ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ANÁLISIS DEL SESGO DE SELECCIÓN EN EL OTORGAMIENTO DE TARJETAS DE CRÉDITO

PROYECTO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE

INGENIERO MATEMÁTICO

por

Franklin Vinicio Mosquera Muñoz
fmosquera777@gmail.com

DIRIGIDO POR:

Dr. Holger Capa Santos
holger.capa@epn.edu.ec

J u n i o 2 0 1 3

Declaración

Yo Franklin Vinicio Mosquera Muñoz, declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Franklin Vinicio Mosquera Muñoz

Certificación

Certifico que el presente trabajo fue desarrollado por Franklin Vinicio Mosquera Muñoz bajo mi supervisión.

Dr. Holger Capa Santos
Director del Proyecto

Agradecimientos

A todos quienes hicieron posible este proyecto, en especial a mi familia y amigos por su incondicional apoyo.

Dedicatoria

A mis Padres, que con su dedicación y entrega, día a día han hecho posibles cada uno de mis logros.

Índice general

Resumen

El objetivo principal de este documento es estudiar el sesgo de selección existente en el proceso otorgamiento de una tarjeta de crédito en una institución financiera ecuatoriana. Lo que se quiere con el estudio es corregir el sesgo existente y, más específicamente, medir cuán significativo es en la estimación de la probabilidad de incumplimiento al compararlo frente a un modelo clásico.

Para nuestros propósitos el documento se ha estructurado de la siguiente manera: en el capítulo 1 se presenta un enfoque general del modelo propuesto por William Greene en 1998, que considera el sesgo de selección en el análisis del incumplimiento de los clientes cuando se les otorga una tarjeta de crédito; adicionalmente, en este capítulo se presenta el esquema de estimación que se dará al modelo a través de la propuesta de Heckman (1979). El capítulo 2, por otro lado, presenta a detalle la metodología usada para la construcción de un modelo clásico para la estimación de la probabilidad buscada; en este capítulo se definirán los períodos de observación de las variables explicativas, así como la manera de escoger las mejores variables para tener un mayor poder predictivo sobre la variable respuesta.

Una vez definida la metodología clásica para el tramiento del problema, en el capítulo 3, se presenta la aplicación de la nueva metodología propuesta por Greene junto con el esquema de estimación desarrollado en el capítulo 1, presentando de forma detallada el procedimiento seguido previo a la obtención de los resultados. Finalmente, en el capítulo 4, se realiza un análisis de los resultados obtenidos tanto en el capítulo 2 como en el capítulo 3; se presentan varias conclusiones y recomendaciones al respecto, en base a estos resultados, en el capítulo 5.

Cabe recalcar, que el documento cuenta con anexos como soporte a cada uno de los capítulos que lo conforman, así como demostraciones no incluidas en la bibliografía, que fueron necesarias en el desarrollo de los procedimientos de estimación y las propiedades de los modelos condicionales aquí tratados.

Notación Básica

Notaciones Generales

$E(X)$	esperanza matemática de la variable aleatoria X.
$E(X \mid Y)$	esperanza matemática de la variable aleatoria X condicionada por Y.
$Var(X)$	varianza de la variable aleatoria X.
$X \sim$ Distribución	la variable X sigue la distribución especificada después de $\sim$.
$N(\mu, \sigma^2)$	distribución normal con media μ y varianza σ^2.
$f_X(x)$	función de densidad de la variable aleatoria X.
$F_X(x)$	función de distribución de la variable aleatoria X.
$\phi(z)$	$f_X(z)$ cuando $X \sim N(0,1)$.
$\Phi(z)$	$F_X(z)$ con $X \sim N(0,1)$.

Vectores y Matrices

$\mathbf{x}$	vector $x \in \mathrm{R}^n$.
$\mathbf{A}$	matriz de $\mathrm{R}^{n \times m}$.
A'	transpuesta de $\mathbf{A}$.
$\lvert A \rvert$	determinante de la matriz A.

Capítulo 1

Análisis del Modelo de Greene

Las instituciones financieras califican a sus clientes al momento de darles u ofrecerles alguno de sus productos; este tipo de calificaciones está basado en modelos estadísticos que permiten decidir si la persona en análisis será un buen o mal cliente para la institución en el futuro. La mayoría de este tipo de modelos son muy buenos predictores y con alto poder discriminatorio; sin embargo, no toman en cuenta un problema latente que existe en dicha predicción: la construcción de la muestra.

Dado que para la construcción de cualquier modelo de este tipo se utilizan datos históricos, la construcción de la muestra presenta un problema: únicamente toma en cuenta a aquellos individuos que ya fueron seleccionados por la institución para ser beneficiarios de alguno de sus productos.

Al analizar entonces un cliente con un modelo construido de esta forma, únicamente se están tomando en cuenta algunos aspectos de comportamiento de todos los posibles, ya que la información de aquellos individuos que no fueron aceptados debido al criterio de selección impuesto por la entidad financiera, no se encuentra disponible. Este problema es el que se conoce con el nombre de *Sesgo de Selección*. Para entender mejor este problema se verá un ejemplo concreto: toda entidad financiera está interesada en evaluar la probabilidad de incumplimiento de un préstamo, una vez que la solicitud ha sido aprobada. Matemáticamente la institución está interesada en medir

$$P(\text{Incumplimiento}|\text{Aprobación})$$

Observando entonces la Figura 1 se puede entender el problema latente que existe en el proceso de selección natural de los clientes de cualquier entidad financiera:

Los clientes potenciales de la institución pueden ser cualesquiera de los individuos de la población P. De esta población $P_1 \cup P_2 \subset P$ se acercaron a la entidad para solicitar un préstamo. La entidad sometió a su proceso habitual de análisis a cada una de las solicitudes de los individuos en cuestión, de donde únicamente las solicitudes de P_1 fueron aceptadas, mientras que las de P_2 fueron rechazadas. Luego de este proceso, la institución necesita evaluar el comportamiento de sus clientes a quienes otorgó su producto financiero. Es aquí donde la entidad necesita estimar la probabilidad de incumplimiento analizando todos aquellos individuos que incumplirían sus pagos en algún momento determinado; es más, lo que se quiere es poder indentificar las características de estos individuos, para poder determinar a priori si un individuo que se acerca a la entidad será un buen o mal cliente. En consecuencia, de acuerdo a esta lógica se debería tener información de los individuos tanto de P_1 como de P_2, ya que realmente éstos fueron los individuos que se acercaron a la entidad; sin embargo, la información de los individuos de P_2 no se encuentra disponible, ya que

únicamente se tiene la información de quienes fueron aceptados como clientes; es decir, el modelo se construye en base a P_1. En consecuencia, el modelo a-posteriori produce una estimación sesgada de la probabilidad de incumplimiento.

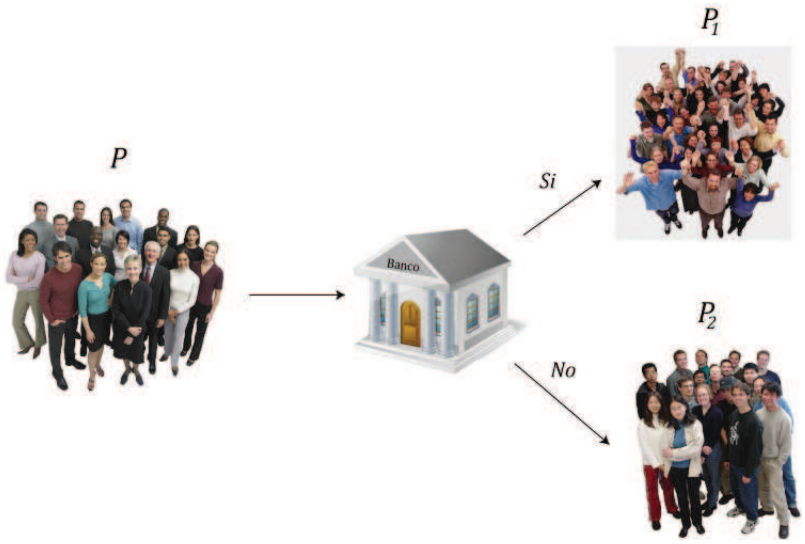

Figura 1. Ilustración del proceso de solicitud/aprobación de un crédito

Por lo tanto se puede definir formalmente al sesgo de selección como sigue:

Definición 1. *(Sesgo de Selección) Es el sesgo*[1] *presentado en el estimador de un modelo estadístico, debido al caracter no aleatorio de la construcción de la muestra.*

1.1. El Modelo de Greene

Toda institución financiera está interesada en calcular la probabilidad de que un cliente mantenga algún tipo de deuda en alguno de sus productos (Probabilidad de Incumplimiento). La esencia de la formulación de Greene (1998) descansa sobre un modelo de elección discreta con regresión latente

$$D_i^* = \beta' \mathbf{x}_i + \epsilon_i \tag{1.1}$$

D_i^*: mide la tendencia que el cliente i tiene al incumplimiento (cantidad de problemas en la que se encuentra el cliente i).
$\mathbf{x}_i$: atributos medidos sobre el cliente i que influyen sobre D_i^*.
ϵ_i: errores, presentes debido a factores no cuantificables que no se toma en cuenta en el estudio.

Utilizando la manera clásica del tratamiento de este problema, se puede definir entonces

$$D_i = 1 \text{ si } D_i^* \geq 0 \tag{1.2}$$

$$D_i = 0 \text{ sino} \tag{1.3}$$

siendo la probabilidad de interés simplemente

$$P_i = P(D_i = 1 \,|\mathbf{x}_i)$$

[1]Diferencia entre el valor esperado del estimador y el verdadero valor del parámetro (Error).

Asumiendo normalidad sobre los ϵ_i en la regresión latente ($\epsilon_i \sim N(0,1)$), se tiene

$$\begin{aligned} P(D_i = 1\,|\mathbf{x}_i) &= P(D_i^* \geq 0\,|\mathbf{x}_i) \\ &= P(\epsilon_i \geq -\beta'\mathbf{x}_i) \\ &= P(\epsilon_i \leq \beta'\mathbf{x}_i) \\ &= \Phi(\beta'\mathbf{x}_i) = \hat{p} \end{aligned}$$

Donde Φ es la función de distribución de la ley normal estándar. Desembocando así en un modelo probit clásico para el cálculo de la probabilidad de incumplimiento. Sin embargo, considerando el criterio de selección existente, la probabilidad que realmente nos interesa calcular es

$$P(D_i = 1\,|\text{Ind\_prod}_i = 1, \mathbf{x}_i)$$

Donde $\text{Ind\_prod}_i = 1$ denota la aceptación del cliente i por parte de la institución para ser beneficiario de alguno de sus productos. De modo que utilizando la misma regresión latente (1.1) se tiene

$$\begin{aligned} P(D_i = 1\,|\text{Ind\_prod}_i = 1, \mathbf{x}_i) &= P(D_i^* \geq 0\,|\text{Ind\_prod}_i = 1, \mathbf{x}_i) \\ &= P(\epsilon_i \leq \beta'\mathbf{x}_i\,|\text{Ind\_prod}_i = 1) \end{aligned}$$

En consecuencia para poder asumir que justamente

$$P(D_i = 1\,|\text{Ind\_prod}_i = 1, \mathbf{x}_i) = \hat{p} \tag{1.4}$$

se debería tener independencia entre D_i^* y $\text{Ind\_prod}_i = 1$; pero esto es poco probable, ya que el acceso a alguno de los productos de la entidad está explícitamente garantizado con algún tipo de evaluación sobre el incumplimiento del cliente y otros parámetros propios del producto ($\mathbf{z}_i$); de modo que, en general, se tiene

$$P(\text{Ind\_prod}_i = 1) = g(\mathbf{x}_i, \mathbf{z}_i)$$

Se puede incurrir en dos tipos de errores al considerar (1.4) como verdadera:

- Aún cuando $\text{P}(\text{Ind\_prod}_i = 1)$ dependa tan solo de los $\mathbf{x}_i$ y no de $\mathbf{z}_i$, y la utilización de la distribución normal sea correcta (lo cual es improbable), β no es el vector de coeficientes del modelo. Así, la estimación es sesgada con un modelo Probit.
- Si $\text{P}(\text{Ind\_prod}_i = 1)$ depende de $\mathbf{z}_i$, ésta entra en la distribución conjunta y en el cálculo de la probabilidad condicional, produciendo un sesgo en el estimador por la omisión de $\mathbf{z}_i$.

Así, la consideración de un modelo clásico[2] produce una estimación sesgada de la probabilidad de incumplimiento. De modo que en general se tiene que:

$$P(D_i = 1\,|\text{Ind\_prod}_i = 1, \mathbf{x}_i) \neq \hat{p}$$

[2]Logit o Probit, pues pese a no realizar el análisis para el modelo logit, el problema persiste, ya que la única diferencia es la función tomada para la estimación de la probabilidad (función logística).

1.2. Definición del Modelo de Greene

Se define entonces el modelo propuesto por Greene (1998) que toma en cuenta el criterio de selección en la estimación de la probabilidad de incumplimiento, y viene dado por el siguiente modelo Probit bivariante:

$$\begin{aligned}
D_i^* &= \beta'\mathbf{x}_i + \epsilon_i \text{ (Ec. del Incumplimiento)} \\
D_i &= 1 \text{ ssi } D_i^* > 0,\ D_i = 0 \text{ caso contrario} \\
\text{Ind_prod}_i{}^* &= \gamma'\mathbf{z}_i + u_i \text{ (Ec. del Producto)} \\
\text{Ind_prod}_i &= 1 \text{ ssi Ind_prod}_i{}^* > 0,\ \text{Ind_prod}_i = 0 \text{ caso contrario} \\
[\epsilon_i, u_i] &\sim N_2(0, \Sigma) \\
E(\epsilon_i \epsilon_{i'}) &= 0 \\
E(u_i u_{i'}) &= 0 \\
E(\epsilon_i u_{i'}) &= 0 \text{ para } i \neq i'
\end{aligned}$$

Donde:
D_i^*: mide la tendencia al incumplimiento del cliente i,
$\text{Ind\_prod}_i{}^*$: mide el nivel de aceptación por parte de la institución para el cliente i,
$\mathbf{x}_i$: atributos medidos sobre el cliente i que influyen sobre D_i^*,
$\mathbf{z}_i$: parámetros propios del producto que se miden sobre el cliente i en el momento de la aplicación,
D_i y $\mathbf{x}_i$ son observados si $\text{Ind\_prod}_i = 1$
Ind_prod_i y $\mathbf{z}_i$ se observan en todos los aplicantes y $\Sigma = \begin{pmatrix} 1 & \rho_{\epsilon u} \\ \rho_{\epsilon u} & 1 \end{pmatrix}$

Así

$$\begin{aligned}
P(D_i = 1 \,|\text{Ind_prod}_i = 1) &= \frac{P(D_i = 1, \text{Ind_prod}_i = 1)}{P(\text{Ind_prod}_i = 1)} \\
&= \frac{P(D_i^* > 0, \text{Ind_prod}_i{}^* > 0)}{P(\text{Ind_prod}_i{}^* > 0)} \\
&= \frac{P(\epsilon_i < \beta'\mathbf{x}_i, u_i < \gamma'\mathbf{z}_i)}{P(u_i < \gamma'\mathbf{z}_i)}
\end{aligned}$$

Siendo ésta la probabilidad de interés considerando el criterio de selección impuesto por la entidad financiera. Ahora, el siguiente paso es estimar los parámetros del modelo, cuyo procedimiento se detalla en la siguiente sección.

