección del modelo en regresión lineal.
- Los resultados no tienen cambios ante la presencia de transformaciones uno a uno de las variables explicativas.
- Los árboles dan buenos resultados cuando la relación entre las variables explicativas con la variable de respuesta forma regiones fácilmente replicables con rectángulos. Además, se puede hacer combinaciones de las variables explicativas para hacer el ajuste del árbol sobre ellas.
- Entre más variables categóricas existan o entre más posibles valores tengan dichas variables resulta más práctico el uso de árboles, pues con otras metodologías se debe crear demasiadas variables indicadoras para ajustar un modelo. Esto no significa que de mejores resultados, solamente sugiere que un primer acercamiento puede ser con árboles de regresión por su practicidad.
- Los árboles de regresión tienen mayor capacidad descriptiva, pues se pueden representar mediante un dendrograma, que gráficamente es sencillo analizar. En cambio los métodos tradicionales generan una ecuación, que no en todos los casos describen de manera sencilla la relación entre variables y presentan dificultades para ser graficados debido a la dimensión (número de variables explicativas).

Desventajas

- La elección del árbol de tamaño óptimo es complicada, debido a que la regla de una desviación estándar no siempre funciona.
- Cambios pequeños en las variables explicativas no necesariamente se refleja en cambios pequeños en la predicción, pues la superficie de predicción no es suave, sino que es la unión de tantos planos como nodos terminales existan en el árbol.
- Si las variables explicativas de una nueva observación toman valores fuera del rango observado dentro de la muestra puede suceder que no exista un nodo terminal que describa adecuadamente a la nueva observación.

V. Bibliografía

- Acuña, Edgar. Notas de clase: Árboles de Decisión, University of Puerto Rico, Mayagüez campus [12 de cidiembre del 2010], URL: http://math.uprm.edu/~edgar/trees1.pdf y http://math.uprm.edu/~edgar/trees2.pdf

- Breiman Leo, Friedman Jerome H., Stone Charles J., Olshen Richard A. Classification and Regression Trees. Chapman & Hall, 1993, Capítulos 1, 2, 3, 4, 5 y 8.

- Qian Song S., King Ryan, Richardson Curtis. Technical Report: Two Statistical Methods for the Detection of Environmental Thresholds. The Cadmus Group y Duke University Wetland Center, 2001, p. 4 y 5.

Software

- Brian Ripley. (2010). tree: Classification and regression trees. R package versión 1.0-28. http://CRAN.R-project.org/package=tree

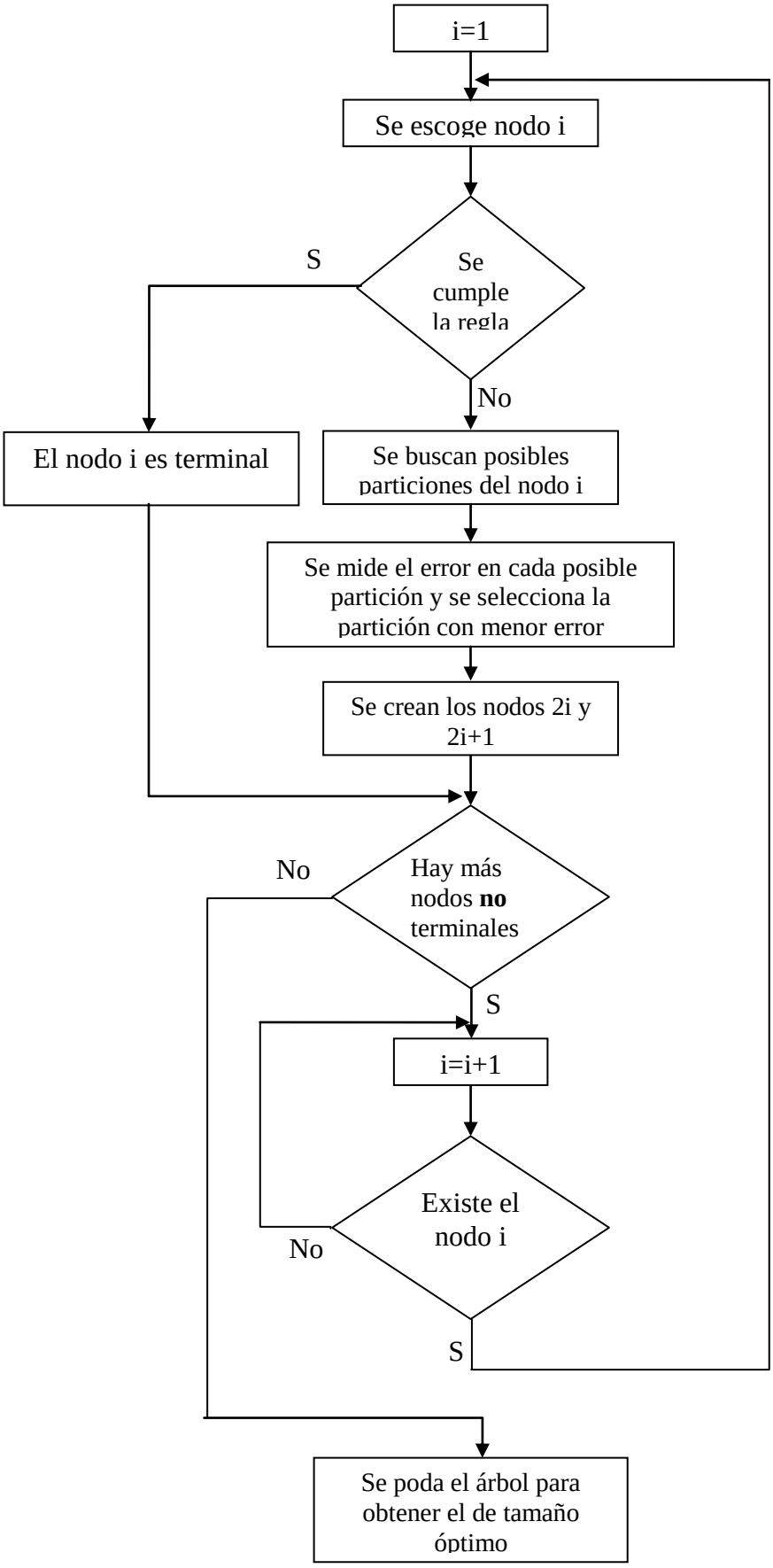
- Microsoft Corporation. Microsoft Excel 2007.

- R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

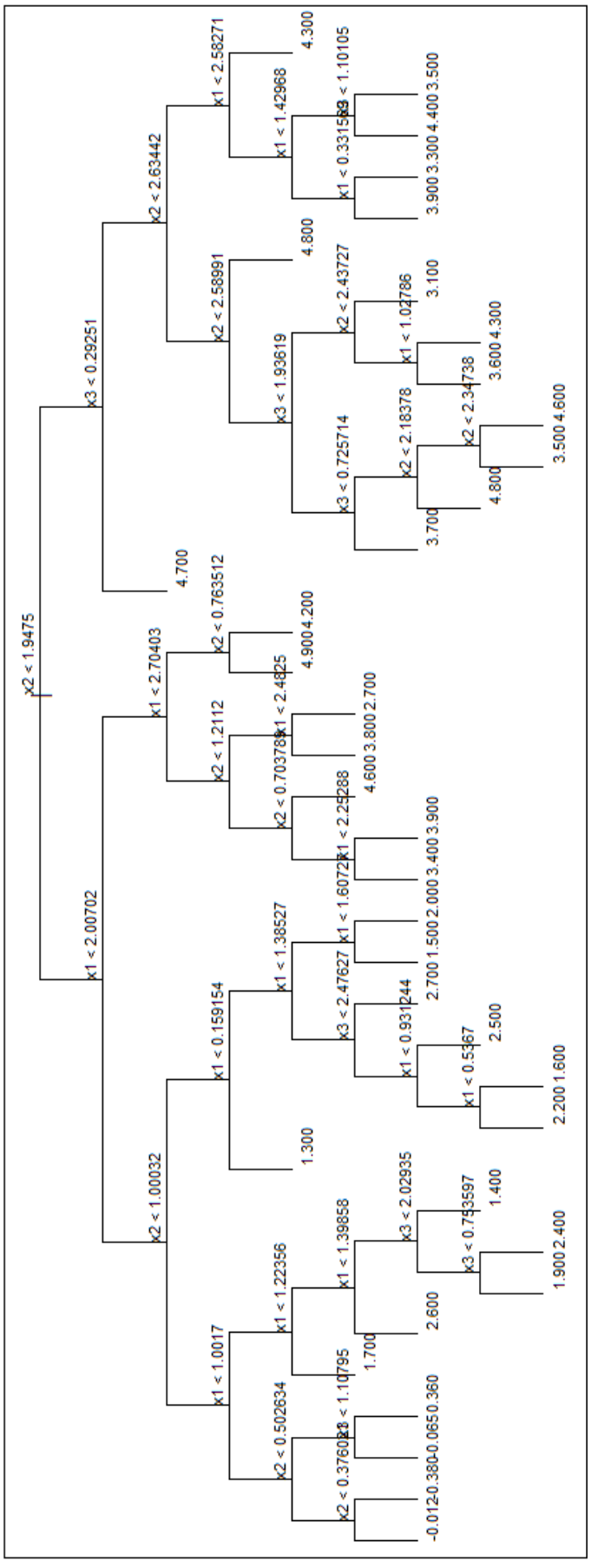
- Thomas Baier (2010). rcom: R COM Client Interface and internal COM Server. R package versión 2.2-3.1. http://CRAN.R-project.org/package=rcom

- Thomas Baier (2009). rscproxy: statconn: provides portable C-style interface to R (StatConnector). R package versión 1.3-1. http://CRAN.R-project.org/package=rscproxy

Anexo 1:
Diagrama de flujo del procedimiento



Anexo 2: Gráfica del árbol T_{max}



Satisfacción de clientes: Una aplicación del análisis CHAID.

Erandeni Juárez López
erandeni_jl@yahoo.com
Teléfono: (+52 55) 56022718

Resumen

Se presenta una breve explicación de la teoría detrás del análisis CHAID así como una aplicación.

I. Introducción

Hoy en día, mantener los clientes en un negocio o empresa es muy importante. Como negocio se espera que la calidad de servicio que se brinda a los clientes sea la mejor pero cuando se tienen demasiados, se puede disminuir. Al suceder esto, se puede provocar que los clientes dejen de requerir de los servicios de la empresa y busquen en la competencia un mejor servicio.

Cada cliente tiene necesidades diferentes y puede estar satisfecho con el servicio que se le otorga o a disgusto en algunos aspectos. La insatisfacción en los clientes siempre va a ser un indicador de que se está realizando algo erróneo, provocando que la empresa quiera saber las causas de este malestar en sus clientes.

En este trabajo se analiza la satisfacción de los clientes de una empresa distribuidora de productos de limpieza y cuidado personal que llamaremos "A" [5] cumpliendo con dos objetivos:

1. Explicar las razones por las cuales un cliente está satisfecho o no con la empresa "A" por medio de un análisis estadístico llamado

[5] Por problemas de confidencialidad no se menciona el nombre de la empresa ni cifras exactas sobre sus resultados.

CHAID (acrónimo de *Chi-squared Automatic Interaction Detector*, desarrollado por Gordon V. Kass en 1980).

2. Explicar de manera sintética los pasos para llevar a cabo el análisis de CHAID, exponiendo las distintas opciones que se presentan durante el proceso para obtener el resultado que mejor explique la realidad.

Este trabajo está organizado en tres capítulos. En el primer capítulo se da una descripción de los datos que se van a analizar. El segundo trata de la teoría del CHAID, donde se expone el filtro y las distintas reglas que se aplican durante el análisis para obtener el resultado final. En el último, se plantean los distintos resultados obtenidos al realizar el CHAID y se interpreta el elegido para explicar la satisfacción de los clientes.

Con la utilización del CHAID se pretende encontrar una forma sencilla de explicarle a la distribuidora "A" los aspectos que hacen que un cliente esté satisfecho o no con sus servicios.

II. Marco teórico de la base de datos.

La distribuidora de productos de limpieza y cuidado personal "A" es una empresa mexicana que tiene más de 10 años de existencia y más de 5 sucursales ubicadas en diferentes estados de la República Mexicana.

"A" cuenta con un amplio surtido de productos como: jabón para trastes, jabón para ropa, detergente, suavizante, shampoo, jabón de tocador, toallas sanitarias, entre otros, y busca satisfacer las necesidades de un sector de la población distribuyendo los productos principalmente a pequeños comercios y mini supers que se encuentran en las principales ciudades del país.

El objetivo principal que "A" busca es lograr una entrega oportuna y una excelente calidad en el surtido y revisión de sus productos, intentando ser el distribuidor principal de sus clientes y atraer a otros. Esto implica el contar con una amplia capacidad de surtido, almacenes y suficientes unidades de

reparto para así cubrir un importante número de comercios, brindando una buena calidad en el servicio.
La distribuidora "A" hace un levantamiento anual de información de las preferencias y características de sus clientes.

La encuesta tiene como objetivos el conocer:

- La distribución de ventas de sus clientes.
- Las necesidades del mercado, por ejemplo, el tipo de producto que mayor demanda ha tenido últimamente.
- Preferencia de los clientes respecto a la competencia de "A".
- La satisfacción de los clientes de "A" respecto a los servicios que "A" le ofrece.

Para el 2009 la encuesta fue realizada a más de 4000 comercios, con la metodología de cara a cara. Estos comercios constituyen una muestra aleatoria del total de clientes de "A", teniendo un error de estimación muestral del 2% sobre la proporción de clientes no insatisfechos, es decir, los clientes totalmente satisfechos, satisfechos y ni satisfechos ni insatisfechos.[6]
Los encargados de levantar la encuesta son los repartidores de productos de la empresa "A". Ésta es aplicada a negocios que son clientes de "A" y la persona a la que se le pide contestarla, es el encargado de la compra de productos a los distribuidores dentro del negocio.

La encuesta consta de 21 preguntas divididas en 4 partes:

1. Datos generales del entrevistado, del negocio y del entrevistador.

[6] El 2% de error se calcula por medio de la fórmula de error muestral:

$$e = \sqrt{\frac{z^2_{1-\alpha} * \hat{p}(1-\hat{p})}{n}}$$

En la que n es el tamaño de la muestra, $\hat{p}$ es el estimador de la proporción de los clientes no satisfechos y α se utiliza para determinar la probabilidad de éxito en la estimación de la proporción de los clientes no insatisfechos que se desea.

2. Distribuidores: características de la competencia de la empresa "A".

3. Clientes: necesidades, características y satisfacción de los clientes de "A".

4. Cliente final: características de las compras y preferencias de productos del mercado.

Las preguntas de la encuesta que son objeto de estudio en este trabajo provienen de la 3ra parte de la encuesta y son:

-¿Qué tan satisfecho se siente con "A" en los siguientes aspectos?

a. Todos los servicios de "A"

b. Servicio ofrecido por el vendedor: actitud del vendedor al entregar su pedido y al hacerle observaciones sobre su trabajo y/o el pedido recibido.

c. Servicio telefónico: atención que recibe al llamar para realizar su pedido.

d. Puntualidad

e. Reconocimiento por fidelidad: qué tanto "A" reconoce su antigüedad como cliente.

f. Pedidos completos

g. Facilidades de pago

h. Condición de productos recibidos

i. Publicidad de productos: publicidad obsequiada por parte de "A" para promocionar los productos en su negocio.

Las posibles respuestas para todas las preguntas (de la ***a*** a la ***i***) son:

1 = Totalmente Insatisfecho

2 = Insatisfecho

3 = Ni satisfecho, ni insatisfecho

4 = Satisfecho

5 = Totalmente Satisfecho

Los servicios o apoyos mencionados en los incisos *a* al *i* son los que la empresa "A" considera que influyen directamente en la decisión de un negocio para escogerlos como distribuidor principal y, que el fallo en alguno

de estos puntos pueden hacer que el cliente se sienta muy insatisfecho con el servicio ofrecido y dejar de ser cliente de "A".

III. Teoría del CHAID

El análisis de segmentación es una técnica estadística que permite seleccionar las variables que son relevantes en la explicación de otra variable específica, por ejemplo, qué servicios o apoyos son relevantes para los clientes para explicar la satisfacción en general con la empresa "A".