1.3. Estimación del Modelo de Greene

Considérese entonces el modelo de Greene dado por las ecuaciones

$$D_i^* = \beta'\mathbf{x}_i + \epsilon_i \tag{1.5}$$

$$\text{Ind\_prod}_{i'}{}^* = \gamma'\mathbf{z}_{i'} + u_{i'} \tag{1.6}$$

$$[\epsilon_i, u_i] \sim N_2(0, \Sigma) \tag{1.7}$$

Para $i \neq i'$:

$$E(\epsilon_i \epsilon_{i'}) = 0 \tag{1.8}$$

$$E(u_i u_{i'}) = 0 \tag{1.9}$$

$$E(\epsilon_i u_{i'}) = 0 \tag{1.10}$$

Donde $i = 1, \ldots, I_1$, $i' = 1, \ldots, I_2$ con I_2: número de individuos solicitantes del producto financiero, I_1: número de individuos a quienes se aprobó el producto financiero (observe que $I_2 > I_1$ debido al proceso de selección impuesto por la institución) además

$$\Sigma = \begin{pmatrix} 1 & \rho_{\epsilon u} \\ \rho_{\epsilon u} & 1 \end{pmatrix}$$

Para estimar el modelo de Greene se recurre al modelo condicional, para poder incluir en la estimación el criterio de selección impuesto por la entidad.

$$E(D_i^* \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^*) = \beta'\mathbf{x}_i + E(\epsilon_i \,|x_i\ , \text{Ind\_prod}_i{}^*)$$

donde $E(\epsilon_i \,|\mathbf{x}_i\ , \text{Ind\_prod}_i{}^*) \neq 0$ porque existen datos perdidos[3]. En consecuencia

$$E(D_i^* \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^*) = \beta'\mathbf{x}_i + E(\epsilon_i \,|\mathbf{x}_i\ , u_i > -\gamma'\mathbf{z}_i)$$

De aquí se sigue que la función de regresión depende tanto de $\mathbf{x}_i$ como de $\mathbf{z}_i$. Bajo los supuestos de normalidad y usando los resultados de la propiedad 3 i) del Anexo A con $X = \epsilon_i$, $Y = u_i$ y $z = -\gamma'\mathbf{z}_i$ se tiene entonces

$$\begin{aligned} E(\epsilon_i \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0) &= E(\epsilon_i \,|u_i > -\gamma'\mathbf{z}_i) \\ &= \rho_{\epsilon u}\frac{\phi(-\gamma'\mathbf{z}_i)}{\Phi(\gamma'\mathbf{z}_i)} \end{aligned}$$

Donde ϕ es la función de densidad de la distribución normal estándar. De igual manera, para el modelo condicional en (1.6) usando ahora la propiedad iii) del Anexo A con $X = u_i$ y $z = -\gamma'\mathbf{z}_i$:

$$\begin{aligned} E(u_i \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0) &= E(u_i \,|u_i > -\gamma'\mathbf{z}_i) \\ &= \frac{\phi(-\gamma'\mathbf{z}_i)}{\Phi(\gamma'\mathbf{z}_i)} \end{aligned}$$

Escribiendo $W_i = -\gamma'\mathbf{z}_i$ y $\lambda_i = \dfrac{\phi(W_i)}{\Phi(-W_i)}$, se tiene

$$E(\epsilon_i \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0) = \rho_{\epsilon u}\lambda_i \tag{1.11}$$

$$E(u_i \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0) = \lambda_i \tag{1.12}$$

Donde λ_i se conoce como el inverso de la razón de Mills.

Así, el modelo estadístico condicional completo bajo el supuesto de normalidad es entonces:

$$\begin{aligned} D_i^* &= E(D_i^* \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0) + V_{1i} \\ \text{Ind\_prod}_i{}^* &= E(\text{Ind\_prod}_i{}^* \,|\mathbf{z}_i,\ \text{Ind\_prod}_i{}^* > 0) + V_{2i} \end{aligned}$$

Reemplazando los resultados anteriores se tiene

$$D_i^* = \beta'\mathbf{x}_i + \rho_{\epsilon u}\lambda_i + V_{1i} \tag{1.13}$$

$$\text{Ind\_prod}_i{}^* = \gamma'\mathbf{z}_i + \lambda_i + V_{2i} \tag{1.14}$$

V_{1i} y V_{2i} son nuevos errores que recojen factores que podrían ser determinantes en la selección de la muestra; además, V_{1i} y V_{2i} cumplen con las siguientes propiedades:

[3]Información de los individuos que fueron rechazados.

Despejando V_{1i} de 1.13, V_{2i} de 1.14 y usando 1.5, 1.6, 1.11 y 1.12 se tiene:

$$\begin{aligned} E(V_{1i} \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= E(D_i^* - \beta'\mathbf{x}_i - \rho_{\epsilon u}\lambda_i \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(\beta'\mathbf{x}_i + \epsilon_i - \beta'\mathbf{x}_i - \rho_{\epsilon u}\lambda_i \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) - \rho_{\epsilon u}\lambda_i \\ &= \rho_{\epsilon u}\lambda_i - \rho_{\epsilon u}\lambda_i \\ &= 0 \end{aligned}$$

$$\begin{aligned} E(V_{2i} \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= E(\text{Ind\_prod}_i{}^* - \gamma'\mathbf{z}_i - \lambda_i \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(\gamma'\mathbf{z}_i + u_i - \gamma'\mathbf{z}_i - \lambda_i \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(u_i - \lambda_i \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(u_i \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) - \lambda_i \\ &= \lambda_i - \lambda_i \\ &= 0 \end{aligned}$$

Además si se aplica el resultado de la propiedad 3 iv) y v) del Anexo A, con $X = \epsilon_i$, $X = u_i$ respectivamente, se sigue que

$$\begin{aligned} E(V_{1i}^2 \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= E((\epsilon_i - \rho\lambda_i)^2 \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i^2 - 2\epsilon_i\rho_{\epsilon u}\lambda_i + \rho_{\epsilon u}^2\lambda_i^2 \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i^2 \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) - 2\rho_{\epsilon u}\lambda_i E(\epsilon_i \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) + \rho_{\epsilon u}^2\lambda_i^2 \\ &= (1 - \rho_{\epsilon u}^2) + \rho^2(W_i\lambda_i + 1) - \rho_{\epsilon u}^2\lambda_i^2 \\ &= (1 - \rho_{\epsilon u}^2) + \rho_{\epsilon u}^2(1 + W_i\lambda_i - \lambda_i^2) \end{aligned}$$

$$\begin{aligned} E(V_{2i}^2 \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= E((u_i - \lambda_i)^2 \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(u_i^2 - 2u_i\lambda_i + \lambda_i^2 \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) \\ &= E(u_i^2 \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) - 2\lambda_i E(u_i \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) + \lambda_i^2 \\ &= W_i\lambda_i + 1 - 2\lambda_i^2 + \lambda_i^2 \\ &= 1 + W_i\lambda_i - \lambda_i^2 \end{aligned}$$

De igual manera aplicando la propiedad 3 ii) del Anexo A, con $X = \epsilon_i$ y $Y = u_i$, y teniendo en cuenta 1.8 se tiene:

$$\begin{aligned} E(V_{1i}V_{2i} \mid u_i > -\gamma'\mathbf{z}_i) &= E((\epsilon_i - \rho_{\epsilon u_i})(u_i - \lambda_i) \mid u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i u_i - \lambda_i\epsilon_i - \rho_{\epsilon u_i}\lambda_i u_i + \rho_{\epsilon u_i}\lambda_i^2 \mid u_i > -\gamma'\mathbf{z}_i) \\ &= \rho_{\epsilon u}(W_i\lambda_i + 1) - \rho_{\epsilon u_i}\lambda_i^2 \\ &= \rho_{\epsilon u}(1 + W_i\lambda_i - \lambda_i^2) \end{aligned}$$

$$\begin{aligned} E(V_{1i}V_{2i'} \mid u_i > -\gamma'\mathbf{z}_i) &= E((\epsilon_i - \rho_{\epsilon u_i})(u_{i'} - \lambda_{i'}) \mid u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i u_{i'} - \lambda_{i'}\epsilon_i - \rho_{\epsilon u_i}\lambda_i u_{i'} + \rho_{\epsilon u_i}\lambda_i\lambda_{i'} \mid u_i > -\gamma'\mathbf{z}_i) \\ &= -\rho_{\epsilon u_i}\lambda_i\lambda_{i'} + \rho_{\epsilon u_i}\lambda_i\lambda_{i'} \\ &= 0 \end{aligned}$$

$$\begin{aligned} E(V_{1i}V_{1i'} \mid u_i > -\gamma'\mathbf{z}_i) &= E((\epsilon_i - \rho_{\epsilon u_i})(\epsilon_{i'} - \rho_{\epsilon u_{i'}}) \mid u_i > -\gamma'\mathbf{z}_i) \\ &= E(\epsilon_i\epsilon_{i'} - \lambda_{i'}\epsilon_i\rho_{\epsilon u_i} + \rho^2_{\epsilon u_i}\lambda_i\lambda_{i'} - \rho_{\epsilon u_i}\lambda_i\epsilon_{i'} \mid u_i > -\gamma'\mathbf{z}_i) \\ &= -\rho_{\epsilon u_i}\lambda_i\lambda_{i'} + \rho_{\epsilon u_i}\lambda_i\lambda_{i'} \\ &= 0 \end{aligned}$$

$$\begin{aligned} E(V_{2i}V_{2i'} \mid u_i > -\gamma'\mathbf{z}_i) &= E((u_i - \lambda_i)(u_{i'} - \lambda_{i'}) \mid u_i > -\gamma'\mathbf{z}_i) \\ &= E(u_iu_{i'} - \lambda_{i'}u_i + \lambda_i\lambda_{i'} - \lambda_iu_{i'} \mid u_i > -\gamma'\mathbf{z}_i) \\ &= -\lambda_i\lambda_{i'} + \lambda_i\lambda_{i'} \\ &= 0 \end{aligned}$$

De modo que en resumen se tiene que V_{1i} y V_{2i} cumplen con:

$$\begin{aligned} E(V_{1i} \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= 0 \\ E(V_{2i} \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= 0 \\ E(V_{ji}V_{j'i'} \mid \mathbf{x}_i, \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= 0 \text{ para } i \neq i' \\ E(V^2_{1i} \mid \mathbf{x}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= (1 - \rho^2_{\epsilon u}) + \rho^2_{\epsilon u}(1 + W_i\lambda_i - \lambda^2_i) \\ E(V^2_{2i} \mid \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= 1 + W_i\lambda_i - \lambda^2_i \\ E(V_{1i}V_{2i} \mid \mathbf{x}_i, \mathbf{z}_i, \lambda_i, u_i > -\gamma'\mathbf{z}_i) &= \rho_{\epsilon u}(1 + W_i\lambda_i - \lambda^2_i) \end{aligned}$$

Del análisis anterior se sigue que el problema del sesgo de selección se reduce a la omisión de regresores importantes en 1.13; es decir, que cuando se utiliza la manera clásica de estimación de la probabilidad de incumplimiento, sin tomar en cuenta el problema latente que existe al tomar la muestra de construcción, se puede correr el riesgo de eliminar regresores influyentes en la predicción del incumplimiento de un cliente a futuro, entendiéndose a λ_i como este conjunto de regresores omitidos.

En consecuencia, si se lograra estimar λ_i, su estimación $\hat{\lambda}_i$ se podría incluir como regresor en 1.13 corrigiendo el problema del sesgo de selección. En la siguiente sección, se presenta la manera de abordar la estimación de λ_i, de modo que se pueda obtener la estimación de la probabilidad condicional que tiene lugar en la nueva formulación, $P(D_i = 1 \,|\mathbf{x}_i,\ \text{Ind\_prod}_i{}^* > 0)$.

1.4. Estimación de la Probabilidad de Incumplimiento

Una vez desarrollado el modelo condicional, se presenta el esquema de estimación para la nueva formulación propuesta por Greene. Como se vió en la sección anterior, el problema del sesgo de selección se reduce a la omisión de λ_i en la ecuación:

$$D^*_i = \beta'\mathbf{x}_i + \rho_{\epsilon u}\lambda_i + V_{1i} \tag{1.15}$$

Donde $D_i = 1$ si y solo si $D^*_i > 0$ caso contrario $D_i = 0$ ($D^*_i \leq 0$), además D_i y $\mathbf{x}_i$ son observados si $\text{Ind\_prod}_i = 1$ ($\text{Ind\_prod}^*_i > 0$).

Para estimar este regresor influyente en 1.15, se utiliza el esquema a dos pasos dado por Heckman (1979):

1. Estimar con el método de máxima verosimilitud los parámetros de la probabilidad de que $\text{Ind\_prod}_i = 1$ ($\text{Ind\_prod}_i^* > 0$) a través de un análisis probit y considerando todos los aplicantes.
2. Del estimador $\hat{\gamma}$ del paso anterior, se puede estimar W_i con $\hat{W}_i$ y por ende λ_i con $\hat{\lambda}_i = f(\hat{\gamma}, \hat{W}_i)$; todos estos estimadores son consistentes[4] ya que $\hat{\gamma}$ es estimador consistente de γ al ser de máxima verosimilitud.

Para la estimación de la probabilidad buscada: D_i puede ser vista como una variable aleatoria que toma dos valores, 1 y 0, con probalidad p_i y $1 - p_i$ respectivamente, siendo $p_i = P(D_i = 1 \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i)$. En consecuencia, D_i^* se puede definir como sigue

$$D_i^* = \begin{cases} a^2 & pi \\ 0 & 1 - p_i \\ -a^2 & 1 - p_i \end{cases}$$

con $a \neq 0$; tomando ahora esperanzas a ambos lados de 1.15 y considerando que $E(V_{1i} \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i) = 0$, se tiene

$$\begin{aligned} E(D_i^* \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i) &= \beta'\mathbf{x}_i + \rho_{\epsilon u}\lambda_i + E(V_{1i} \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i) \\ &= \beta'\mathbf{x}_i + \rho_{\epsilon u}\lambda_i \end{aligned} \tag{1.16}$$

Por otro lado, D_i^* es una variable aleatoria, de modo que

$$\begin{aligned} E(D_i^* \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i) &= a^2 p_i - a^2(1 - p_i) \\ &= a^2(2p_i - 1) \end{aligned} \tag{1.17}$$

Igualando 1.17 con 1.16, se sigue que

$$p_i = \frac{1}{2} + \frac{1}{2a^2}\beta'\mathbf{x}_i + \frac{\rho_{\epsilon u}}{2a^2}\lambda_i$$

Haciendo $\tilde{\beta} = \frac{1}{2a^2}(\beta_1 + a^2, \beta_2, \ldots, \beta_k)$ y $\tilde{\rho} = \frac{\rho_{\epsilon u}}{2a^2}$, se tiene entonces

$$p_i = \tilde{\beta}'\mathbf{x}_i + \tilde{\rho}\lambda_i \tag{1.18}$$

siendo 1.18 una expresión equivalente a 1.15. Llegando así a un modelo clásico para la estimación de la probabilidad de incumplimento, en el siguiente capítulo, se tratará la manera más adecuada de estimar los parámetros de 1.18, además se explicará porque no se debe utilizar la estimación por mínimos cuadrados ordinarios en este tipo de formulaciones.

[4]Sea $\hat{\theta}_n$ estimador de θ, se dice que $\hat{\theta}_n$ es estimador consistente de θ cuando

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

Capítulo 2

Construcción del Modelo Clásico

Este capítulo tiene como objetivo describir los procesos y la metodología relacionada al desarrollo de un modelo clásico para el cálculo de la estimación de la probabilidad de incumplimiento. La construcción del modelo en cuestión, como se verá en las secciones posteriores, será dividida en cinco fases secuenciales:

1. Selección y consistencia de la muestra.
2. Definición de buenos y malos clientes.
3. Definición de variables explicativas.
4. Construcción del modelo.
5. Pruebas de validación.

En la fase de selección y consistencia de la muestra, se procede a seleccionar los clientes con los cuales se construirá y validará el modelo; en esta fase se definirá una tasa de incumplimiento mensual histórica, que permitirá tomar una muestra de clientes representativa que tenga la suficiente madurez (clientes con un tiempo prudente en la institución) y con un número de clientes representativo acorde a la población de la institución financiera. Luego, en la definición de buenos y malos clientes, se encontrará el criterio más apropiado para poder saber cuándo un cliente es bueno o malo para la institución financiera, de modo que pueda ser definida la marca de cliente bueno-malo, una de las variables más importantes del modelo. Con la muestra de clientes junto con la marca bueno-malo, se procederá a seleccionar el conjunto de variables más adecuadas para poder construir y validar el modelo, obteniendo así la estimación de la probabilidad de incumplimiento buscada.

Cabe recalcar que la información para la contrucción del modelo fue proporcionada por la institución financiera que apoyó el proyecto.

2.1. Selección y Consistencia de la Muestra

El objetivo en este apartado es definir la muestra de clientes con la que se trabajará en las secciones posteriores; en otras palabras, se quiere determinar adecuadamente el período de observación donde se muestra el comportamiento real de pago de los clientes.

Existen dos complicaciones al momento de seleccionar la muestra: el primero es que la muestra debe ser representativa de todos aquellos potenciales clientes. El segundo, es que la muestra debe incorporar información suficiente sobre los diferentes tipos de comportamiento de pago, de modo que sea posible identificar características que se reflejen a futuro en los nuevos postulantes de tarjeta de crédito.

El problema emerge ante la necesidad de que la muestra sea representativa de los potenciales clientes, lo que hace preferir seleccionar grupos recientes. No obstante, si se pretende poder diferenciar entre buenos y malos clientes se necesita contar con un periodo razonable de historia de pago de los mismos, lo que implica mayor tiempo desde la otorgación.

2.1.1. Determinación del Período de Observación. Se desea entonces determinar la fecha incial y la fecha final del período de observación, de modo que sean las más adecuadas; es decir, que este período de tiempo permita obtener clientes con suficiente tiempo en la institución (madurez) y que a la vez contenga clientes representativos. En consecuencia la fecha inicial deberá ser tomada lo más cercana a la fecha de desarrollo[1].

Para cumplir con este propósito se contruye la tasa de incumplimiento, la que recoge en porcentaje, la razón de clientes, que desde el mes de apertura de su tarjeta hasta enero 2013 (fecha de desarrollo) deterioraron su comportamiento de pago (clientes malos)[2], respecto de todos los clientes que obtuvieron su tarjeta en el mismo mes[3]; esta tasa se define como sigue:

(2.1)

$$\text{Tasa de Incumplimiento}_i = \frac{\text{Número de clientes malos desembolsados en el mes i}}{\text{Número total de clientes desembolsados en el mes i}} \cdot 100$$

Con el objetivo de que la muestra sea lo más representativa a la composición actual de cartera de tarjeta de crédito, y que refleje el comportamiento de pago de los posibles clientes a futuro, se realiza un análisis anual histórico de la composición en porcentaje de participación que ha tenido la institución en las diferentes gamas o tipos de tarjetas[4] (alta, media y baja) a lo largo del tiempo. En la figura 1 se puede ver el resultado de este análisis, donde claramente se observa que a partir del año 2011 la participación de la institución en las diferentes gamas cambia radicalmente; la institución comienza a tener mayor participación en gama alta y media, dejando de lado a la gama baja, la cual había sido muy representitiva en años anteriores. En consecuencia, la fecha de referencia a la cual se calcula la tasa de incumplimiento es Mayo 2011, fecha a la cual los directivos de la institución deciden aplicar esta medida.

Se calcula entonces la tasa de incumplimiento para cada mes de colocación a partir de mayo 2011 (fecha a partir de la cual se asegura la representatividad de la muestra); cabe recalcar que para el cálculo de esta tasa, se definió una nueva variable denominada *atraso*, que es función de la mora, pre-mora, el estado de la tarjeta y los pagos vencidos, variables usadas por la institución para calificar el comportamiento de pago de sus tarjetahabientes, las cuales al estar medidas en diferentes unidades de tiempo (mora y pre-mora en días, estado de la tarjeta en código y pagos vencidos

[1]La fecha de desarrollo es la fecha a la cual se decide construir el modelo; para este caso es enero 2013.

[2]Para definir si un cliente deterioró o no su comportamiento de pago se utilizará el criterio experto de acuerdo a la experiencia del departamento de cobranza de la institución y la construcción de una nueva variable *atraso*.

[3]A este tipo de análisis se lo conoce como Análisis de Cosecha.

[4]La clasificación en gamas se realiza de acuerdo a la marca de tarjeta: alta (black, signature y platinium), media (gold) y baja (clásica).

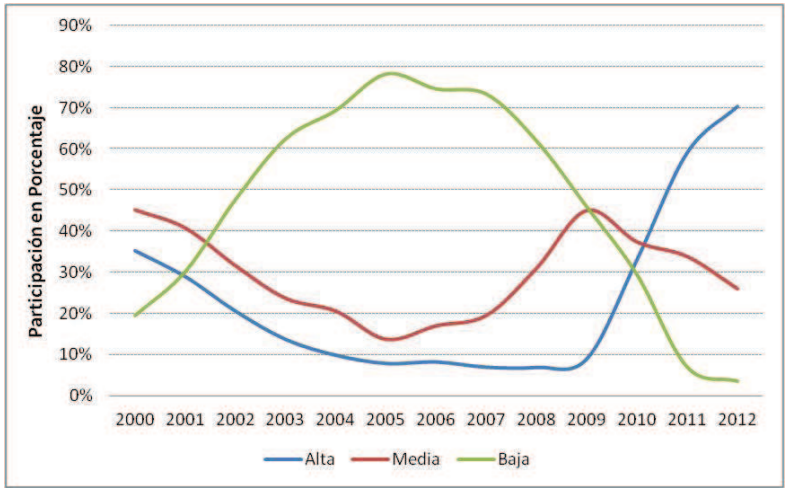

Figura 1. Composición histórica anual de la cartera de tarjeta de crédito.

en número), debían de ser condensadas en una sola definición (ver Anexo D.1). Esta nueva variable se calculó mensualmente por cliente, desde la fecha de apertura de la tarjeta hasta la fecha de desarrollo (enero 2013); de estas nuevas variables se tomó el valor máximo y se comparó si era o no mayor a $t = 30$ días[5], de modo que se pueda determinar si un cliente ha deteriorado o no su comportamiento de pago (numerador de la tasa de incumplimiento); luego, al clasificar a todos clientes de acuerdo al mes de apertura de la tarjeta se puede obtener el número de clientes desembolsados mensualmente (denominador de la tasa de incumplimiento).

El resultado de este cálculo se presenta en la figura 2. Como se puede observar, mientras más cercano se esté a la fecha de desarrollo la tasa de incumplimiento tiende a cero; lo que es consecuencia de que para la mayoría de clientes en estos meses no se posee información suficiente en cuanto a su comportamiento de pago, para poder determinar la condición de deterioro (atraso máximo > 30 días), mientras que en meses más lejanos de la fecha de desarrollo se tiene clientes con mayor información; de modo que, la tasa de incumpliento mantiene niveles similares. A la serie temporal generada por el cálculo de la tasa de incumplimiento en los distintos meses de colocación se la somete a un análisis de cambio estructural, de modo que se pueda conocer la fecha en la cual la tasa comienza a caer hacia cero; es decir, conocer hasta qué tiempo se tiene información suficiente para poder determinar si un cliente deteriora o no su comportamiento de pago, fecha a partir de cual se definirá el tiempo de madurez de un cliente en la institución.

2.1.1.1. Tiempo de Madurez. Para determinar el tiempo de madurez, o el tiempo en el cual los clientes muestran su comportamiento real de pago, se utilizará la prueba de cambio estructural CUSUM, herramienta estadísitica que se describe en el Anexo D.2 y una de cuyas aplicaciones consiste en técnicas diseñadas para mostrar posibles desviaciones o cambios de estructura en una regresión lineal; la regresión con

[5]Se toma como referencia $t = 30$ días, siendo éste el número de días a partir del cual la gestión de cobranzas se intensifica, ya que es el punto en el cual acorde a la experiencia en el negocio, el cliente comienza a caer en un proceso de no recuperación.

la que se trabajará para este caso es la tasa de incumplimiento en función del tiempo.

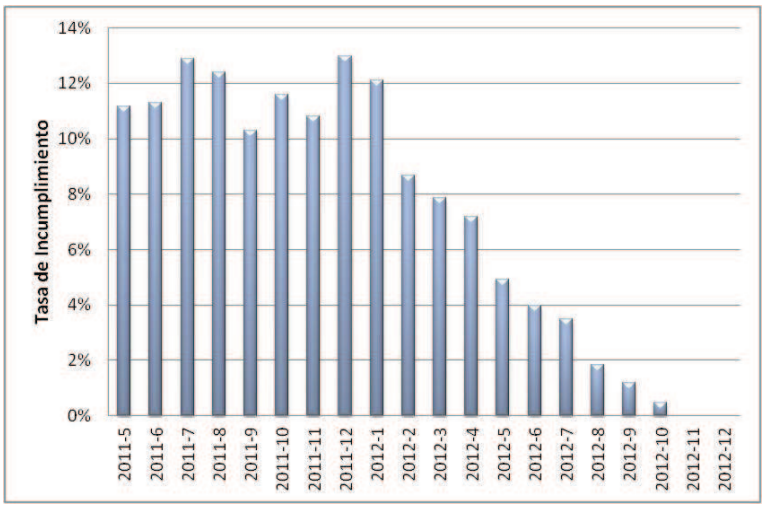

Figura 2. Tasa de incumplimiento

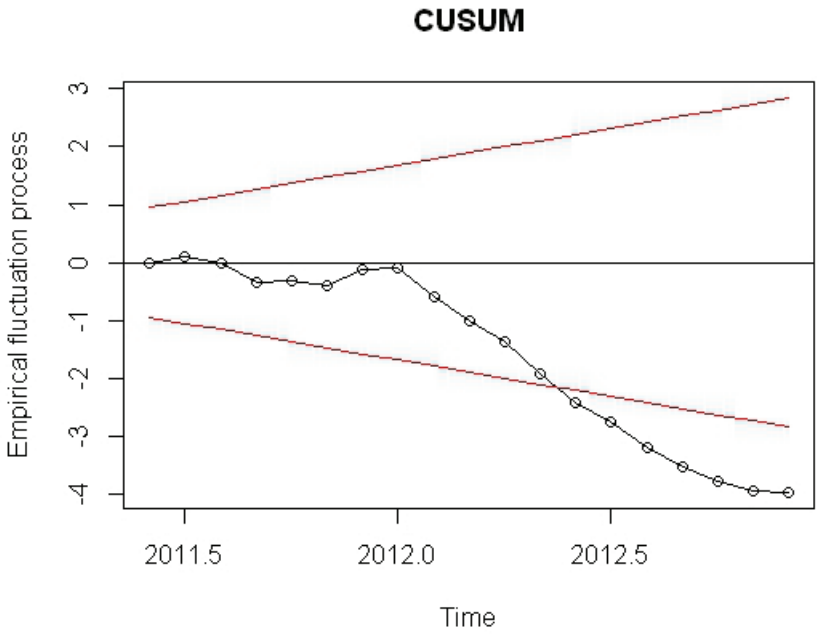

Figura 3. Resultado de la prueba CUSUM.

En la figura 3, se muestra la salida gráfica del software R para la prueba CUSUM. Se puede ver que un cambio estructural tiene lugar a partir de enero 2012, ya que se observa en el gráfico del proceso $W_n(t)$ (ver Anexo D.2), asociado al modelo: Tasa de Incumplimiento$_i$ $= \beta_1 + \beta_2 \mathrm{t}_i + u_i$ $(i = 1, \ldots, 20)$, una clara desviación de su media, desviación que se mantiene a partir de este instante hasta pasar totalmente las bandas de confianza; esto se puede corroborar con el valor $p = 0{,}001048$, producto

del contraste de las hipótesis asociado a la prueba:

$$\begin{aligned} H_0 &: \text{No existe un cambio estructural} \\ H_1 &: \text{Existe un cambio estructural} \end{aligned}$$

donde se rechaza H_0 con un nivel de significación del 5 %. De igual manera se calcula el valor p en el intervalo mayo 2011 - enero 2012, dando como resultado $p = 0{,}9019$ que permite aceptar H_0; por lo tanto, la fecha en la cual la tasa de incumplimiento comienza a caer a cero es enero 2012.

Teniendo en mente el resultado anterior y además que la fecha de desarrollo es enero 2013, se debe esperar entonces al menos 12 meses para que el cliente muestre su comportamiento real de pago; en consecuencia, el tiempo de madurez es igual a 12 meses.

Ahora, como el objetivo del modelo es predecir el comportamiento de pago futuro (variable dependiente $\mathbf{y}$) en función de variables independientes $\mathbf{X}$, se deben tomar entonces dos períodos de observaciones con magnitud igual al tiempo de madurez, para poder asegurar suficiente información, tanto para $\mathbf{X}$ como para $\mathbf{y}$. Se definen entonces los períodos:

- **Período 1** (enero 2011 a diciembre 2011): tiempo en el que se miden y observan las variables independientes denotadas por $\mathbf{X}$
- **Período 2** (enero 2012 a diciembre 2012): tiempo en el cual se mide y observa la variable dependiente denotada por $\mathbf{y}$.

Los clientes sobre los cuales se medirán tanto las variables $\mathbf{X}$ como $\mathbf{y}$, serán todos cuya tarjeta se encuentre activa a diciembre 2011, y que además tengan al menos un estado emitido tanto en el período 1 como en el período 2; con esta muestra de clientes será con la que se trabajará en las secciones posteriores, ya que cumple con las condiciones de madurez y representatividad deseadas.

2.2. Definición de Buenos y Malos Clientes

El siguiente paso de la metodología propuesta es obtener de alguna manera un criterio apropiado para establecer si un cliente es bueno o malo; esta definición es determinante en el desarrollo del modelo estadístico a obtenerse debido a que de esta categorización resulta la variable dependiente $\mathbf{y}$.

La manera de establecer estos conceptos es a través de criterios expertos dados por los ejecutivos de la entidad donde se desarrolla el modelo, o analizando cuadros de estadística descriptiva basados en variables auxiliares que relacionen el comportamiento de pago de los clientes, tales como mora máxima histórica, mora promedio, contadores de mora (reincidencia), entre otros, que indiquen cómo se clasifica al cliente al interior de una institución financiera; sin embargo, el planteamiento que se presenta utiliza la variable *atraso* definida en la sección anterior, la cual es construida mensualmente en el período de observación de la variable $\mathbf{y}$ (enero 2012 -

diciembre 2012), y el valor máximo así como el promedio de estas 12 nuevas variables es contrastado con la rentabilidad generada por el cliente (RORAC[6]).

Las figuras 4 y 5 muestran los resultados obtenidos del análisis sobre la muestra de clientes definidos en el apartado anterior. Como se puede observar, en ambas figuras se han consolidado a los clientes de acuerdo al número de atraso promedio o máximo, respectivamente, y calculado en cada caso el índice RORAC; además, se ha trazado una línea de rentabilidad esperada definida de acuerdo a juicio experto (RORAC=25 %), la cual se entiende como el límite mínimo de rentabilidad que la institución desearía obtener de todos los clientes. A partir de este criterio se pueden definir los cortes para las definiciones de buen y mal cliente, donde un buen cliente es aquel que tiene un índice de rentabilidad por encima de la esperada; por el contrario, un cliente malo sería aquel cuyo RORAC está por debajo del esperado.