A la variable que se desea explicar se le llama la variable dependiente (Y), y las variables que explican el comportamiento de Y se les llama variables explicativas o predictoras ($X_1, \ldots, X_k$). La escala de las variables puede ser de razón o categórica ya sea nominal u ordinal. En nuestro caso la variable Y será "¿Qué tan satisfecho se siente con "A" en todos los servicios? y las variables explicativas serán el resto; siendo todas las variables que se utilizan dentro de este trabajo de tipo ordinal.

En el análisis de segmentación se tiene la finalidad de dividir el conjunto de individuos objeto de estudio, los clientes de "A", en grupos de acuerdo a las características de las variables predictoras, siendo estos grupos homogéneos y mutuamente excluyentes entre sí, para así poder predecir o explicar el comportamiento de la variable dependiente Y.

Una de las formas de encontrar las variables que mejor explican la satisfacción de los servicios en general (Y), y con las cuáles se forman los grupos, se hace por medio de un análisis llamado CHAID.

El CHAID es muy utilizado como una técnica exploratoria de datos o como una opción para analizar datos cuando éstos no cumplen con los supuestos estadísticos para realizar otros análisis. Si los datos son de tipo categórico no se basa en ninguna distribución probabilística de los datos.

El CHAID utiliza variables nominales u ordinales pero también puede utilizar variables continuas sólo que para éstas crea categorías de manera que las convierte en ordinales.

Una de las ventajas del CHAID es que la relación entre la variable dependiente y las variables predictoras se visualiza mediante la imagen de un diagrama de árbol, conocido como árbol de clasificación o decisión.

El árbol de clasificación tiene como base un nodo inicial formado por todos los datos. Después se tiene un primer criterio que divide a este nodo en dos o más grupos llamados nodos hijos (Gráfico 1).

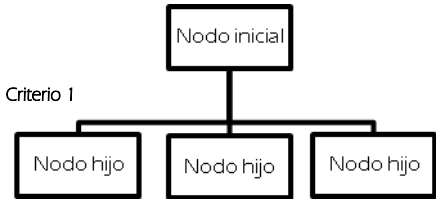


Gráfico 1

Para cada nodo hijo se busca otro criterio para dividir los datos nuevamente, por lo cual, los nodos hijos se vuelven nodos padres y de éstos salen nuevos nodos hijos creando así la estructura de un árbol (Gráfico 2).

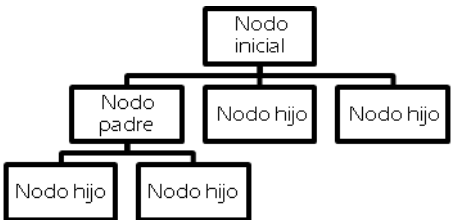


Gráfico 2

Un árbol de decisión no utiliza un modelo estadístico formal, utiliza un algoritmo para clasificar los datos mediante los valores de las variables.

Existen diferentes algoritmos para construir estos árboles, las diferencias principales entre éstos radican en las estrategias de terminación, en la toma de decisiones y en la regla adoptada de clasificar los datos, es decir, de particionar los nodos.

En el caso del CHAID se selecciona la variable que mejor explica a la variable Y, X_j, como primer criterio para dividir al nodo inicial de forma que cada nodo hijo esté compuesto por un grupo de valores homogéneos de X_j. Este proceso se repite hasta que el árbol se ha desarrollado por completo.

Para identificar las mejores variables predictoras se realizan pruebas estadísticas que a su vez dependen de la escala de Y.

Se pueden elegir distintas opciones al realizar el algoritmo de CHAID. Estas opciones incluyen: la posibilidad de elegir el criterio a cumplir para unir categorías, alfa-merge (α_m) o para separar, alfa-split (α_s), la profundidad que se desea que tenga el árbol o el número de individuos que debe de tener como mínimo cada nodo.

A continuación se describe, en forma general, los pasos del algoritmo del CHAID:

1. Para cada variable predictora X_i, se quiere disminuir el número de categorías que posee, reduciendo la complejidad de la segmentación sin tener pérdida de información. El procedimiento dependerá del tipo de escala de las variables:

   a. Variables nominales: Cada valor de la variable pronosticadora puede ser agregado a cualquier otro valor de la misma variable.

   b. Variables ordinales: Un valor de la variable sólo puede ser agregado a otro si es contiguo en la escala. Por ejemplo, se puede fusionar las categorías <<muy insatisfecho>> con <<insatisfecho>> pero no <<muy insatisfecho>> con <<totalmente satisfecho>>.

   c. Variables con valores perdidos: los valores perdidos se tratan como otra categoría <<no contesta/ no sabe>> y puede agregarse a cualquier otra categoría.

   Para la formación de grupos de categorías homogéneas los pasos son los siguientes:

   i. Se forman todos los pares posibles de categorías. El número de pares dependerá del tipo de variable que se tenga.

      Por ejemplo, en la variable de puntualidad, que presenta cinco posibles valores (1 = Totalmente Insatisfecho, 2 = Insatisfecho, 3 = Ni satisfecho, ni insatisfecho, 4 = Satisfecho, 5 = Totalmente Satisfecho) el número posible de pares serían 4, pues es una variable ordinal.

Tabla 1.

Pares posibles de la variable puntualidad
Muy insatisfecho-Insatisfecho
Insatisfecho-Ni satisfecho/ Ni insatisfecho
Ni satisfecho/ Ni insatisfecho-Satisfecho
Satisfecho-Totalmente satisfecho

ii. Para cada par posible de X_i se realizará una prueba respecto a Y para saber si ese par de categorías debe fundirse o no. La prueba usada depende de la escala de Y.

a. Si Y es continua se realiza una prueba de hipótesis llamada prueba F, donde se prueba si las medias de Y para diferentes categorías de X_i son las mismas o no[7]. La hipótesis nula es que las medias son iguales para las diferentes categorías de X_i. En esta prueba se calcula el estadístico *F* y se compara contra tablas de la distribución *F* de Snedecor, obteniendo así un valor de probabilidad llamado valor-p[8].

b. Si Y es nominal, se calcula una tabla de contingencia, formada por las categorías de X_i y las categorías de Y para llevar a cabo una prueba de hipótesis llamada chi-cuadrada de independencia[9].

[7] La prueba *F* tiene como supuestos:
- Para cada categoría de X_i, la distribución de la variable Y es normal.
- La desviación estándar de la distribución de Y es la misma para cada grupo.
- La población de cada categoría X_i, es independiente.

[8] El valor-p basado en la prueba- *F* es calculado por:

$$\rho = P(F(I-1, N-I) > F) \quad \text{donde:} \quad F = \frac{\sum_{i=1}^{I} N_i(\bar{y}_i - \bar{\bar{y}})^2/(I-1)}{\sum_{i=1}^{I}(N_i-1)s(y)_i^2/(N-I)}$$

Para mayor detalle consultar [9].

[9] La prueba chi-cuadrada tiene como supuestos:
- Los datos provienen de una muestra aleatoria.

La prueba chi-cuadrada contrasta si el par de categorías de X_i son heterogéneas respecto a Y, es decir, si son dependientes. La hipótesis nula es que las categorías de X_i son homogéneas respecto a Y.

Existen 2 estadísticos que se pueden calcular para llevar a cabo esta prueba: estadístico chi-cuadrado de Pearson X^2 o el estadístico de la razón de verosimilitud G^2, los valores de ambos son calculados por medio de la tabla de contingencia previamente mencionada.

Para cualquiera de los estadísticos que se calcule se realiza una comparación con la distribución chi-cuadrada χ^2 obteniendo así el respectivo valor-p[10].

c. Si Y es ordinal (como es el caso de nuestro estudio), al igual que cuando Y es nominal, se realiza una tabla de contingencias y se calcula el estadístico de la razón de verosimilitud H^2. La diferencia reside en que este estadístico se calcula utilizando un modelo de asociación de Y desarrollado por Goodman [6] llamado modelo de efectos por renglón[11]. De la misma manera al obtener el estadístico se realizará la prueba chi-cuadrada obteniendo así su respectivo valor-p[12].

- Las variables tienen escala de tipo nominal u ordinal.
- Las frecuencias esperadas de la tabla de contingencia sean mayores a 5.

[10] Las fórmulas correspondientes a los estadísticos y los valores-p son:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij}-N_{ij}^*)^2}{N_{ij}^*} \qquad p = P(\chi_d^2 > X^2)$$

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}N_{ij}ln\left(\frac{N_{ij}}{N_{ij}^*}\right) \qquad p = P(\chi_d^2 > G^2)$$

Donde N_{ij} es la frecuencia observada, N_{ij}^* es la frecuencia esperada y χ_d^2 sigue la distribución de una chi-cuadrada con d=(I-1)(J-1) grados de libertad. Consultar [9] para mayor detalle.

[11] El modelo de efectos por renglón tiene los supuestos siguientes:

- Los datos que se utilizan para realizar la tabla de contingencia provienen de una muestra aleatoria.
- La tabla de contingencia no tiene celdas vacías .

Estos supuestos son cumplidos por nuestros datos.

El estadístico de razón de verosimilitud se calcula: $H^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{f}_{ij}ln\left(\frac{\hat{f}_{ij}}{\hat{f}_{ij}^*}\right)$

Donde f_{ij}^* son las frecuencias esperadas bajo el modelo de independencia de columnas y renglones y f_{ij} son las frecuencias esperadas bajo el modelo de efectos por renglón. Consultar [6] y [8] para mayor detalle.

[12] El valor-p se calcula como $p = P(\chi_{I-1}^2 > H^2)$. Consultar [6] y [8] para mayor detalle.

Después de realizar alguna de las pruebas anteriores, se escoge el valor-p que sea mayor entre todos los pares posibles. El elegido se compara contra un nivel alfa preestablecido llamado α_m (alfa-merge), si éste es mayor a α_m entonces nuestro valor-p no será significativo, es decir, el del valor-p no es suficiente para rechazar la hipótesis nula y, por lo tanto, se aceptará provocando que se fusionen el par de categorías y creando una nueva categoría; si es menor nos dice que el valor-p es significativo, es decir, su valor es suficiente para poder rechazar la hipótesis nula y que las categorías no se fusionen, esto implica que son heterogéneas entre sí respecto a los valores de Y.

iii. Si se ha fusionado un determinado par de categorías, se procede a realizar nuevas fusiones de los valores de X_i, esta vez con una categoría menos pues dos categorías ya han sido fusionadas.

iv. El proceso termina cuando ya no se pueden realizar fusiones, esto es porque los p-valores de las pruebas que realizamos son menores que la α_m establecida.

v. Si el analista considera que alguna categoría contiene muy pocas observaciones para el análisis, se puede fusionar con la categoría más similar por medio de una prueba de hipótesis.

2. Después de hacer la combinación de categorías para cada variable X_i, se procede a encontrar la variable que mejor pronostica a Y, es decir, la variable X_i que por medio de sus valores puede formar grupos homogéneos dentro de si para explicar de mejor manera los valores de Y.

Para cada X_i, con sus nuevas categorías, se procede a realizar una prueba de hipótesis donde la hipótesis nula es si X_i es independiente de Y. Esta prueba se realiza con el mismo procedimiento explicado anteriormente, dependiendo del tipo de variable que sea Y. Realizada la prueba correspondiente se obtiene el valor-p de cada X_i.

La probabilidad de que el valor-p sea significativo aumenta con la realización de las pruebas de hipótesis realizadas previamente para fundir categorías. Una forma de contrarrestar este aumento es multiplicando el valor-p por el ajuste propuesto por Bonferroni [13].

3. Seleccionar la variable predictora X_i cuyo valor-p sea el menor y se compara con el nivel alfa preestablecido α_s (alfa-split).

a. Si el valor-p escogido es menor o igual que α_s entonces se rechaza la hipótesis nula, y X_i se convierte en la variable predictora de Y y se realiza la segmentación de los datos conforme al número de categorías de X_i.

b. Si el valor-p es mayor no se realiza la segmentación.

4. Si se realizó la segmentación, se procede a la ejecución de sucesivas segmentaciones para cada uno de los grupos formados por la anterior segmentación, siguiendo los mismos pasos del 1 al 4 hasta que se cumpla alguna de las condiciones del filtro de segmentación o de alguna de las reglas de parada.

---

[13] Bonferroni propone que al realizar B pruebas de significancia, cada una con significancia α_i de rechazar la hipótesis nula, la significancia total (α_T) debe de ser menor o igual que la suma de cada una de las significaciones, esto es,

$$\alpha_T \leq \sum_{i=1}^{B} \alpha_i$$

Evitando así el riesgo de rechazar inadecuadamente una hipótesis por realizar múltiples pruebas de significancia. Para aplicar esta desigualdad se debe multiplicar el valor-p final por el número posible de pruebas de significancia B realizadas, calculándose a partir del número de categorías iniciales (I) de la variable X_i y el número de categorías (r) restantes tras la fusión. El cálculo de B depende del tipo de escala de la variable.

$$B = \begin{cases} \binom{I-1}{r-1} & ordinal \\ \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!\,(r-v)!} & nominal \\ \binom{I-2}{r-2} + r\binom{I-2}{r-1} & ordinal\ con\ valores\ perdidos \end{cases}$$

El filtro y las reglas de parada son criterios que se establecen para que el análisis de segmentación tenga límites pues si no se puede dar el caso de que se produzca una gran cantidad de nodos de tamaño muy pequeño y difíciles de interpretar.

Filtro de segmentación:

Se encarga de no permitir segmentaciones que no sean estadísticamente significativas.

Este filtro puede ser aplicado en la agrupación de categorías de una variable y en la selección del mejor pronosticador.

En el proceso de agrupación de categorías se desea determinar la significación mínima (α_m) para que dos categorías de una variable queden englobadas en el mismo grupo. La α_m de 0.05 es el valor utilizado más comúnmente[14]. Si la significación (valor-p) de la prueba estadística de dos categorías de la variable independiente X_i es menor que este valor, se rechaza la hipótesis nula provocando que las 2 susodichas categorías queden separadas y se pueda proseguir con la segmentación. En cambio, si el valor es superior a 0.05 las categorías se funden, si quedan agrupadas todas las categorías de todas las variables, la segmentación se detiene.

Los valores extremos permiten comprender con mayor eficacia el efecto del filtro de segmentación. Si se escoge el mayor valor posible de la α_m entonces, la agrupación o reducción de categorías de las variables se convierte en imposible y, siempre que haya significación entre pronosticador y variable dependiente, la segmentación se formará con una determinada variable tantos grupos como categorías se tengan, tendiendo a crear una segmentación más frondosa, más amplia.

Si en vez de poner el nivel de significancia muy grande se sitúa en un valor bajo, por ejemplo .0001, en lugar de producirse más subdivisiones entre los grupos, se generarían menos divisiones entre las categorías, con el riesgo de que una determinada variable no funcione como un buen pronosticador pues todas sus categorías podrían quedar fundidas.

[14] Este valor fue mencionado por R.A. Fisher en 1925 en su libro *Statistical Methods do Reserach Workers* donde comentó que en ese nivel (.05) se tiene el tamaño de dos desviaciones estándar y que es conveniente tomar ese punto para determinar si es significante o no Este libro fue de gran impacto y empezó a ser utilizado de manera convencional. Mayor detalle consultar [15] y [16].

En el proceso de selección de variables, se tiene una forma directa de finalizar la segmentación, pues después de encontrar el pronosticador con menor significación, si no es inferior al límite establecido α_s no se tendrá ningún otro pronosticador que cumpla también con esta propiedad, por lo que el proceso de división de los datos termina.

Si α_s=1, la segmentación se produce por todas las variables existentes pero si α_s= 0.0 no se produce ni siquiera en el 1er nivel pues la significación de un pronosticador por muy pequeña que sea siempre es superior a cero.

Reglas de parada:

1. *Por tamaño:*

   Su principal objetivo es evitar que se formen grupos muy pequeños durante el proceso de segmentación.

   La regla de tamaño puede aplicarse en dos momentos: antes de la segmentación (Na, nodo padre) y después de la segmentación (Nd, nodo hijo).

   En el caso del nodo padre, la segmentación se detiene si el nodo que se quiere separar tiene un tamaño menor a Na.

   En el caso del nodo hijo, no se puede formar un grupo si no tiene un número establecido de componentes, es decir, si al crear un nuevo nodo (nodo hijo) su tamaño es menor a Nd no se crea el nodo y se funde con la categoría más similar.

2. *Por nivel:*

   Consiste en determinar un nivel (Ns, profundidad) máximo de segmentación. Por nivel se entiende cada una de las franjas horizontales del árbol.

   La primera franja horizontal corresponde al nodo principal, la segunda a la primera segmentación, la tercera a la segunda y así sucesivamente.