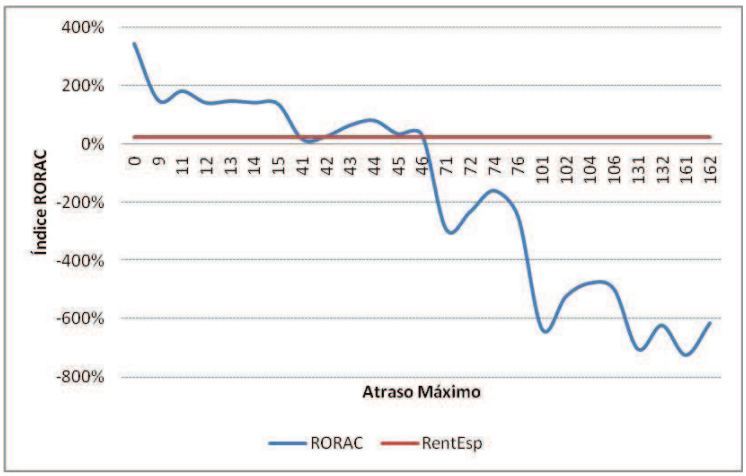

Figura 4. Atraso Máximo vs. RORAC

De acuerdo a este criterio se puede observar, en la figura 4, que un cliente bueno de acuerdo al criterio de atraso máximo sería cuando el cliente tiene un atraso máximo menor a 41 días, mientras que un cliente con un atraso máximo superior a los 46 días estaría categorizado como malo; por otro lado, analizando la figura 5 se tiene que, en función al atraso promedio, un cliente bueno sería aquel cuyo atraso promedio es menor a 11 días, mientras que el cliente malo sería aquel cuyo atraso promedio supera los 14 días. Cabe recalcar, que las secciones donde la rentabilidad fluctúa sobre la esperada, definirán los cortes para aquellos clientes catalogados como indeterminados.

Para poder unificar las definiciones de buen y mal cliente acorde a las variables atraso máximo y promedio, se cruzan estas dos variables. Este resultado se presenta

[6]El modelo de Rentabilidad por cliente, recoge todos los ingresos, comisiones y gastos que el cliente genera de sus operaciones y transacciones con los distintos canales de atención del banco, resultado de lo cual genera un estado de resultados antes de impuestos, agregando además ajustes por liquidez de acuerdo al órgano regulador y estrategias que el banco desea mantener para las captaciones y colocaciones, incorporando además el valor de la provisión específica acorde a la calificación que por operación crediticia mantiene el cliente, llegando así a determinar el índice de rentabilidad RORAC (rentabilidad ajustada al riesgo).

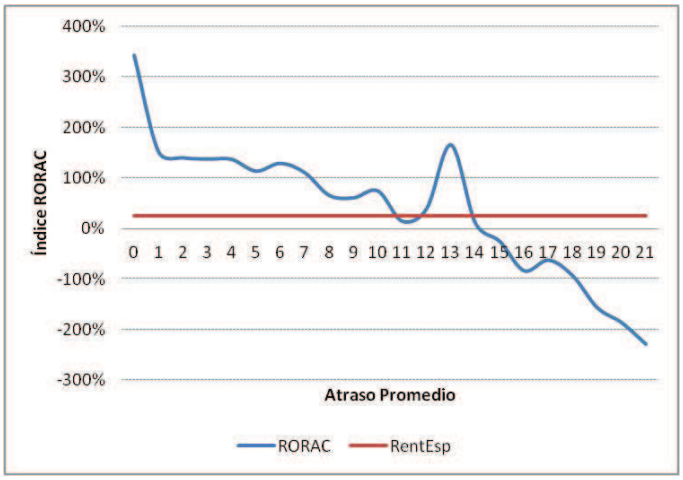

Figura 5. Atraso Promedio vs. RORAC

Atraso Máximo	Atraso Promedio					
	0-9		10-14		>14	
	Clientes	Rent.	Clientes	Rent.	Clientes	Rent.
0-35	44.183	168 %	782	171 %	79	86 %
41-46	1.960	61 %	978	34 %	728	11 %
>46	3	-354 %	218	-177 %	5.550	-506 %

Cuadro 1. Atraso Máximo vs. Atraso Promedio.

en el cuadro 1, donde se puede apreciar una matriz en la cual se dispone de distintos rangos para las variables atraso promedio y atraso máximo, producto de los cortes definidos anteriormente para ambas variables; las casillas producto del cruce de estas variables contienen el número de clientes presentes en cada categoría, así como el valor de la rentabilidad acumulada que estos clientes mantienen. La idea de cruzar estas dos variables es obtener una definición que nos permita mitigar aquellos clientes donde el atraso máximo es muy ácido o exigente, es decir tratar de rescatar aquellos clientes cuyo máximo sea alto pero mantengan un promedio bajo; sin embargo, se puede observar en el cuadro 1 que los clientes que se quisiera salvar, es decir aquellos con atraso máximo mayor a 46 y promedio menor a 9, poseen una rentabilidad negativa y a más de ello no son representativos (3 clientes).

Se han utilizado entonces distintos colores para generar un criterio para seleccionar lo que se considera buen cliente (casillas verdes), mal cliente (casillas rojas) e indeterminados (casillas amarillas). Las definiciones, como se observa, toman únicamente como referencia al atraso máximo, ya que el promedio en todos sus cortes frente al máximo no posee un ordenamiento en rentabilidad tan marcado como lo tiene el atraso máximo frente a cada uno de los cortes del atraso promedio. En consecuencia, las definiciones para clientes buenos, indeterminados y malos quedan caracterizadas de la siguiente manera:

- **Bueno:** Atraso Máximo ≤ 35.
- **Indeterminado:** $41 \leq$ Atraso Máximo ≤ 46.
- **Malo:** Atraso Máximo > 46

Cada una de las definiciones fue validada con respecto a la mora ampliada[7] promedio tomada a diciembre 2012, criterio con el cual las instituciones financieras miden el riesgo de crédito de sus clientes. El resultado de este análisis se presenta en el cuadro 2, junto con el número de clientes en cada categoría.

Definición	Mora Ampliada	Clientes
Bueno	0,09 %	45.042
Indeterminado	21,35 %	3.673
Malo	74,66 %	5.766

Cuadro 2. Mora ampliada por definición de bueno, malo e indeterminado.

Observando el cuadro 2 se tiene que se ha trabajado sobre un universo de 54.481 tarjetas, de las cuales la mayor parte son calificadas como buenas; además, al analizar el valor de la mora ampliada en cada grupo se puede ver un claro orden de acuerdo al esperado; es decir, que las tarjetas calificadas como buenas mantienen un valor en este indicador muy bajo, con respecto aquellas que se han catalogado como malas; de igual manera, la mora ampliada de las tarjetas indeterminadas se ubica en el medio de los dos valores, como se esperaba.

2.3. Definición de Variables Explicativas

Se presenta en esta sección las variables $\mathbf{X}$, medidas sobre el período 1 (enero 2011 - diciembre 2011), que serán utilizadas para la construcción del modelo en cuestión. Cada una de las variables ha sido tomada de bases de datos internas de la institución financiera y se describen en el cuadro 3. Como se puede apreciar, existen variables tanto de comportamiento de pago (atraso máximo, atraso promedio, calificación del tarjetahabiente, entre otras) así como sociodemográficas (estado civil, género y edad).

Luego de examinar la consistencia de la información en cada una de las variables a través de estadísticos descriptivos tales como valores máximos, mínimos, media y varianza, se han eliminado cierto número de observaciones que presentaban inconsistencias, quedando de un total de 54.481 registros, 52.357 registros válidos. Cabe recalcar, que aquellos registros que en las variables: saldo promedio, máximo saldo, mínimo saldo y promedio pago, mantenían valores superiores e inferiores a los percentiles 99 % y 1 %, respectivamente, fueron reemplazados por el valor del percentil correspondiente, es decir se hizo una corrección a los registros válidos para tener mayor consistencia en la información.

Según las definiciones de clientes buenos, malos e indeterminados en los 52.357 registros, se tiene 42.924 clientes buenos, 5.863 clientes malos y 3.570 indeterminados. Para la construcción del modelo no se tomarán en cuenta los clientes indeterminados

[7]La mora ampliada se define como:

$$\frac{\text{Saldo Vencido} + \text{Saldo que no devenga interés}}{\text{Saldo Total}}$$

Variables	**Descripción**
Estado Civil	Estado civil a diciembre 2011 (soltero, casado, union libre, divorciado y viudo).
Genero	Femenino o Masculino.
Edad	Edad del cliente a diciembre 2011.
PeorCalif	Peor calificación en el período 1 (A, AA, AAA, B, C, D, E, F, G o N.)
Gama	Tipo de tarjeta que posee el cliente (Alta, Media y Baja).
Antiguedad	Fecha de apertura - diciembre 2011.
AtrMax	Máximo de las variables atraso en el período 1.
AtrProm	Promedio de las variables atraso en el período 1.
Actividad	Número de estados emitidos en el período 1.
Inactividad	12 - Actividad.
SaldoProm	Promedio del saldo mantenido en el período 1.
MaxSaldo	Máximo del saldo mantenido en el período 1.
MinSaldo	Mínimo del saldo mantenido en el período 1.
NumPagos	Número de meses que el campo pago fue distinto de cero en el período 1.
PagoProm	Promedio de los pagos realizados por el cliente en el período 1.
PagoSaldo	PagoProm/SaldoProm.

Cuadro 3. Variables **X**.

debido a que no se puede determinar si son buenos o malos clientes; en consecuencia, se tiene un universo total de 48.787 registros sobre los cuales se construirá y validará el modelo. La proporción de malos a buenos en dicho universo de clientes es del 14 %; es decir, por cada 100 clientes buenos se tienen 14 malos.

Como el número de clientes buenos no es similar al número de clientes malos y con el objetivo de que los resultados que arroje el modelo no estén influenciados principalmente por el mayor número de clientes buenos, es necesario obtener una muestra de clientes buenos con número similar al de clientes malos. Por otro lado, para que los resultados obtenidos a partir de la muestra se puedan extender a la población, la muestra debe ser representativa de la misma en lo que se refiere a las variables en estudio; esto es, la distribución de las variables en la muestra debe ser aproximadamente igual a la distribución que tienen estas variables en la población. La representatividad en estadística se logra con el tipo de muestreo adecuado (método de seleccionar la muestra) y con estadísticos de prueba que corroboren aquella hipótesis.

De igual manera, no hay que olvidar que los 52.357 registros deben ser divididos en dos grupos: muestra de construcción y muestra de validación. Para la validación del modelo se guardarán 14.636 registros (aproximadamente el 30 % del universo de registros válidos); la manera de obtener la muestra de validación será a través de un muestreo aleatorio simple, teniendo de esta manera 34.151 registros válidos que servirán para la construcción del modelo.

2.3.1. Selección y Consistencia de la Muestra de Buenos. Se procede a seleccionar y validar la consistencia de la muestra de buenos, la muestra será obtenida de los clientes catalogados como buenos en la muestra de construcción; es decir, 30.092 clientes. El tamaño de la muestra será igual al número de clientes malos; es decir 4.059. Por lo tanto, se debe obtener una muestra de tamaño 4.059 de un total de 30.092 clientes.

La forma en la cual se seleccionará la muestra será a través de un muestreo aleatorio simple, al encontrarnos sobre un universo donde todos los individuos son iguales (todos son clientes buenos). Como se mencionó anteriormente, la representatividad estadística de la muestra se logra a través de un tipo de muestreo adecuado, el cual ya se determinó, y con pruebas estadísticas que permitan rechazar o no la hipótesis de que tanto la distribución de la muestra como la de la población sean semejantes. Para ello se procedió a comprobar la idoneidad de la muestra extraída por medio de dos pruebas de bondad de ajuste: Prueba de Kolmogorov-Smirnov (K-S) y Prueba Ji-cuadrado[8].

Continuas	**Valor p (K-S)**	**Discretas**	**Valor p (Ji-cuadrado)**
Edad	0,55	Gama	0,99
Antiguedad	0,53	EstadoCivil	0,98
SaldoProm	0,39	Genero	0,96
MaxSaldo	0,08	NumPagos	0,97
MinSaldo	0,88	Actividad	0,95
PagoProm	0,36	Inactividad	0,99
PagoSaldo	0,88	PeorCalif	0,90

Cuadro 4. Resultados de las pruebas de bondad de ajuste K-S y Ji-Cuadrado

Las pruebas de bondad de ajuste tienen por objetivo determinar si los datos (muestra de buenos clientes) se ajustan a una determinada distribución (población de buenos clientes). Estas pruebas se basan en una comparación entre las funciones de distribución acumulada que se observan en la muestra ordenada y la distribución propuesta, bajo la hipótesis nula de que ambas son iguales. Si esta comparación revela una diferencia suficientemente grande entre las funciones de distribución muestral y propuesta, entonces la hipótesis nula se rechaza.

Para poder validar la consistencia de la muestra, se dividió a las variables en continuas y discretas; el motivo de esta división se debe a que se utilizará la prueba K-S sobre las variables continuas y Ji-cuadrado sobre las variables discretas. Los resultados de este análisis se presentan en el cuadro 4 donde se tiene que, tanto para las variables continuas como para las variables discretas, el valor p para las pruebas K-S y Ji-cuadrado, respectivamente, son mayores al 5 %, de manera que se puede concluir que todas las variables en la muestra extraída conservan la distribución de la población al 95 % de confianza; es decir, las variables de la muestra son representativas de la población.

[8]Para mayor detalle sobre estas pruebas ver el Anexo D.3.

2.4. Construcción del Modelo

Una vez que se ha definido el conjunto de variables a utilizarse, se debe determinar el mejor modelo para poder calcular la estimación de la probabilidad de incumplimiento en función de la marca de cliente bueno y malo, y las variables previamente seleccionadas en la sección anterior. Para ello, en este apartado se verá la mejor manera de abordar este problema, comenzando primero con un modelo de regresión discriminante y desembocando en un modelo de regresión logístico, siendo este último el más adecuado para los objetivos buscados. De igual manera, se presenta una metodología basada en árboles de decisión para definir y seleccionar las mejores variables a incluir en el modelo seleccionado.

2.4.1. Modelo de Regresión Discriminante. El problema de discriminar entre los grupos de buenos y malos clientes puede abordarse de la siguiente manera: sea y_i una nueva variable que toma únicamente dos valores, 0 cuando el cliente es catalogado como *Bueno* y 1 cuando el cliente es *Malo*. El problema de discriminación es equivalente a predecir el valor de la variable y_i.

Se construye un modelo que permita prever el valor de la variable binaria y_i, en función de ciertas características $\mathbf{x}_i = (x_{1i}, \ldots, x_{(k-1)i})$. Suponiendo que se dispone de una muestra de n obervaciones del tipo $(y_i, \mathbf{x}_i)$, donde $y_i = 0$ cuando el cliente pertenece a la población de buenos y 1 cuando es parte de la población de malos. La primera formulación para el modelo es la regresión lineal:

$$y_i = \beta_1 + \beta' \mathbf{x}_i + u_i \ (i = 1, \ldots, n) \tag{2.2}$$

Donde $\beta = (\beta_2, \ldots, \beta_k)$ y $u_i \sim N(0, 1)$. Luego, con el objetivo de asociar la probabilidad de pertenencia de un individuo a algún grupo, se toma esperanzas a cada lado de la igualdad en la expresión anterior cuando $\mathbf{x} = \mathbf{x}_i$

$$E(y_i|\mathbf{x}_i) = \beta_1 + \beta' \mathbf{x}_i$$

Si se llama p_i a la probabilidad de que y_i tome el valor de 1 cuando $\mathbf{x} = \mathbf{x}_i$, es decir,

$$p_i = P(y_i = 1|\mathbf{x}_i)$$

Entonces la esperanza de y_i es

$$E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) \cdot 1 + P(y_i = 0|\mathbf{x}_i) \cdot 0 = p_i.$$

Por lo tanto,

$$p_i = \beta_1 + \beta' \mathbf{x}_i$$

Que es una expresión similar al modelo 2.2 donde la predicción $\hat{p}_i$ estima la probabilidad de que un individuo con características definidas por $\mathbf{x} = \mathbf{x}_i$ pertenezca a la población correspondiente a $y_i = 1$.

El principal inconveniente de esta formulación es que cuando p_i es estimada por mínimos cuadrados ordinarios, no hay ninguna garantía de que la predicción esté entre cero y uno; el modelo puede dar como resultado incluso probabilidades mayores que la unidad. Sin embargo, esto no es un problema insalvable para clasificar a los clientes, pero lo es si se quiere interpretar el resultado de la regla de clasificación

como una probabilidad de pertenencia a cada población. En ese sentido, el modelo logístico puede conducir a mejores resultados.

2.4.2. El Modelo Logístico. Para que el modelo proporcione directamente la probabilidad de pertenecer a algún grupo de clientes, se debe transformar la variable respuesta de alguna manera para garantizar que la predicción esté acotada entre cero y uno. Si se toma,

$$p_i = F(\beta_1 + \beta' \mathbf{x}_i) \tag{2.3}$$

de modo que F cumpla la reestricción impuesta se tendría el resultado deseado. La clase de funciones con este tipo de propiedades (crecientes y acotadas entre cero y uno), son las funciones de distribución; en consecuencia, el problema se resuelve al tomar como F cualquier función de distribución. El modelo logístico corrige la estimación de la probabilidad mencionada, usando la función logística en 2.3 de modo que,

$$p_i = \frac{1}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}$$

En la siguiente sección, se trata la estimación de los parámetros del modelo logístico. Cabe mencionar que este modelo será usado para la estimación de la probabilidad de incumplimiento buscada, ya que es el más adecuado y el que mejor se ajusta a la realidad del problema de discriminar entre clientes buenos y malos.

2.4.3. Estimación del Modelo Logístico. Como se vió en la sección anterior, el modelo logístico representa la probabilidad condicional de y_i dado $\mathbf{x}_i$ como sigue

$$p_i = \frac{1}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}$$

La variable endógena a ser estimada en el modelo es y_i. La función de distribución condicional es la función de densidad de Bernoulli:

$$f(y_i|\mathbf{x}_i, \beta_1, \beta) = \left(\frac{1}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}\right)^{y_i} \left(\frac{\exp(-\beta_1 - \beta' \mathbf{x}_i)}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}\right)^{(1-y_i)}$$

La estimación de los parámetros (β_1, β) se realizará por máxima verosimilitud. En consecuencia, calculando el logarítmo de la función de verosimilitud se tiene

$$\begin{aligned} l(y_i|\mathbf{x}_i, \beta_1, \beta) &= \sum_{i=1}^{n} \left(y_i \log\left(\frac{1}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}\right) + (1 - y_i) \log\left(\frac{\exp(-\beta_1 - \beta' \mathbf{x}_i)}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)}\right)\right) \\ &= \sum_{i=1}^{n} \left(-\log(1 + \exp(-\beta_1 - \beta' \mathbf{x}_i) - (1 - y_i)(\beta_1 + \beta' \mathbf{x}_i))\right) \end{aligned}$$

Derivando esta última expresión respecto a los parámetros e igualando a cero se obtienen los estimadores de máxima verosimilitud:

$$\frac{\partial l(y_i|\mathbf{x}_i,\beta_1,\beta)}{\partial \beta_1} = \sum_{i=1}^{n}\left(y_i - \frac{1}{1+\exp(-\beta_1-\beta'\mathbf{x}_i)}\right) = 0$$

$$\frac{\partial l(y_i|\mathbf{x}_i,\beta_1,\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\left(y_i - \frac{x_{ij}}{1+\exp(-\beta_1-\beta'\mathbf{x}_i)}\right) = 0 \quad j=2,3,\ldots,k$$

El sistema de ecuaciones es no lineal y para encontrar una solución adecuada se recurre a métodos numéricos. Se puede demostrar que la solución se obtiene resolviendo de forma iterada el sistema:

$$X'\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = X'\mathbf{W}\mathbf{Z}^{(m-1)}$$

los exponentes $(m-1)$, (m) significan que: el vector Z es calculado en el paso $m-1$ y b en el paso m. $\mathbf{X}$ es la matriz de diseño

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \ldots & x_{ik} \\ \vdots & \vdots & \ddots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

$\mathbf{W}$ es una matriz diagonal de la forma

$$\mathbf{W} = \text{diag}\{\pi_1(1-\pi_1),\ldots,\pi_n(1-\pi_n)\}$$

Z es un vector de R^k cuyas componentes son de la forma

$$z_i = \beta_1 + \beta'\mathbf{x}_i + \frac{y_i-\pi_i}{w_i}$$

$$\pi_i = \frac{\exp(b_1 + b'^{(m-1)}x_i)}{1+\exp(b_1+b'^{(m-1)}x_i)}$$

Para comenzar el proceso iterativo se debe dar un valor al vector $b^0 = (b_1, b)$ donde $b = (b_2,\ldots,b_k)$. Para este valor, se recomienda la estimación por mínimos cuadrados de los parámetros del modelo 2.2.

2.4.4. Construcción de las Variables para el Modelo. Una vez escogido el modelo a ser utilizado (modelo logístico), se procede entonces a definir las variables para la estimación de los parámetros y la obtención de la probabilidad de incumplimiento buscada.

En la sección 2.2, se definió cuando un cliente es bueno y cuando es catalogado como malo; se crea entonces la variable y_i para cada cliente, teniendo en cuenta la siguiente regla: si un cliente está catalogado como bueno $y_i = 0$, mientras que si el cliente cumple con la definición de malo $y_i = 1$.

Ahora, para definir las características que influyen en que un cliente sea bueno o malo, es decir el vector $\mathbf{x}_i$, se usarán las variables explicativas de la sección 2.3. Cada una de las variables será sometida a un análisis de dependencia con la variable y_i. En ese sentido, se utilizarán árboles de decisión[9], herramienta de minería de datos que

[9]Para mayor información sobre árboles de decisión ver [**2**].

se destaca por su sencillez y facilidad de interpretación. Dentro de la gama de algoritmos para árboles de decisión se utilizó el método CHAID. El algoritmo CHAID basa sus reglas de partición en el estadístico de prueba Ji-cuadrado, que permite en cada nodo evaluar la dependencia de la variable objetivo respecto a todas las variables incluidas en el modelo; además, el algoritmo agrupa cada variable por categorías.

La metodología que se plantea es construir nuevas variables indicadoras[10] en base a estas agrupaciones, de modo que se tenga nuevas variables donde la proporción de malos a buenos difiera de la población inicial. Cabe recalcar, que este procedimiento se realizará sobre la muestra de buenos obtenida en la sección 2.3.1 sumando los individuos catalogados como malos en la muestra de construcción, obteniendo un total de 8.118 clientes, donde la proporción de malos a buenos es igual a la unidad y sobre los cuales se estimará el modelo.

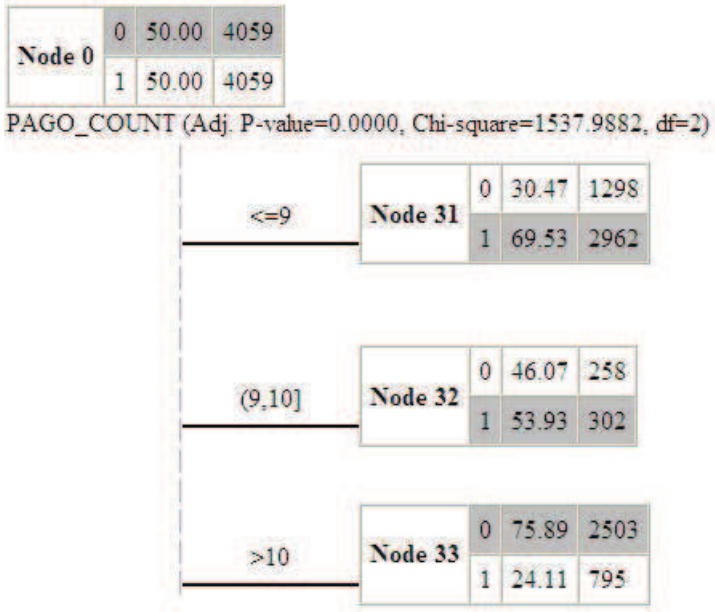

Figura 6. Arbol de decisión para la variable NumPagos.

En la figura 6 se muestra un ejemplo de la metodología para la variable *NumPagos*. Como se observa el algoritmo CHAID ha agrupado a la variable en tres categorías de clientes: aquellos que tienen menos de 9 pagos, entre 9 y 10 pagos y los que poseen más de 10 pagos. La manera en la cual se selecciona la regla para la construcción de las variables indicadoras es: tomar la condición que hace que la proporción de la población malos a buenos sea diferente de la población inicial; en este caso, en el grupo de clientes con número de pagos inferior a 9 la proporción malos a buenos es 2,28; por otro lado, los clientes que pagan entre 9 y 10 veces mantienen casi la misma proporción que la población inicial 1,17 mientras que los clientes con número de pagos superior a 10 poseen una proporción de malos a buenos de 0,32; por lo tanto, se construyen dos variables indicadoras: la primera toma el valor de 1

[10]Un variable indicador es aquella que únicamente toma dos valores 0 o 1.

cuando los clientes tienen un número de pagos inferior a 9, mientras que la segunda adquiere este valor si el número de pagos del cliente es mayor a 10, ambas toman el valor de 0, respectivamente, si las condiciones no se cumplen. Hay que mencionar además, que la variable se construye siempre y cuando el número de clientes a los cuales aplica la regla sea de al menos el 5 % de la población total[11].

La idea de construir estas variables indicadores es tener un conjunto de características que tengan relación o bien con clientes malos o clientes buenos, en el cuadro 5, se presentan los cruces producto de la aplicación del algoritmo CHAID siguiendo el procedimiento descrito anteriormente, además en este cuadro se puede ver la descripción de la regla aplicada a la variable y una columna Tipo, indicando a que grupo de clientes apunta la variable construida.

Variable	**Descripción**	**Tipo**
Ind1	Edad ≤ 38	Malo
Ind2	Edad > 49	Bueno
Ind3	PeorCalif= A,AA,AAA,B o N	Bueno
Ind4	PeorCalif= D,E,F o G	Malo
Ind5	Antigüedad≤ 330	Bueno
Ind6	Antigüedad $> 2{,}289$	Bueno
Ind7	Antigüedad entre 330 y 461	Malo
Ind8	Antigüedad entre 605 y 1369	Malo
Ind9	AtrMax≤ 12	Bueno
Ind10	AtrMax> 13	Malo
Ind11	AtrProm ≤ 1	Bueno
Ind12	AtrProm > 4	Malo
Ind13	Actividad ≤ 11	Bueno
Ind14	MinSaldo≤ 0	Bueno
Ind15	MinSaldo entre 0 y 76	Malo
Ind16	MinSaldo $> 2{,}727$	Malo
Ind17	NumPago ≤ 9	Malo
Ind18	NumPago > 10	Bueno
Ind19	PagoProm ≤ 80	Malo
Ind20	PagoProm > 80	Bueno
Ind21	PagoSaldo $\leq 9\,\%$	Malo
Ind22	PagoSaldo $> 9\,\%$	Bueno
Ind23	Baja, Edad ≤ 38	Malo
Ind24	Media, Edad ≤ 27	Malo
Ind25	Alta, Edad ≤ 30	Malo
Ind26	Alta, Edad > 27	Bueno
Ind27	Media, Edad > 38	Bueno

Cuadro 5. Variables indicadoras obtenidas a través del algoritmo CHAID.

[11]En el Anexo C, se encuentran todos los árboles de decisión usados para la construcción de las variables descritas en el cuadro 5.

2.4.5. Búsqueda del Mejor del Modelo. Una vez que se ha encontrado el conjunto de variables explicativas con mayor poder discriminatorio, es necesario estimar varios modelos y buscar el mejor, en base a criterios estadísticos. Para la selección del mejor modelo de entre las 27 variables indicadoras, construidas en la sección anterior, se usará el método de Regresión Paso a Paso (*Stepwise*).

La regresión paso a paso es un método por el cual, un solo regresor es añadido o eliminado de la regresión. El procedimiento, consiste en fijar dos niveles F (fractiles de la ley de Fisher) F_{ent}, F_{sal}. A cada paso un regresor en particular es elminado si la regresión, producto de su inclusión, produce una razón F (ver Anexo D.4) menor o igual a F_{sal}. Si ninguna variable es eliminada por este método, el regresor se introduce a condición de que la razón F, sea superior o igual a F_{ent}.

En el cuadro 6, se presentan las variables significativas luego del proceso de regresión paso a paso; hay que mencionar que el modelo fue resultado de una serie de corridas del método de manera que, los coeficientes asociados a las variables seleccionadas por el método sean significativos al 5 %, de acuerdo a la razón t[12], lo que significa que el método no necesariamente da como resultado un modelo donde todos los coeficientes son significativos con un nivel de significación fijado, ya que únicamente valida la prueba de hipótesis asociada a la razón F. El procedimiento seguido fue excluir las variables cuyo coeficiente no fue significativo y volver a utilizar el método de regresión paso a paso para elegir un nuevo modelo de entre las restantes. Este procedimiento fue realizado hasta que los coeficientes de las variables consideradas fueron significativos al 5 %.

Por otro lado, en el modelo logístico, al igual que en el modelo de regresión lineal, la dependencia lineal entre las variables explicativas (multicolinealidad) causa que los parámetros estimados sean inestables. Si ésto sucede, la varianza de los parámetros es sobreestimada y como consecuencia se puede aceptar que un coeficiente es significativo cuando en realidad no lo es. Se utiliza entonces, el índice de condicionamiento para evaluar el nivel de multicolinealidad en las variables retenidas por el modelo, el que se define como

$$IC = \sqrt{\frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}}}$$

donde $\lambda_{\text{máx}}$, y $\lambda_{\text{mín}}$, son los valores propios, máximo y mínimo respectivamente, de la matriz de correlaciones de las variables incluidas en el modelo. No hay ningún acuerdo total sobre el uso de este índice, pero se puede recurrir a la siguiente regla o criterio para tomar una decisión: si el valor del índice es mayor a 15 la multicolinealidad es fuerte, entre 10 y 15 la multicolinealidad es moderada y menor que 10 la multicolinealidad no es un problema. Para el caso presente se tiene que

[12]La razón

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \quad j = 2, 3, \ldots, k$$

a diferencia de la razón F, contrasta $H_0 : \beta_j = 0$ contra $H_1 : \beta \neq 0$. El estadístico t_j sigue asintóticamente una ley de Student con $n - k$ grados de libertad. Se rechaza H_0 cuando t_j es mayor al fractil de la ley de nivel $1 - \frac{\alpha}{2}$, siendo α el nivel de significación de la prueba.

Variable	Descripción	Coeficiente	Tipo
Ind1	Edad ≤ 38	0,25	Malo
Ind2	Edad >49	−0,23	Bueno
Ind4	PeorCalif= D,E,F o G	0,47	Malo
Ind6	Antigüedad >2289	−0,35	Bueno
Ind7	Antigüedad entre 330 y 461	0,26	Malo
Ind9	AtrMax <= 12	−0,99	Bueno
Ind11	AtrProm ≤ 1	−0,51	Bueno
Ind12	AtrProm >4	1,49	Malo
Ind14	MinSaldo ≤ 0	−0,37	Bueno
Ind16	MinSaldo>2727	0,71	Malo
Ind17	NumPago ≤ 9	0,73	Malo
Ind19	PagoProm ≤ 80	0,41	Malo
Ind21	PagoSaldo $\leq 9\,\%$	1,18	Malo
Ind26	Alta, Edad >27	−0,21	Bueno
Constante	Constante en el modelo	−0,62	

Cuadro 6. Coeficientes del modelo luego del procedimiento de regresión paso a paso.

$$IC = 5{,}29$$

de modo que se puede concluir que la multicolinealidad no es un problema en las variables que se ha retenido para el modelo en cuestión.