   Este filtro evita que se formen múltiples segmentaciones en segmentos grandes de los datos. Asimismo, contribuye a simplificar los resultados en

la medida en que reduce directamente el número de variables necesarias para predecir la variable dependiente.

3. *Por pureza:*
   - Si un nodo es puro, es decir, todos los casos del nodo tienen el mismo valor para la variable dependiente Y, el nodo no será dividido y se detendrá la segmentación.
   - Si en la segmentación de un nodo todos los nodos hijos tienen los mismos valores de Y y se ha llegado al nivel de profundidad deseado, el nodo no se divide.

Los datos que se tienen que analizar para encontrar la satisfacción de los clientes de "A" son de tipo ordinal y se desea explicar la respuesta de una variable en función de otras, es por esto que el CHAID resulta útil para este trabajo proporcionándonos además una forma

IV. Resultados

En éste capítulo se exponen algunos de los árboles CHAID que se pueden obtener aplicando distintos parámetros en las reglas de parada y el filtro de segmentación y también se obtiene el mejor resultado para explicar la satisfacción de los clientes de la empresa "A" logrando que este resultado sea sencillo de explicar y entender por el cliente.

Nuestra variable dependiente Y será:

¿Qué tan satisfecho se siente con "A" en: todos los servicios?. Siendo de tipo ordinal con los posibles siguientes valores:

1 = Totalmente Insatisfecho

2 = Insatisfecho

3 = Ni satisfecho, ni insatisfecho

4 = Satisfecho

5 = Totalmente Satisfecho

Esta variable es de orden creciente pues cada número indica que la satisfacción es mayor que la del número anterior.

Las variables con las que explicamos a Y son:

-¿Qué tan satisfecho se siente con "A" en los siguientes aspectos…?

X_1 = Servicio ofrecido por el vendedor

X_2 = Servicio telefónico

X_3 = Puntualidad

X_4 = Reconocimiento por fidelidad

X_5 = Pedidos completos

X_6 = Facilidades de pago

X_7 = Condición de productos recibidos

X_8 = Publicidad de productos

En éste caso, como le mencionamos anteriormente, nuestras variables explicativas también son ordinales y tienen los mismos valores de respuesta que la Y:

1 = Totalmente Insatisfecho

2 = Insatisfecho

3 = Ni satisfecho, ni insatisfecho

4 = Satisfecho

5 = Totalmente Satisfecho

El análisis CHAID se realizó con el paquete estadístico SPSS 16.0. En éste se indica qué tipo de variables son las que se utilizan, pues dependiendo de los tipos se realizan las pruebas de hipótesis adecuadas.

En nuestro caso como Y es variable ordinal, se calcula el estadístico de la razón de verosimilitud H^2 y se compara con el valor de la distribución chi-cuadrada para obtener el valor-p. En todos los árboles se usa el ajuste Bonferroni para las significaciones de las pruebas.

Se obtienen distintos resultados, dependiendo de las reglas de parada o filtro de segmentación que estipulamos, las diferencias en éstos pueden proporcionarnos árboles muy grandes o muy pequeños.

A continuación se presentan algunos casos para observar la diferencia que se puede obtener de resultados y finalmente se presenta la mejor solución.

Las frecuencias de la variable dependiente a explicar son:

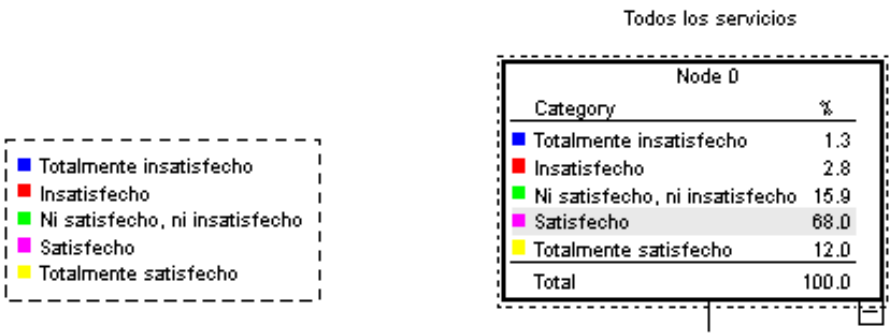


Gráfico 3

Se puede observar que el 68% de las personas están satisfechas con los servicios que le ofrece "A" y sólo el 4.1% están insatisfechos o totalmente insatisfechos con los servicios. Este bloque es muy importante pues aunque representan un muy pequeño porcentaje se puede detectar en qué servicios o apoyos son en los que "A" está fallando y qué provoca que los clientes no estén satisfechos; teniendo como fin principal especial atención hacia éstos, sin descuidar los servicios donde los clientes están satisfechos.

Cabe mencionar que los negocios a los cuales se les hace la encuesta son clientes de "A", esto puede provocar que se tenga un sesgo en las respuestas y que los clientes no hayan sido sinceros totalmente.

Resultado 1.

Para este primer resultado se escogen los siguientes parámetros:

- Alfa-merge α_m = .05
- Alfa-splitting α_s = .05
- Tamaño nodo hijo Nd = 1%
- Tamaño nodo padre Na = 2%
- Nivel de profundidad del árbol Ns = 8

El alfa-merge y alfa-splitting se definieron en un nivel convencional para comparar si un valor-p es significativo o no.

Los tamaños de Nd y Na se situaron en el 1% y 2% aproximadamente del tamaño de la muestra, con estos niveles se permite tener nodos hijos de mínimo el 1% de los clientes.

El nivel de profundidad es de 8, con este nivel estamos permitiendo que todas las variables que son significativas para explicar a Y sean utilizadas en la solución.

El resultado al fijar estos niveles corresponde a un árbol muy frondoso pues tiene una profundidad de 6 niveles, y está compuesto por 75 nodos en total.

Este resultado es muy detallado y resulta muy difícil para el cliente poder entender toda la información al mismo tiempo.

Resultado 2.

Se presenta un árbol reducido, donde los parámetros que se eligieron son:

- Alfa-merge α_m = .01
- Alfa-splitting α_s = .01
- Tamaño nodo hijo Nd = 10% encuestados
- Tamaño nodo padre Na = 20% encuestados.
- Nivel de profundidad del árbol Ns = 1

Con el alfa-merge y el alfa-splitting de .01, se es más exigente al fundir categorías y al encontrar un pronosticador. El nivel de profundidad se fija en 1 permitiendo que sólo una variable X_i sea la que explica a Y. El tamaño de

los nodos son 10% y 20%, respectivamente, del total de los encuestados, permitiendo con esto que los nodos tengan mínimo el 10% de los clientes.

Al analizar el resultado de este árbol se observa que no es posible explicar adecuadamente la satisfacción en general de los clientes pues, como se puede observar en el gráfico 4, los tres nodos hijos nos conducen a que los clientes están satisfechos o totalmente satisfechos. Esto se da pues el número de clientes que están insatisfechos y totalmente insatisfechos son menos del 10% de los encuestados y no estamos permitiendo que ningún nodo se forme con menos del 10% de los clientes.