2.5. Pruebas de Validación sobre el Modelo Final

Una vez que se ha encontrado el mejor modelo para la estimación de la probabilidad de incumplimiento buscada, éste debe ser validado en otro conjunto de individuos de similares caraterísticas, extraído de la misma muestra. En esta sección se usará la muestra de 14.636 clientes, definida en la sección 2.3 como muestra de validación. Sobre ésta muestra se medirán los indicadores:

- Coeficiente de Gini.
- Área bajo la curva ROC.

2.5.0.1. Coeficiente de Gini. Las medidas de concentración, como es el caso del coeficiente de Gini, tratan de explicar el mayor o menor grado de igualdad en la distribución de la variable en estudio. Este coeficiente es un número entre 0 y 1, en donde 0 se corresponde con la perfecta igualdad y 1 se corresponde con la perfecta desigualdad. En ese sentido se esperaría que la distribución entre las dos categorías de la variable independiente sean lo más heterogéneas posible. Esto se traduce en un alto valor del coeficiente. En el caso del modelo, éste presenta un coeficiente de Gini de 0,76, lo cual indica que el modelo final es eficaz al momento de discriminar entre buenos y malos clientes.

2.5.0.2. Área Bajo la Curva ROC (AUROC). La curva ROC (acrónimo de Receiver Operating Characteristic) es una representación gráfica de la razón de verdaderos positivos (aciertos del modelo) frente a la razón de falsos positivos (desaciertos del modelo) según varíe el umbral de discriminación (valor a partir del cual se dice

que un caso es positivo).

El mejor punto posible de predicción se situará en un punto de la esquina superior izquierda o coordenada (0,1), indicando una clasificación perfecta. Por el contario, una clasificación totalmente aleatoria daría un punto a lo largo de la línea diagonal de no discriminación, desde el extremo izquierdo hasta la esquina superior derecha.

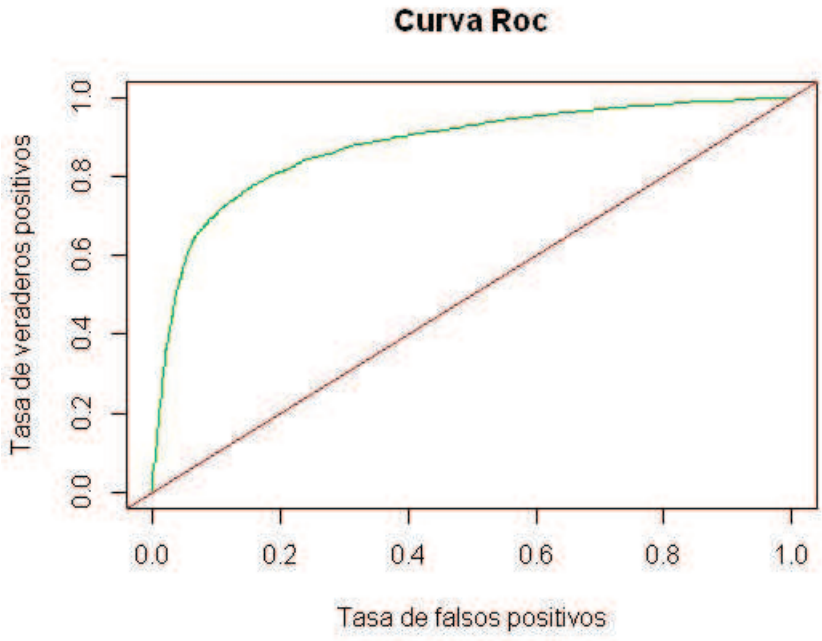

Figura 7. Curva ROC para la muestra de validación.

La curva ROC se puede usar para generar estadísticos que resumen el rendimiento de algún clasificador (en este caso el modelo logístico). Uno de estos estadísticos es el así llamado área bajo la curva ROC; este índice, para el caso del modelo final, se puede interpretar como la probabilidad que el modelo identifique a un cliente como malo de manera correcta.

El área bajo la curva es una medida conjunta de la eficiencia discriminatoria del modelo. Para el caso del modelo final seleccionado se tiene un valor AUROC de 0,88, lo que indica una capacidad discriminatoria satisfactoria; de igual manera en la figura 7 se presenta el gráfico de la curva ROC del modelo aplicado sobre la muestra de validación.

Capítulo 3

Construcción del Modelo de Greene

Este capítulo contiene a detalle la metodología utilizada para la estimación de la probabilidad de incumplimiento considerando el problema del sesgo de selección, formulación propuesta por Greene (1998). La construcción de este modelo será dividida en tres fases secuenciales:

1. Identificación del universo de análisis.
2. Identificación de las variables explicativas del modelo de Greene.
3. Estimación y validación del modelo de Greene.

En la primera fase se identificará el universo de clientes sobre el cual se construirá y validará el modelo en cuestión. En la fase siguiente, una vez definido el universo de clientes, se procede a medir sobre este grupo de clientes las variables explicativas que servirán de entrada para la estimación del modelo de Greene. Finalmente, una vez identificado el universo de clientes objetivo junto con las variables medidas sobre cada uno, se estiman los parámetros del modelo siguiendo el esquema de estimación tratado en el capítulo 1. Este esquema será divido en tres fases secuenciales, describiendo en cada una todo el procedimiento seguido hasta llegar a la estimación de la probabilidad buscada. Hay que recalcar que, el poder discriminatorio de la estimación de la probabilidad de incumplimiento proporcionada por el modelo final, será validado a través de los indicadores Gini y AUROC.

Al igual que en el capítulo anterior, la información que se utiliza para la construcción y estimación del modelo de Greene fue proporcionada por la institución financiera que apoyó el proyecto.

3.1. Identificación del Universo de Análisis

El objetivo de este apartado es identificar el universo de clientes que nos servirá para construir y validar el modelo en cuestión. Como se vió en el capítulo 1, el modelo de Greene está compuesto por las siguientes ecuaciones:

$$
\begin{aligned}
D_i^* &= \beta' \mathbf{x}_i + \epsilon_i \text{ (Ec. del Incumplimiento)} \\
D_i &= 1 \text{ ssi } D_i^* > 0,\ D_i = 0 \text{ caso contrario} \\
\text{Ind\_prod}_i{}^* &= \gamma' \mathbf{z}_i + u_i \text{ (Ec. del Producto)} \\
\text{Ind\_prod}_i &= 1 \text{ ssi Ind\_prod}_i{}^* > 0,\ \text{Ind\_prod}_i = 0 \text{ caso contrario}
\end{aligned} \tag{3.1}
$$

Donde:
D_i^*: mide la tendencia al incumplimiento del cliente i,
Ind_prod$_i{}^*$: mide el nivel de aceptación por parte de la instutición para el cliente i,
$\mathbf{x}_i$: atributos medidos sobre el cliente i que influyen sobre D_i^*,
$\mathbf{z}_i$: parámetros propios del producto que se miden sobre el cliente i en el momento de la aplicación,

D_i es medida siempre y cuando Ind_p$_i = 1$

En consecuencia, se necesita identificar clientes que se hayan acercado a la institución financiera a solicitar una tarjeta de crédito en un instante de tiempo determinado. De este grupo de clientes se requiere identificar: todos aquellos que lograron obtener su tarjeta de crédito (Ind_p$_i = 1$) y los que fallaron en proceso de aplicación (Ind_p$_i = 0$), teniendo así la variable la variable dependiente ligada a la ecuación del producto. Ahora, por la condición de que D_i es medida siempre y cuando Ind_p$_i = 1$, se necesita entonces información suficiente de comportamiento de pago para aquellos clientes donde Ind_p$_i = 1$.

Del capítulo anterior, se sabe que se requiere al menos 12 meses para que un cliente en la institución financiera muestre su comportamiento real de pago (tiempo de madurez), en esa línea, la muestra de clientes, que se usó en el capítulo 2 para la construcción del modelo clásico, cumple con esta condición de madurez; además, estos clientes tienen la marca $D_i = 1$ y $D_i = 0$, definición de buen y mal cliente para el modelo clásico; sin embargo, carece de información ya que no posee clientes que fallaron en el proceso de acceder a una tarjeta de crédito.

Para solucionar este problema, se toma como universo de análisis a todos los clientes, usados para la construcción del modelo clásico, que obtuvieron su tarjeta a enero 2011 sumando a este grupo, aquellos que al mismo instante de tiempo fallaron en el proceso de aplicación, asegurando de esta manera un universo de clientes, donde se pueden medir todas las variables necesarias para la formulación propuesta por Greene.

3.2. Identificación de las Variables Explicativas para el Modelo de Greene

Una vez identificado el grupo de clientes sobre los cuales se va a trabajar, se definen las variables que van a ser medidas sobre estos clientes.

Acorde a la formulación dada en 5.3, se tienen dos grupos de variables a ser indentificados: el primero, que contiene a todas aquellas variables que miden la tendencia de un cliente hacia el incumpliento ($\mathbf{x}_i$), y el segundo, posee todas las características medidas sobre el cliente en el momento de la aplicación y que influyen en la aceptación o rechazo en el otorgamiento del producto[1] al cliente ($\mathbf{z}_i$).

El primer grupo de variables, ya se identificó previamente en la construcción del modelo clásico, y corresponde a las 27 variables indicadoras que fueron construidas utilizando la metodología de árboles de decisión (ver cuadro 5 del capítulo 2). Por otro lado, el segundo grupo de variables, se presenta en el cuadro 1. Como se puede observar, en su totalidad son variables que tienen que ver con el estado del cliente en el sistema financiero externo a la institución, así como en el sistema comercial (casas comerciales). Cabe recalcar que este grupo de variables serán transformadas a través de árboles de decisión, siguiendo la metodología propuesta en el capítulo 2,

[1]Para este caso puntual el producto que se está analizando es: tarjeta de crédito.

de modo que se puede obtener variables indicadoras con alto poder discriminatorio sobre $\text{Ind\_p}_i = 1$.

Variables	**Descripción**
IngresoEstimado	Ingreso estimado de acuerdo a un modelo propio de la institución.
CuotaEstimada	Cuota promedio que mantiene el cliente en sus créditos activos.
Score	Puntaje que otorga el buró de crédito al cliente (0 - 1000).
PeorCalificacionSBS	Peor calificacion de la Superintendencia de Bancos en los últimos 6 meses.
MayorPlazoVencidoSICOM	Mayor plazo vencido histórico en el sistema comercial.
MayorValorVencidoSICOM	Mayor valor vencido histórico en el sistema comercial.
Tarjetas Activas	Número de tarjetas activas a enero 2011.
CreditoConsumo	Si tiene o no un crédito de consumo activo a enero 2011.

Cuadro 1. Variables que influyen en la aceptación o rechazo de una tarjeta de crédito en la institución financiera en cuestión.

3.3. Estimación del Modelo de Greene

En esta sección se seguirá el esquema de estimación descrito en el capítulo 1, para el modelo condicional resultante de la formulación de Greene:

$$\begin{aligned} D_i^* &= \beta' \mathbf{x}_i + \rho_{\epsilon u} \lambda_i + V_{1i} \\ \text{Ind\_prod}_i{}^* &= \gamma' \mathbf{z}_i + \lambda_i + V_{2i} \end{aligned}$$

El esquema del capítulo 1, divide a la estimación del modelo condicional en tres partes:

1. Estimación de la probabilidad de que $\text{Ind\_prod}_i = 1$.
2. Estimación de λ_i.
3. Estimación de la probabilidad de interés.

Antes del desarrollo de esquema anterior, se identifica al universo de clientes definido en la sección 3.1. Siguiendo entonces, la metodología propuesta para la obtención del grupo de clientes objetivo y analizando la consistencia de la información en cada una de las variables descritas en la tabla 1, se tiene que, de los 52.357 clientes que se tomaron en cuenta, para la construcción del modelo clásico, 17.188 aperturaron su tarjeta en enero 2011. A este grupo de clientes se le suma alrededor de 2.056 clientes que fallaron en el proceso de obtención de su tarjeta de crédito a esa fecha. Teniendo un total de 19.244 registros válidos para la estimación y validación del modelo en cuestión.

Definido entonces el universo de 19.244 clientes, se procede a desarrollar en las secciones posteriores la metodología utilizada para el desarrollo de las tres fases del esquema de estimación propuesto.

3.3.1. Estimación de la probabilidad de que Ind_prod$_i = 1$. Se estima entonces la probabilidad de aceptación que tiene un cliente de tarjeta de crédito considerando el universo de análisis ya definido. A través de un análisis probit con ecuación latente

$$\text{Ind\_p}_i^* = \gamma' \mathbf{z}_i + u_i \tag{3.2}$$

donde $u_i \sim N(0,1)$ y $\text{Ind\_p}_i = 1$ si y solo si $\text{Ind\_p}_i^* > 0$. Para este caso, se sabe que no existe ningún tipo de problema en la muestra para la estimación de los parámetros del modelo, todos los aplicantes se acercaron al azar a la institución sin ningún tipo de influencia. En consecuencia,

$$\begin{aligned} P(\text{Ind\_prod}_i = 1) &= P(\text{Ind\_prod}_i{}^* > 0) \\ &= P(u_i > -\gamma' \mathbf{z}_i) \\ &= P(u_i < \gamma' \mathbf{z}_i) \\ &= \Phi(\gamma' \mathbf{z}_i) \end{aligned}$$

Llegando así a una tranformación de la variable respuesta para garantizar que la predicción esté acotada entre cero y uno. El modelo probit, a diferencia del modelo logístico, utiliza la función de distribución de la ley normal de media cero y varianza unidad para acotar la predicción buscada. La estimación de los parámetros se realiza por máxima verosimilitud de manera completamente análoga al modelo logístico, pero considerando

$$p_i = \Phi(\gamma' \mathbf{z}_i)$$

Como el número de clientes a quienes se les otorgó una tarjeta de crédito (17.188 clientes) no es similar al número de clientes que fallaron en la obtención de una (2.056 clientes), y con el objetivo de que la estimación no esté influenciada por aquellos clientes que son mayoría, se debe tomar una muestra representativa de clientes que son mayoría de tamaño similar al número de clientes que no lograron obtener una tarjeta de crédito. Es decir, se debe tomar una muestra representativa de tamaño igual a 2.056 clientes, teniendo así un universo de 4.112 clientes sobre el cual se construirá el modelo probit. La manera en que se obtendrá esta muestra será a través de un muestreo aleatorio simple sobre los 17.188 clientes que lograron obtener una tarjeta de crédito a enero 2011.

Continuas	**Valor p (K-S)**
IngresoEstimado	0,97
CuotaEstimada	0,98
Score	0,86
MayorPlazoVencidoSICOM	0,99
MayorValorVencidoSICOM	0,98
Tarjetas Activas	0,97
Discretas	**Valor p (Ji-Cuadrado)**
PeorCalificacionSBS	0,97
CreditoConsumo	0,98

Cuadro 2. Resultados de las pruebas de bondad de ajuste K-S y Ji-Cuadrado

La representatividad de la muestra, al igual que en el capítulo 2, es validada estadísticamente a través de las pruebas K-S y Ji-cuadrado. Las variables fueron divididas en discretas y continuas. En el cuadro 2, se puede observar el valor p asociado a cada una de las pruebas, en los distintos grupos de variables, en todos los casos es mayor al 5 % de modo que se puede concluir que las variables en la muestra extraida conservan la distribución de la población al 95 % de confianza.

Para la construcción del modelo probit se utilizarán cada una de las variables del cuadro 1; sin embargo, estas variables serán transformadas en indicadoras utilizando la metodología de árboles de decisión desarrollada en el capítulo 2. En el cuadro 3, se muestran las variables indicadoras producto de la aplicación de la metodología, al igual que en el capítulo 2, el cuadro contiene una columna tipo, para saber a que grupo de clientes apunta la variable indicadora[2]. En este caso Si, corresponde al grupo de clientes que lograron obtener una tarjeta, mientras que No, al grupo de clientes que fallaron al proceso de aplicación.

Variable	**Descripción**	**Tipo**
IndP1	IngresoEstimado >1.625	Si
IndP2	IngresoEstimado =0	No
IndP3	CuotaEstimada >445.8	Si
IndP4	CuotaEstimada =0	No
IndP5	Score >880	Si
IndP6	Score $\leq$ 880	No
IndP7	PeorCalificacionSBS = A,B	Si
IndP8	PeorCalificacionSBS = C,D,E,F,G	No
IndP9	MayorPlazoVencidoSICOM =0	Si
IndP10	MayorPlazoVencidoSICOM >30	No
IndP11	Tarjetas Activas $\leq$ 1	Si
IndP12	Tarjetas Activas >4	No
IndP13	MayorValorVencidoSICOM $\leq$ 9	Si
IndP14	MayorValorVencidoSICOM >33,87	No
IndP15	CreditoConsumo=0	Si

Cuadro 3. Variables indicadores obtenidas a través del algoritmo Chaid.

Una vez encontrado el conjunto de variables con mayor poder discriminatorio en la variable dependiente, se procede a estimar los parámetros del modelo, a través del método de Regresión Paso a Paso, metodología que se utilizó en el capítulo anterior y que será usada de manera totalmente análoga en la búsqueda del mejor modelo para la estimación de la probabilidad de aceptación. El cuadro 4, muestra los coeficientes significativos del modelo final, producto de varias corridas del método. Cabe recalcar, que al igual que en el capítulo anterior, el método fue utilizado una y otra vez hasta obtener un modelo significativo tanto para la prueba F como para la razón t en cada coeficiente, el nivel de significación usado fue del 5 %.

[2]En el Anexo C, se presentan todos los árboles de decisión que se utilizaron para la construcción de las variables indicadores del cuadro 3.

Variable	Descripcion	Coeficiente	Tipo
IndP3	−0,23	CuotaEstimada >445,8	Si
IndP4	−3,22	CuotaEstimada =0	No
IndP5	0,41	Score >880	Si
IndP7	0,64	PeorCalificacionSBS = A,B	Si
IndP10	−0,68	MayorPlazoVencidoSICOM >30	No
IndP12	−0,91	Tarjetas Activas >4	No
IndP14	−0,31	MayorValorVencidoSICOM >33,87	No
IndP15	1,69	CreditoConsumo =0	Si
Constante	0,71	Constante en el modelo	

Cuadro 4. Coeficientes del modelo final luego del procedimiento de regresión paso a paso.

Se analizó también la multicolinealidad en las variables retenidas a través del índice de condicionamiento, con los mismo criterios del capítulo 2, donde se obtuvo un $IC = 5,47$ de modo que se puede concluir que la multicolinealidad no es un problema en el grupo de variables retenidas y por ende en el modelo final estimado, teniendo así una estimación significativa de los parámetros del modelo 3.2.

3.3.2. Estimación de λ_i. De la estimación de los parámetros del modelo anterior ($\hat{\gamma}$), se obtiene la estimación de λ_i como sigue: del capítulo 1 se sabe que

$$\lambda_i = \frac{\phi(-\gamma'\mathbf{z}_i)}{\Phi(\gamma'\mathbf{z}_i)} \tag{3.3}$$

de modo que reemplazando $\hat{\gamma}$ en 3.3, se puede obtener

$$\hat{\lambda_i} = \frac{\phi(-\hat{\gamma}'\mathbf{z}_i)}{\Phi(\hat{\gamma}'\mathbf{z}_i)}$$

Cabe recalcar, que la estimación de λ_i se realiza para el grupo de clientes que lograron obtener su tarjeta de crédito a enero 2011 (17.188 clientes), al ser el grupo de clientes objetivo sobre el cual se estimará $P(D_i = 1\,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i)$, en la sección siguiente.

3.3.3. Estimación de la Probabilidad de Interés. El objetivo de esta apartado es estimar la probabilidad de incumplimiento asociada al modelo de Greene, es decir, $p_i = P(D_i = 1\,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i)$. En el esquema de estimación del capítulo 1, para conseguir la estimación buscada se parte de la siguiente ecuación del modelo condicional

$$D_i^* = \beta'\mathbf{x}_i + \rho_{\epsilon u}\lambda_i + V_{1i} \tag{3.4}$$

llegando a

$$p_i = \tilde{\beta}'\mathbf{x}_i + \tilde{\rho}\lambda_i \tag{3.5}$$

ecuación equivalente a 3.4, y de donde se puede concluir que, incluyendo la estimación de λ_i en 3.5 y realizando una transformación adecuada sobre p_i, se puede acotar la estimación de la probabilidad deseada en el intervalo $(0, 1)$. Tal y como se menciona en el capítulo 1, la transformación a ser utilizada es la función logística

desembocando así en la estimación de un modelo logístico para conseguir la probabilidad p_i de interés.

La estimación del modelo logístico se realizará sobre los 17.188 clientes, es decir, esta estimación se realizará sobre el grupo de clientes que fueron exitosos en la obtención de su tarjeta de crédito a enero 2011. Para la validación del modelo, al igual que en el capítulo 2, se guardarán alrededor de 5.000 clientes, obtenidos a través de un muestreo aleatorio simple sobre los 17.188 clientes. Teniendo así 12.188 clientes sobre los cuales se construirá el modelo logístico.

Definido entonces el grupo de clientes sobre el cual se va a trabajar, se analiza la composición de clientes buenos y malos dentro de este grupo. Como se observa en el cuadro 5, el número de clientes buenos es superior al número de clientes malos. De modo que, para que los resultados no estén influenciados por el número de clientes buenos, se debe tomar una muestra estadísticamnete representativa sobre este grupo de clientes, de modo que el tamaño de la muestra sea similar al tamaño del grupo de clientes malos.

Tipo	**Clientes**
Bueno	11.213
Malo	975
Total	12.188

Cuadro 5. Composición de los clientes para el modelo logístico.

Con la consideración anterior, se extrae una muestra de tamaño 975 sobre los clientes catalogados como buenos, al igual que en el capítulo 2 se valida estadísticamente esta muestra a través de las pruebas K-S y Ji-cuadrado. Las variables a ser utilizadas para este modelo son, como se identicó anteriormente las 27 variables indicadoras, que fueron construidas de manera que poseen el mayor poder discriminatorio con la marca de buen y mal cliente. En el cuadro 6, se presentan el valor p correspondiente a cada una de las pruebas, como se puede ver, en todos los casos es mayor que el 5 %, de modo que se puede concluir al 95 % que la muestra tomada guarda la misma distribución que tiene la población de buenos clientes, en cada una de las variables consideradas.

Se estima entonces el modelo en cuestión sobre un universo de 4.112 clientes. Universo de clientes construido de tal manera que mantiene una proporción de malo a bueno igual a la unidad; en consecuencia, la estimación de la probabilidad de incumplimiento no se verá influenciada por los clientes que son mayoría. En el cuadro 7, se presentan los coeficientes estimados del modelo logit final, este modelo fue obtenido a través del método de regresión paso a paso, al igual que en el capítulo 2, se validó con un nivel de significación del 5 % la prueba F y la significancia de cada coeficiente a través de la razón t. En este cuadro, se observa además que el coeficiente asociado a la estimación de λ_i es significativo, de modo que al no considerar esta variable como regresor se omite información relacionada con la tendencia que tiene el cliente hacia el incumplimiento.

Discretas	**Valor p (Ji-Cuadrado)**
Ind1	0.99
Ind2	0.97
Ind3	0.99
Ind4	0.99
Ind5	0.95
Ind6	0.99
Ind7	0.99
Ind8	0.99
Ind9	0.98
Ind10	0.99
Ind11	0.99
Ind12	0.98
Ind13	0.97
Ind14	0.97
Ind15	0.99
Ind16	0.97
Ind17	0.95
Ind18	0.96
Ind19	0.99
Ind20	0.99
Ind21	0.99
Ind22	0.99
Ind23	0.99
Ind24	0.99
Ind25	0.97
Ind26	0.97
Ind27	0.98
Continuas	**Valor p (K-S)**
$\hat{\lambda}_i$	0.98

Cuadro 6. Validación de la muestra de clientes buenos para la construcción del modelo logístico.

Se calculó también el índice de condicionamiento sobre las variables retenidas en el modelo final, obteniendo un $IC = 5{,}29$. Concluyendo así, que la multicolinealidad no es un problema que afecte a la estimación de los coeficientes del modelo final.

Por otro lado, sobre los 5.000 clientes que no se consideraron para la construcción y estimación del modelo logit, se calculan los indicadores Gini y AUROC. En el cuadro 8, se tienen los valores de los indicadores mencionados, como se puede apreciar, el indicador de GINI es mayor a 0,5, y el $AUROC$ cercano a 1; en consecuencia, se puede concluir que el poder de discriminación del modelo final es satisfactorio.

Obtenida la estimación de $P(D_i = 1 \,|\mathbf{x}_i, \lambda_i,\ u_i > -\gamma'\mathbf{z}_i)$, se procede en el capítulo siguiente a comparar esta estimación con la dada por el modelo clásico desarrollado en el capítulo 2.

Variable	Descripción	Coeficiente	Tipo
Ind1	Edad ≤ 38	0.57	Malo
Ind2	Edad > 49	-0.37	Bueno
Ind4	PeorCalif= D,E,F o G	0.57	Malo
Ind5	Antigüedad≤ 330	1.84	Bueno
Ind6	Antigüedad $>$ 2,289	-0.95	Bueno
Ind7	Antigüedad entre 330 y 461	0.87	Malo
Ind9	AtrMax≤ 12	-0.82	Bueno
Ind10	AtrMax> 13	-0.45	Malo
Ind11	AtrProm ≤ 1	-0.73	Bueno
Ind12	AtrProm > 4	1.54	Malo
Ind13	Actividad ≤ 11	-0.99	Bueno
Ind14	MinSaldo≤ 0	-0.51	Bueno
Ind15	MinSaldo entre 0 y 76	-0.71	Malo
Ind16	MinSaldo $>$ 2,727	0.69	Malo
Ind17	NumPago ≤ 9	0.72	Malo
Ind23	Baja, Edad ≤ 38	-0.48	Malo
Ind27	Media, Edad > 38	0.65	Bueno
Lambda		-0.18	

Cuadro 7. Coeficientes significativos del modelo logístico final.

Gini	AUROC
0,69	0,84

Cuadro 8. AUROC y GINI en la muestra de validación del modelo final.

Capítulo 4

Comparación entre el Modelo Clásico y la Nueva Propuesta de Greene

Una vez que se ha desarrollado el tratamiento clásico para la estimación de la probabilidad de incumplimiento, así como la nueva propuesta de Greene que considera el sesgo de selección latente en la muestra de construcción, se procede a comparar las estimaciones de ambos modelos, de manera que se pueda determinar el efecto de utilizar tanto el primero como el segundo método. La muestra de clientes sobre la cual se trabajará, serán los 17.188 clientes utilizados para la construcción y estimación del modelo de Greene. Este grupo de clientes fue escogido al tener información necesaria para la estimación de la probabilidad de incumplimiento en ambas metodologías.

En un principio, se comparan ambas estimaciones de modo que se pueda determinar qué tan diferentes son entre sí. Para ello, se calcula la diferencia que existe entre la estimación que se da por la vía clásica contra aquella que proporciona la formulación de Greene. En primer lugar, se analiza esta diferencia de manera global en cada grupo de clientes (clientes buenos y malos) así como en la población total. Luego, se cuenta el número de casos en los que esta diferencia mantiene un valor positivo y también el número de casos en los cuales esta diferencia es negativa, de manera que se pueda determinar qué estimación mantiene un valor superior sobre la muestra de clientes analizados. Por otro lado, el poder discriminatorio entre clientes malos y buenos para ambos modelos, es medido a través de los indicadores Gini y AUROC, así como el error de mala clasificación, definiendo una marca de buen y mal cliente estimada también dentro de la estimación de cada modelo.

Finalmente, se compara la pérdida esperada generada por cada modelo con la pérdida real en la cartera de clientes de tarjeta de crédito a una fecha determinada, de modo que se pueda observar qué modelo mantiene la predicción más cercana al valor real.

4.1. Comparación de las Estimaciones de ambos Modelos

El objetivo de este apartado es medir cuan diferentes son las estimaciones de la probabilidad de incumplimiento dadas tanto por el modelo clásico así como por el modelo de Greene. Para ello, se calcula en primer lugar la diferencia promedio alcanzada por cada modelo en la población total y en los dos grupos de clientes que se utilizaron para la construcción de los mismos (clientes buenos y malos). Para el cálculo de esta diferencia se tomará como referencia la estimación dada por Greene.

Los resultados del procedimiento anterior se presentan en el cuadro 1. Como se puede observar, la estimación dada por la formulación de Greene es superior en

Cliente	Probabilidad (Greene)	Probabilidad (Clásico)	Diferencia
Bueno	32 %	24 %	8 %
Malo	68 %	60 %	8 %
Total	35 %	26 %	9 %

Cuadro 1. Probabilidades promedio por tipo de cliente en los dos modelos.

promedio a la presentada por la vía clásica, tanto en clientes malos como buenos, así como en la población total. Por otro lado, se puede ver además, que la diferencia cuadrática media entre las probabilidades en promedio es la misma tanto para clientes buenos como para los clientes catalogados como malos; sin embargo, la población general no mantiene el mismo comportamiento al crecer en un punto porcentual frente al valor mantenido en ambos grupos de clientes. En consecuencia, se puede decir que al no considerar el sesgo de selección en la estimación de la probabilidad de incumplimiento, en promedio se subestima el valor de la misma.

Cliente	Positiva		Negativa	
	% de clientes	Diferencia	% de clientes	Diferencia
Bueno	77 %	12 %	33 %	6 %
Malo	81 %	71 %	19 %	5 %
Total	77 %	13 %	33 %	6 %

Cuadro 2. Diferencias positivas y negativas entre ambos modelos.

Ahora, con el objetivo de estudiar los valores que toma la diferencia en la muestra de clientes analizados, se procede a analizar el número de veces donde esta diferencia toma un valor positivo así como el número de casos en donde la misma adquiere un valor negativo. En el cuadro 2, se presenta en porcentaje el número de casos mencionados. Se puede observar que un 77 % de los clientes buenos mantienen una diferencia positiva entre las estimaciones; de igual manera, en el grupo de clientes malos un 81 % mantienen este mismo comportamiento, siendo minoría aquellos clientes donde la diferencia presenta valores negativos (33 % y 19 %, respectivamente). En la población general al igual que en cada grupo de clientes, los clientes que presentan una diferencia positiva son aquellos que son mayoría representando un 77 % de la población analizada.

En el cuadro 2, también se puede apreciar el cálculo de la diferencia cuadrática media en los diferentes grupos de clientes, donde se puede ver que cuando los clientes mantienen una diferencia positiva la brecha entre las estimaciones es mayor que en aquellos que poseen una diferencia negativa. En consecuencia, la estimación del modelo de Greene es superior a la dada por el modelo clásico, en la mayoría de los clientes analizados.

4.2. Poder Discriminatorio

El objetivo de esta sección es comparar el poder discriminatorio que tiene cada modelo entre los dos grupos de clientes (buenos y malos). Los indicadores que se medirán sobre las estimaciones de cada modelo serán:

- Coeficiente de Gini.

- Indicador AUROC.

Hay que recalcar, que estos indicadores ya han sido utilizados para evaluar la capacidad discriminatoria en cada modelo pero sobre las respectivas muestras de validación que se guardaron en cada caso. En este apartado, se evaluarán estos indicadores para cada modelo, pero sobre la misma muestra de clientes, de modo que los resultados puedan ser comparables.

Indicador	**Greene**	**Clásico**
Gini	0,69	0,63
AUROC	0,84	0,82

Cuadro 3. AUROC y GINI para ambos modelos.