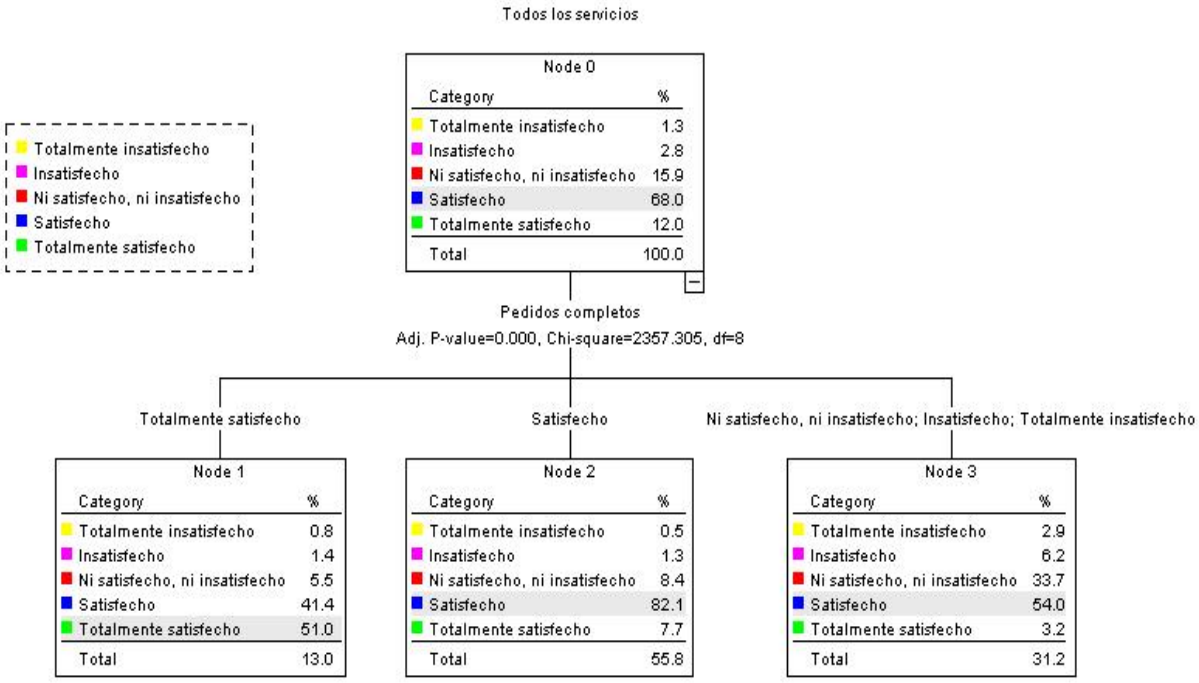


Gráfico 4. Resultado 2

Resultado final:

Para poder obtener el árbol que mejor explique los resultados, se toma en cuenta los objetivos a cumplir:

- Poder explicar el porqué de cada uno de los grados de satisfacción de los clientes.

- Que el resultado sea fácil de explicar y entender hacia el cliente.

Los parámetros que se usaron fueron los siguientes:

- Alfa-merge α_m = .01
- Alfa-splitting α_s = .01
- Tamaño nodo hijo Nd = 50
- Tamaño nodo padre Na = 100
- Nivel de profundidad del árbol Ns = 3

Las $\alpha's$ se fijaron al igual que en el resultado anterior provocando que la segmentación sea más exigente al fundir categorías y encontrar un pronosticador.

Los tamaños de los nodos (Nd, Na) se fijaron en niveles pequeños, menores al 4% de la muestra, pues el objetivo es poder explicar todos los niveles de satisfacción, y el porcentaje de clientes que dijeron estar <<Totalmente insatisfecho>> e <<Insatisfecho>> fue el 4.1%.

La profundidad del árbol la fijamos en 3 con esto, se busca que la interpretación del árbol sea sencilla de explicar al cliente limitando al resultado a sólo tener máximo tres variables explicativas para cada grupo.

El árbol final se puede apreciar en 4 partes, Gráfico 5.1 a 5.4.

En estos gráficos se puede apreciar que hay algunos nodos que tienen un signo de (+) en la parte de abajo. Éstos signos nos indican que hay más nodos hacia abajo, nodos hijos, pero éstos nos predicen el mismo valor del nodo padre que es el que posee el (+), por lo que, no se muestran. Esto se hizo pues es más fácil explicarle al cliente las características de uno de sus tipos de cliente respecto a una variable, que respecto a dos o a tres.

De acuerdo al árbol de CHAID obtenido se puede explicar qué es lo que hace que un cliente esté satisfecho o no con "A".

- Cliente Totalmente satisfecho: en el gráfico 5.1 se puede observar que hay dos tipos de clientes que reunieron las mismas características para estar muy satisfecho con "A":

- Los negocios que están <<Totalmente satisfechos>> con pedidos completos y con la publicidad de productos.
- Y los que están <<Totalmente satisfechos>> con pedidos completos, <<Satisfechos>> o <<Ni satisfecho/insatisfecho>> con la publicidad de productos y <<Totalmente satisfechos>> con el servicio ofrecido por el vendedor.

Esto nos dice que un cliente está totalmente satisfecho con "A" si los pedidos, publicidad y el servicio que recibe son muy buenos. Estos tres servicios/apoyos son los que "A" debe de mantener y no descuidar para tener clientes más satisfechos y mantener a los que ya están totalmente satisfechos.

- Cliente Satisfecho: el 68% de los clientes a los que se les aplicó la encuesta están satisfechos con "A". En el gráfico 5.2 se pueden observar las principales características que reúnen los clientes como lo son:
  - <<Satisfecho>> con los pedidos completos.
  - <<Ni satisfecho/insatisfecho>> con los pedidos completos y <<Satisfecho y totalmente satisfecho>> con el servicio ofrecido por el vendedor.

Al observar a detalle el árbol se puede notar que los clientes clasificados como satisfecho están en 6 de 12 nodos finales esto, por una parte, muestra que muchos nodos llevan a que el cliente está satisfecho, lo cual era de esperarse pues, como se mencionó, son el 68% de todos los clientes. Por otra parte, se obtuvieron 6 perfiles de clientes satisfechos pero para facilidad de comprensión del cliente sólo se escogieron 2 que reunían a la mayoría (84%) de los clientes satisfechos.

Un cliente está satisfecho con "A" si no está insatisfecho con los pedidos completos ni con el servicio que le da el vendedor. Así como en el caso de los clientes totalmente satisfechos, los servicios que más le importan al cliente final son los pedidos completos y que el vendedor tenga buena actitud hacia ellos.

• Cliente Ni satisfecho, ni insatisfecho: éstos componen el 16% del total de los clientes. En el gráfico 5.2 y el 5.3 se encuentran los nodos que explican a este tipo de cliente. Sus características son:

  - <<Ni satisfecho, ni insatisfecho>> con los pedidos completos y de <<Totalmente insatisfecho>> a <<Ni satisfecho, ni insatisfecho>> con el servicio ofrecido por el vendedor.
  - <<Insatisfecho>> con los pedidos completos, de <<Ni satisfecho, ni insatisfecho>> a <<Totalmente satisfecho>> con el servicio ofrecido por el vendedor e <<Insatisfecho y Totalmente insatisfecho>> con la puntualidad.

Estos clientes nos dicen que reciben un servicio ni bueno ni malo, esto es señal de dos cosas, por una parte, que no se está ofreciendo un buen servicio en los pedidos completos ni por parte del vendedor y, por otra parte, que a pesar de esto los clientes no están insatisfechos.

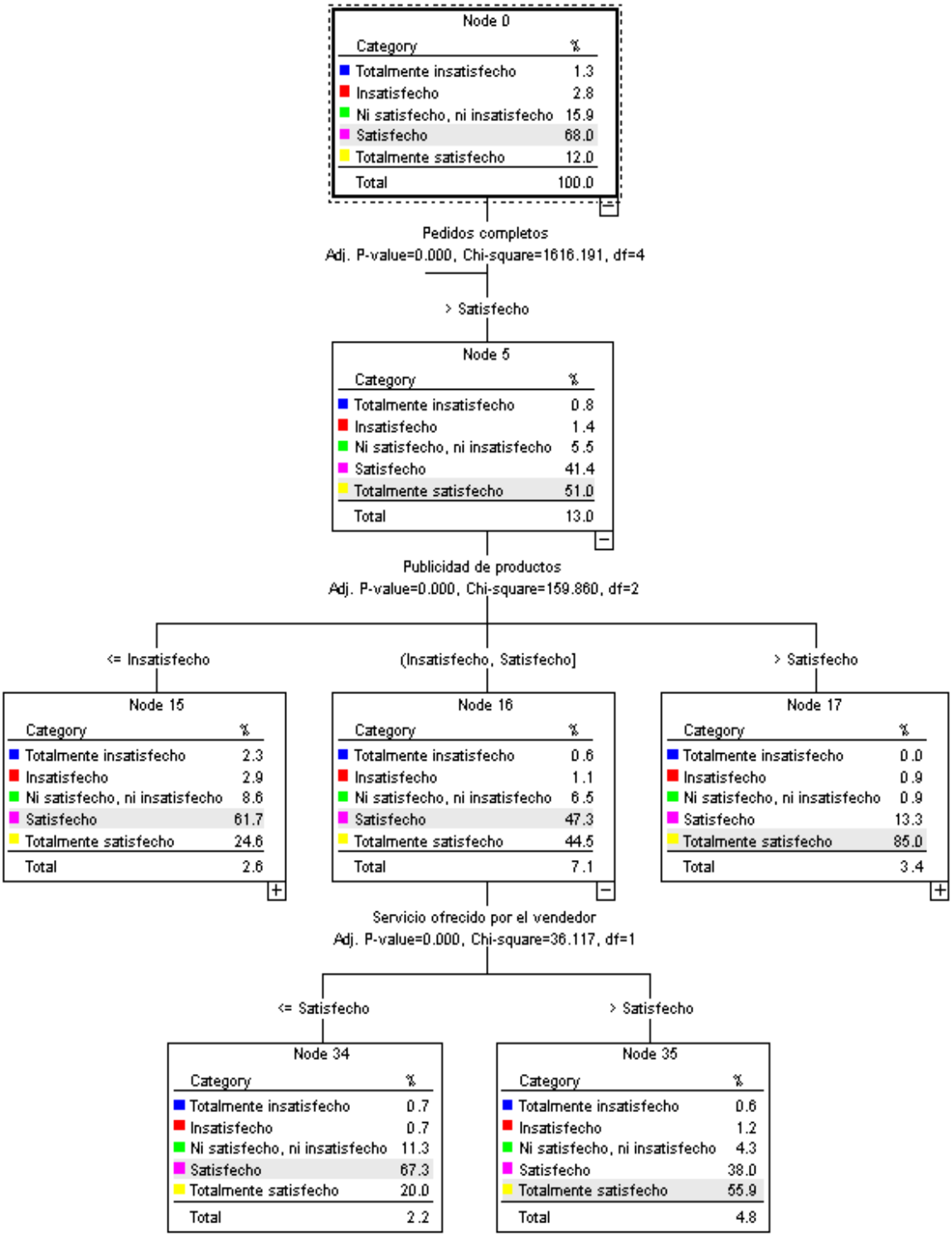


Gráfico 5.1 Parte 1

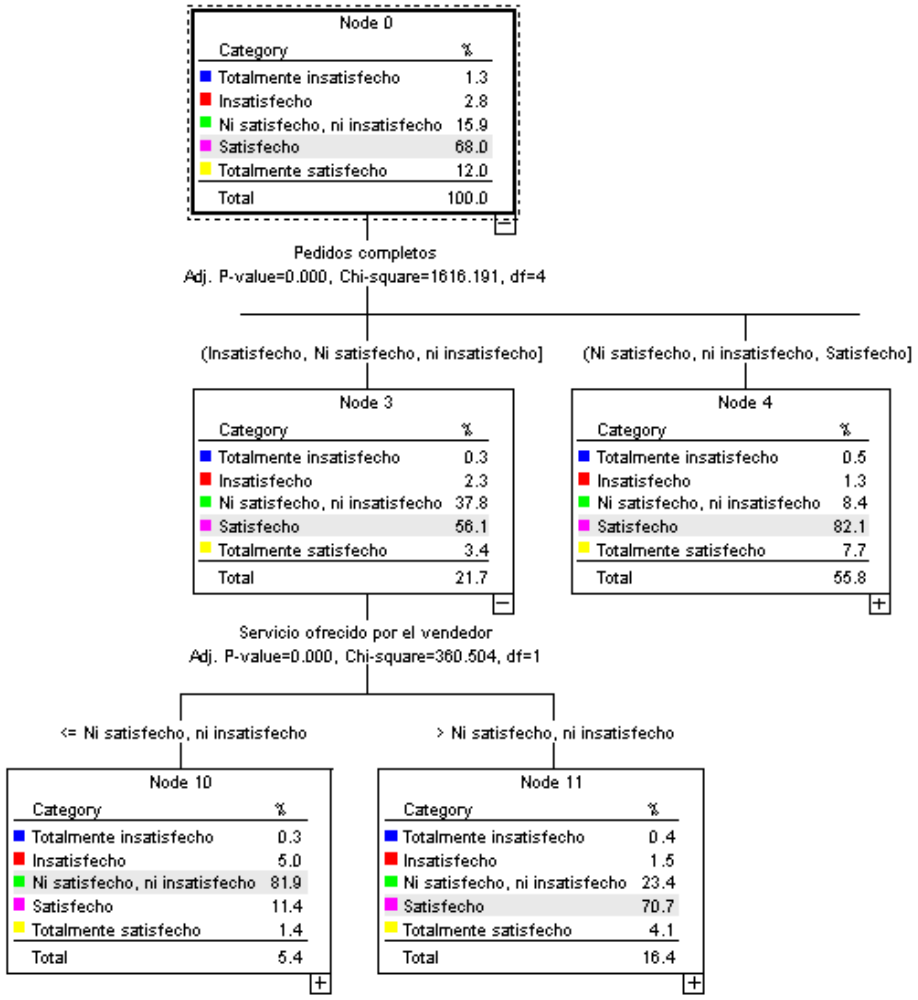


Gráfico 5.2 Parte 2

- Cliente Insatisfecho: éste tipo de cliente representa el 2.8% de todos los clientes. En el gráfico 5.3 se encuentra el nodo que los contiene. Sólo se identificó un perfil representativo para éstos:
  - <<Insatisfecho>> con los pedidos completos e <<Insatisfecho y Totalmente insatisfecho>> con los servicios ofrecidos por el vendedor.

Estos clientes están insatisfechos con "A" si están inconformes con el vendedor y sus pedidos.

- Cliente Totalmente insatisfecho: son el 1.3% de todos los clientes. Las características de éstos se encuentran en el gráfico 5.4 y son:

- <<Totalmente insatisfechos>> con los pedidos completos e <<Insatisfechos y Totalmente insatisfechos>> con las facilidades de pago.

Este es el único grupo de clientes en el que se toma en cuenta las facilidades de pago para su satisfacción, por lo que, si a un negocio no se le da facilidad de pago va a estar totalmente insatisfecho con "A".

V. Conclusiones

El CHAID tiene varias bondades, por una parte, es una herramienta para crear segmentos de los datos, siendo éstos fáciles de interpretar, cosa que no sucede con otros análisis estadísticos. Asimismo, garantiza que los grupos serán distintos y serán los mejores que se puedan encontrar y, por otra parte, con los resultados obtenidos se pueden hacer predicciones sobre datos futuros.

Al presentar los resultados en forma gráfica resulta ser más fácil de entender, siempre y cuando el resultado final no comprenda, de acuerdo a la percepción del analista, demasiados nodos por los cuáles se torne difícil la comprensión.

Ésta herramienta ha resultado ser muy útil en la práctica. En la experiencia del autor es más sencillo para las empresas entender un gráfico, el cual sigue una lógica en su interpretación, a tratar de entender la teoría de un modelo estadístico. Asimismo se tiene la ventaja de poder proporcionar el árbol creado con el algoritmo por si la empresa desea observar con más detalle cada nodo.

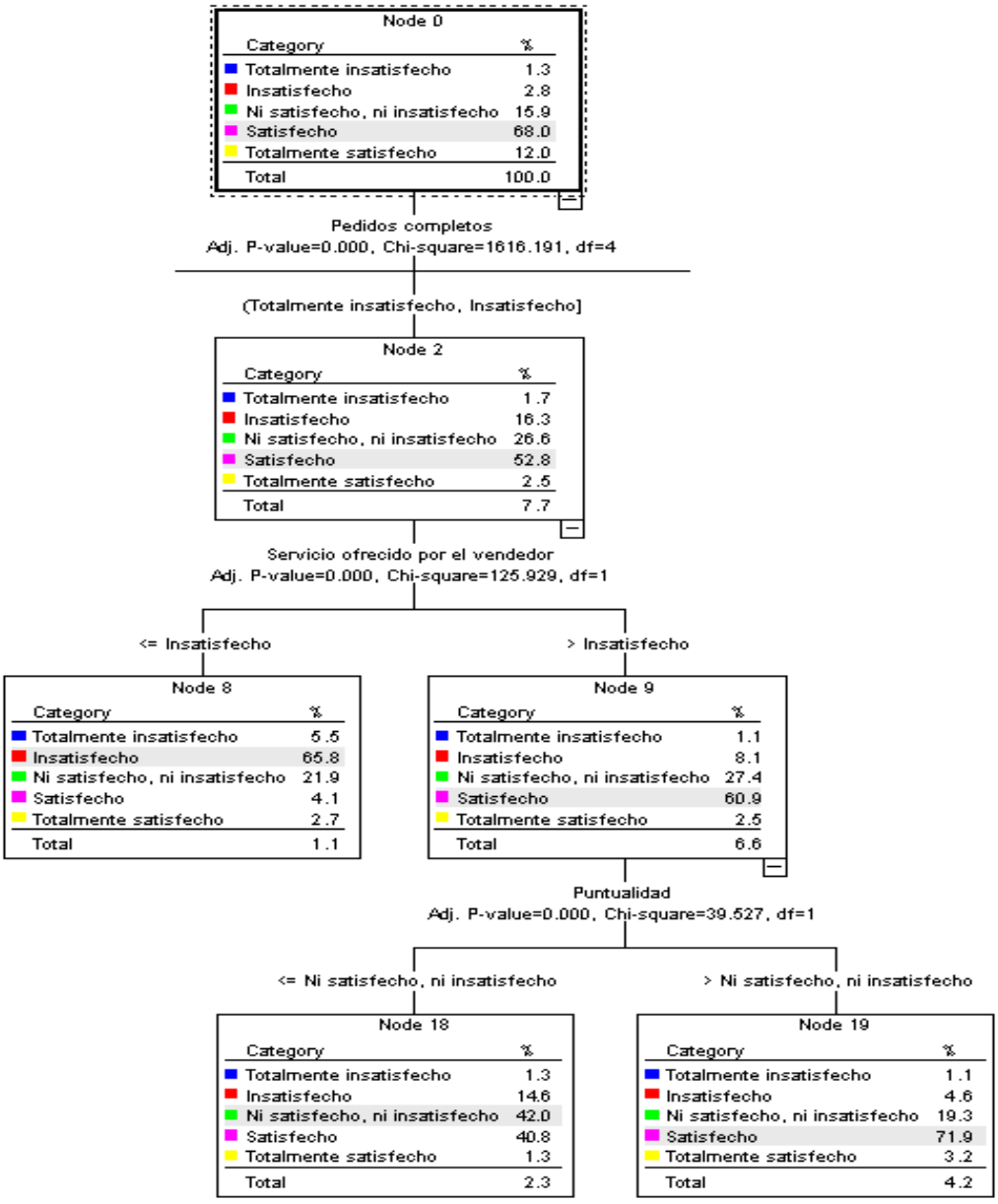


Gráfico 5.3 Parte 3

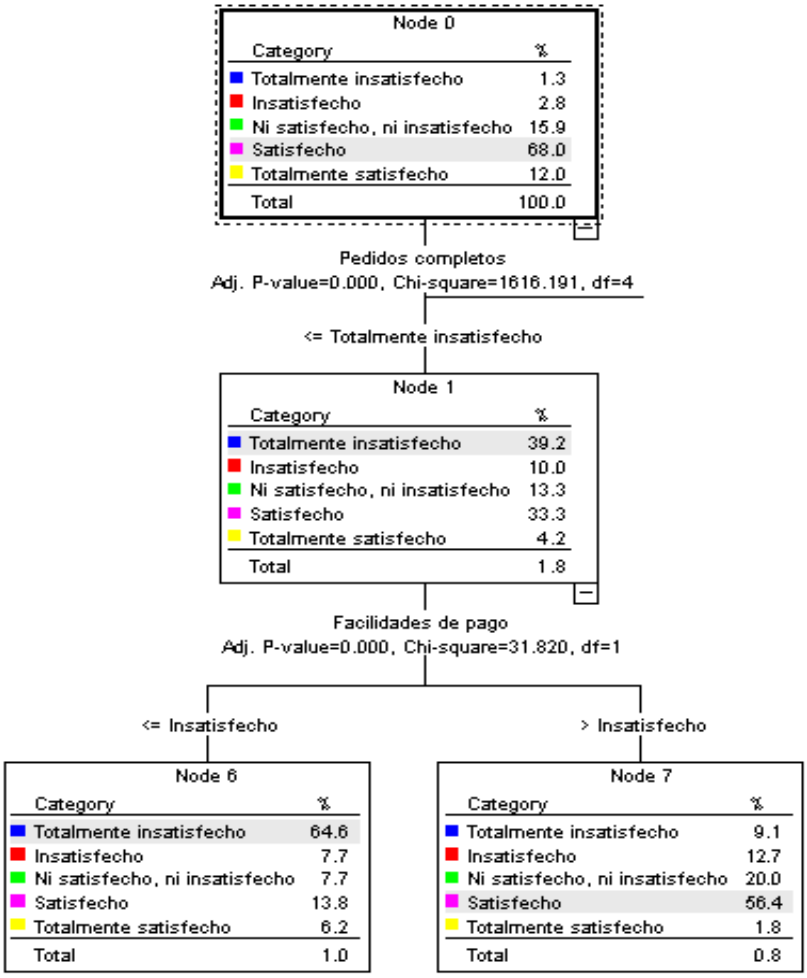


Gráfico 5.4 Parte 4

Por medio del CHAID se brinda un resumen a "A" de las características de sus clientes para que así pueda llevar a cabo acciones para mejorar el servicio ofrecido.

Como se pudo observar en las características de los distintos tipos de clientes, el punto más importante para determinar la satisfacción con los servicios en general de "A" son los pedidos completos. Esto puede parecer lógico pues como cliente de "A", se espera recibir el pedido tal cual se ordenó. Si en este punto no se está satisfecho y es común que haya faltantes entonces el negocio va a buscar otro proveedor que sí le entregue completo, provocando que "A" pierda un cliente y que afecte a futuros prospectos de clientes, es decir, que el negocio insatisfecho con "A" no lo recomiende a terceros e incluso pueda fomentar una mala imagen de ésta.

En cuanto a los clientes insatisfechos y totalmente insatisfechos, a pesar de que son pocos, es muy importante saber sus características para llevar a cabo acciones que permitan aumentar la satisfacción y para evitar que este porcentaje de clientes no aumente con el paso del tiempo.

VI. Bibliografía

1. Ángel, J,(2007), *Estadística general aplicada.* Colombia: Fondo Editorial Universidad EAFIT.

2. Arzamendi, E. O. J., (2001). *Utilización del análisis Chaid para encontrar un perfil ideal de agente de seguros,* Tesina de licenciatura en actuaría, Instituto Tecnológico Autónomo de México.

3. Bakerman, R., Robinson, B. (1994), *Understanding Log-Linear Analysis with ILOG: an interactive approach,* pp 24-25. Lawrence Eribaum Associates, Inc., Publishers.

4. Escobar, M., (1998), Metodología de las ciencias sociales N° 1, *Las aplicaciones del análisis de segmentación: El procedimiento Chaid*, 1, 13-49.

5. Hubbard,R., Bayarri, M.J. (2003), *P values are not error probabilities.* Ministry of Science and Technology of Spain.

6. Gonzalez, M. *CHAID en Investigación de Mercados, Entendiendo a los Segmentos. Millward Brown.*

7. Goodman L.A. (1979), Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, Volumen 74, 367, 537-552.

8. Iñiguez, C. A., Morales, M. G. (2009), *Selección de perfiles de clientes mediante regresión logística para muestras desproporcionadas, validación, monitoreo y aplicación en la proyección de provisiones*. Proyecto previo a la obtención del título de Ingeniero Matemático. Escuela Politécnica Nacional, Ecuador.

9. Harris, R. J. (2001), *A primer of multivariate statistics*. 3ª ed. Estados Unidos: Lawrence Eribaum Associates, Inc., Publishers, pp 13-14.

10. Hoare, R. (2004), *Using CHAID for classification problems*. Conference at the New Zealand Statistical Association. Wellington.

11. Torres, Z. J. (2004), *CHAID y regresión logística aplicados a la segmentación del mercado de tiendas tradicionales del Valle de México con respecto a la presencia de una marca de detergentes para ropa*, Tesis de licenciatura en matemáticas aplicadas, Instituto Tecnológico Autónomo de México.

12. SPSS support. *TREE-CHAID.pdf*. [En línea] http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/algorithms/. [Consulta: 05/08/2010]

13. SPSS support. *SPSS 16.0 Algorithms.pdf,* pp. 744-752. [En línea] http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows/index.html [Consulta: 23/10/2010]

14. SPSS support. *selectpred.pdf*. [En línea] http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/algorithms/. [Consulta: 23/10/2010]

15. SPSS 16.0 Help. CHAID and Exhaustive CHAID Algorithms. 2007.

16. *Why p=0.05?* http://www.jerrydallal.com/LHSP/p05.html. [Consulta: 26/11/2010]