En el cuadro 3, se presentan los valores de los indicadores para cada modelo, como se resaltó anteriormente los indicadores fueron calculados sobre la muestra de 17.188 clientes. Tanto el indicador Gini como el AUROC, muestran valores superiores para el modelo de Greene en comparación con la formulación clásica. Por lo tanto, se puede decir que sobre la muestra de clientes analizados, el modelo de Greene tiene mayor poder discriminatorio entre clientes buenos y malos que el modelo clásico.

Greene	**Bueno**	**Malo**
Bueno	78 %	23 %
Malo	22 %	77 %

Cuadro 4. Porcentajes de buenos y malos retenidos de acuerdo al modelo de Greene.

Clásico	**Bueno**	**Malo**
Bueno	77 %	25 %
Malo	23 %	75 %

Cuadro 5. Porcentajes de buenos y malos retenidos de acuerdo al modelo clásico.

De igual manera, el poder de discriminación entre clientes malos y buenos se puede analizar trazando un punto de corte en cada estimación de modo que se pueda tener de acuerdo al modelo, una marca de buen y mal cliente estimada. Usualmente el punto de corte elegido es aquel que reduce el error de mala clasificación al 20 %, tanto en clientes malos como en clientes buenos. Con esta consideración, se obtienen dos puntos de corte,[1] uno para cada modelo. El primero clasifica a un cliente como bueno de acuerdo a Greene si su probabilidad de no pago estimada es menor al 51 %, mientras para el modelo clásico esta definición se ubica en el 40 %; en ambos casos si la condición se incumple el cliente es marcado como malo.

[1]Para construir estos puntos de corte la población fue dividida en deciles de acuerdo a la estimación de la probabilidad de cada modelo, siendo el valor buscado el decil correspondiente que asegura el porcentaje de error de mala clasificación fijado.

En los cuadros 4 y 5, se tiene el porcentaje de clientes buenos y malos retenidos de acuerdo a cada una de estas definiciones. Acorde a estos cuadros se puede observar que ambas definiciones permiten retener alrededor del 80 % en buenos y malos clientes, teniendo alrededor de un 20 % de clientes mal clasificados. En esa línea se puede observar que el modelo de Greene retiene un 2 % más de clientes malos que el modelo clásico, así como un 1 % más en clientes catalogados como buenos.

En consecuencia, el modelo de Greene frente al modelo clásico mantiene un menor error de clientes mal clasificados. De igual manera, se calculó la perdida real que se produce en el error de tipo 1 para cada uno de los modelos.

Para el modelo de Greene, la pérdida[2] alcanzada por los clientes mal clasificados fue de alrededor de 523.960 dólares, mientras que en el modelo clásico esta pérdida alcanzó un valor de 675.118 dólares. De modo que si se optara por clasificar de acuerdo a la regla de Greene se perdería menos que al optar por la regla dada por el modelo clásico; sin embargo, hay que considerar lo que se dejó de ganar al cometer un error del tipo 2 de acuerdo a la mala clasificación que se tiene por cada modelo. Cabe recalcar, que este tipo información es muy difícil de estimar ya que el cliente tiene una tasa de interés diferente en cada consumo diferido que varía de acuerdo al plazo al que se realice la compra; es por ello que esta estimación no es posible obtener tan fácilmente sin un tratamiento previo de la información de por medio. Sin embargo, en la sección siguiente, a manera de ejercicio, se calculan las pérdidas esperadas sobre la cartera de los 17.188 clientes para cada modelo y se observa que modelo mantiene una mejor estimación sobre la pérdida real generada en ese período para esa cartera de clientes. Cabe recalcar, que no se profundizará sobre este tema al no ser parte del alcance del documento.

4.3. Pérdida Esperada

El objetivo de esta sección es calcular la pérdida esperada para cada modelo en la cartera de 17.188 clientes, y comparar cada uno de estos valores esperados con la pérdida real que los clientes mantienen a una fecha determinada. La fecha a la que hará referencia será diciembre 2012, y al igual que en la sección anterior se entenderá como pérdida real al dinero que el cliente no pagó a lo largo del período enero 2012 - diciembre 2012. En consecuencia, este será el valor con el cual se tendrá que comparar el valor estimado por cada modelo.

La pérdida esperada se define mediante la siguiente expresión[3]:

$$P_e = E(1 - r)P_i \tag{4.1}$$

Donde:
E: nivel de exposición del riesgo de crédito.
r: tasa de recuperación.
P_i: probabilidad de incumplimiento.

[2]Se entiende como pérdida al dinero acumulado que el cliente no pagó durante el período: enero 2012 - diciembre 2012.

[3]Para un mayor detalle sobre esta definición consultar **[5]**

Si se asume que un cliente en particular no tiene nada que ver con algún otro del universo de clientes analizados; es decir, todos los clientes a analizar son independientes entre sí. La pérdida esperada global sería la suma de cada una de las pérdidas esperadas por cliente. Si además se toma como E el saldo que mantiene el cliente en su tarjeta de crédito a diciembre 2012 y como P_i la probabilidad estimada por cada uno de los modelos, tan solo faltaría calcular la tasa de recuperación para poder calcular P_e en 4.1. Si se entiende a r como el porcentaje de clientes que logran recuperarse después de haber incumplido en alguno de sus pagos en un período de tiempo determinado, se podría calcular este valor como sigue: la institución financiera maneja un rango de calificaciones para sus clientes de acuerdo a sus pagos vencidos en un intervalo de 12 meses, lo que se hace es determinar qué calificación tenían los tarjetahabientes activos con saldo vigente, a enero 2012 y a través de una matriz de transición ver qué calificación alcanzaron a diciembre 2012.

Calificación	**Descripción**
A	Cliente con 0 pagos vencidos.
B	Cliente con 1 pago vencido.
C,D,E,F y G	Ciente con más de 1 pago vencido.

Cuadro 6. Calificación interna de un tarjetahabiente.

En el cuadro 6, se tiene la calificación que la institución maneja para cada uno de sus clientes. Como se puede observar, los rangos de calificación van desde A hasta G, y están en función de los pagos que el cliente no realizó (pagos vencidos). De acuerdo a la definición de buen y mal cliente tratada en el capítulo 2, los clientes buenos eran aquellos que tenían un atraso máximo menor a 35 días, mientras que los malos eran todos los clientes cuyo atraso máximo era mayor a 46 días. De acuerdo a la calificación de la institución los clientes buenos mantendrían una calificiación A o B, mientras que los catalogados como malos calificaciones C, D, E, F y G. Estudiando entonces la transición de calificaciones en el período enero 2012 - diciembre 2012 para el universo de clientes definido, se observa que el 91 % de los clientes que a enero 2012 erán A o B a diciembre 2012 siguen teniendo las mismas calificaciones; es decir, únicamente el 9 % degradaron su calificación a esa fecha. Por el contrario, aquellos clientes que a enero 2012 tenían calificaciones C, D, E, F ó G, el 65 % obtuvieron la misma calificación a diciembre 2012, mientras que el 35 % restante pasaron a tener calificaciones A o B. Por lo tanto, r para nuestro caso toma dos valores, $r = 91\,\%$ para clientes buenos y $r = 35\,\%$ para clientes malos.

Modelo	**Pérdida Esperada**	**Pérdida Real**	**Diferencia**
Greene	2.136.003	1.979.014	-156.989
Clásico	1.587.716	1.979.014	391.298

Cuadro 7. Pérdidas esperadas para el modelo de Greene y la formulación clásica.

Los resultados del cálculo de la pérdida esperada para cada uno de los modelos así como el valor de pérdida real en la cartera de clientes se presentan en el cuadro 7; como se puede observar, el modelo de Greene presenta un valor mucho más cercano al valor real de pérdida que alcanzaron los cliente durante ese año.

Capítulo 5

Conclusiones y Recomendaciones

5.1. Conclusiones

De acuerdo a lo tratado en el presente documento se puede concluir lo siguiente:

- La nueva formulación propuesta por Greene se reduce a un modelo condicional dado por las siguientes ecuaciones:

$$D_i^* = \beta' \mathbf{x}_i + \rho_{\epsilon u} \lambda_i + V_{1i} \tag{5.1}$$

$$\text{Ind\_prod}_i{}^* = \gamma' \mathbf{z}_i + \lambda_i + V_{2i} \tag{5.2}$$

  donde, siguiendo el esquema de estimación dado por Heckman (1979) y las suposiciones realizadas sobre el modelo de Greene, se tiene que la estimación de la probabilidad de interés se resume en estimar un modelo logístico donde

$$p_i = \frac{1}{1 + \exp(-\tilde{\beta}' \mathbf{x}_i - \tilde{\rho} \lambda_i)} \tag{5.3}$$

  Por otro lado, la formulación clásica para la estimación de la probabilidad de incumplimiento, al igual que en Greene, desemboca en la estimación de los coeficientes de un modelo logísto donde esta vez

$$p_i = \frac{1}{1 + \exp(-\beta_1 - \beta' \mathbf{x}_i)} \tag{5.4}$$

  En consecuencia, al comparar 5.3 y 5.4, se tiene que el problema del sesgo de selección se reduce a la omisión de regresores importantes en la formulación clásica. Se entiende al inverso de la razón de Mills como un regresor que resume de alguna manera todos los regresores no incluidos en la formulación clásica, pero influyentes en el incumplimiento de un cliente.

- Considerando lo mencionado anteriormente, una manera de medir la significación que tiene el sesgo de selección en la estimación de la probabilidad de incumplimiento, es a través de la razón t, asociada al coeficiente de regresión del inverso de la razón de Mills. En el caso de la institución financiera de la cual se extrajo la información, se puede concluir que el problema del sesgo de selección es significativo al 95 % de confianza, ya que el coeficiente del inverso de la razón de Mills lo es a ese nivel de significación.

- De acuerdo a las comparaciones realizadas en el capítulo 4 entre las estimaciones de la probabilidad de incumplimiento dadas por ambos modelos, se observa que estimar la probabilidad de no pago a través de la metodología propuesta por Greene, es en general superior a la dada por el modelo clásico en un 77 % de la población analizada, mientras que es inferior en tan solo un 33 %. En esa línea, se puede concluir que al estimar la probabilidad de incumplimiento de la manera clásica se subestima en promedio esta probabilidad en un 9 %.

- De igual manera, observando los resultados en los indicadores de Gini y AUROC calculados sobre la misma muestra de clientes para ambos modelos, se tiene que la estimación dada por la formulación de Greene mantiene un mayor poder de discriminación entre clientes malos y buenos frente a la estimación dada por el modelo clásico.
- Finalmente, la pérdida esperada en el modelo de Greene tiene una diferencia mucho menor que la calculada por el modelo clásico, de modo que al utilizar la nueva formulación de Greene se puede estimar mucho mejor la pérdida que se espera de la cartera de clientes un año después, al momento en el que se evalúa al cliente con el modelo.

5.2. Recomendaciones

- Dado que el problema del sesgo de selección se reduce a la omisión de regresores importantes, se podría menorar su incidencia incluyendo en la formulación clásica más variables de comportamiento externo del cliente tanto en el sistema comercial así como en el sitema financiero externo a la institución.
- Otra manera de probar la significancia que posee el sesgo de selección en la estimación de la probabilidad de incumplimiento, es a través de la prueba de hipótesis que plantea Heckman (1979).
- Para probar el poder discriminatorio entre clientes malos y buenos y la estabilidad de los coeficientes de cada modelo, se pueden calcular los indicadores Gini y AUROC y evaluarlos en diferentes muestras medidas en distintas instancias de tiempo.
- Se debería profundizar más en el cálculo de la pérdida real, considerando las variaciones positivas mensuales que tienen las provisiones regulatorias, las cuales se calculan siempre y cuando un cliente incumple en sus pagos, y están determinadas por la Superintendencia de Bancos del Ecuador. En esa línea, se podría obtener la distribución de pérdidas de la cartera de clientes, de modo que se pueda afinar el cálculo de la pérdida esperada así como la obtención de la aproximación de la pérdida inesperada.
- Por último, se debería medir el error de tipo 2 incurrido en cada modelo considerando consumos tanto diferidos como corrientes, para poder conocer realmente lo que se deja de ganar cuando se clasifica a un cliente que es bueno como malo, de acuerdo a la regla de decisión impuesta por la estimación del modelo en cuestión.

Bibliografía

[1] Anderson R., 2007, "Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation", Oxford University Press, Oxford.

[2] Breiman L., Friedman J., Olshen R., Stone C., 1984, "Classification and Regresion Trees", Wadsworth International Group.

[3] Brezis H., 1983, "Analyse Fonctionnelle", MASSON, París.

[4] Castro A., 2008, "Regresión Lineal", Monografías de Matemática y Estadística, Quito.

[5] Diz E., 2006, "Teoría de Riesgo", ECOEDICIONES, Bogotá.

[6] Greene W., 2003, "Econometric Analysis", Prentice Hall, Englewood Cliffs, New Jersey.

[7] Greene W., 1998, "Sample Selection in credit-scoring models", Japan and the World Economy, 10, 299-316.

[8] Heckman J., 1979, "Sample selection bias as a specification error", Econometrica, 47, 153-161.

[9] Johnson N., Kotz S., 1972, "Distribution in Statistics: Continuous Multivariate Distributions", John Wiley and Sons, New York.

[10] Knight K., 2000, "Mathematical Statistics", Chapman& Hall/CRC, New York.

[11] Zeilis A., Leisch F., Hornik K., Kleiber C., 2002, "strucchange: An R Package for Testing Structural Change in Linear Regression Models", CRAN.

[12] Zeilis A., 2000, "p-Values and Alternative Boundaries for CUSUM Tests", Working Paper No. 78, Vienna University of Economics and Business Administration.

ANEXO A

Ley Normal Multivariante (Definiciones y Propiedades)

Definición 1 (Vector Aleatorio)**.** *Un vector aleatorio es un vector de* R^n *donde cada una de sus componentes son variables aleatorias.*

Definición 2 (Ley Normal Multivariante)**.** *Sean:* Σ *una matriz* $n \times n$ *simétrica definida positiva,* μ *un vector de* R^n*. Se dice que el vector*[1]

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

sigue una ley normal multivariante de parámetros (μ, Σ) *si admite por densidad la función:*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\right), \ \forall \boldsymbol{x}$$

Se escribe $\boldsymbol{X} \sim N_n(\mu, \Sigma)$*.*

Propiedad 1. *Si* $(X, Y) \sim N_2(0, \Sigma)$ *con* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *entonces su función de densidad viene dada por*

$$f_{(X,Y)}(x, y) = \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right) \tag{A.1}$$

Demostración. Como $(X, Y) \sim N_2(0, \Sigma)$, se sigue de la definición 2 (haciendo $n = 2$) que

$$f_{(X,Y)}(x, y) = \frac{1}{(2\pi)\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}([x, y]'\Sigma^{-1}[x, y])\right) \ \ x, y \in \mathrm{R} \tag{A.2}$$

Realizando el cálculo de $|\Sigma|$, se tiene que

$$|\Sigma| = 1 - \rho^2 \tag{A.3}$$

De igual manera usando el procedimiento clásico para obtener la inversa de una matriz de dimensión (2×2):

$$\Sigma^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \tag{A.4}$$

[1]Siempre para vectores tomar como referencia la forma del vector $\mathbf{X}$, donde $X' = (X_1, X_2, \ldots, X_n)$. Cabe recalcar que por facilidad se suele escribir el vector $\mathbf{X}$ como su transpuesto; sin embargo, se debe tomar en cuenta para las operaciones que tiene la forma ya mencionada.

Reemplazando A.3 y A.4 en A.2 y realizando las simplificaciones algebraicas respectivas

$$f_{(X,Y)}(x,y) = \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right) \tag{A.5}$$

□

Propiedad 2. *Si* $(X,Y) \sim N_2(0,\Sigma)$ *con* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ *, entonces tanto* X *como* Y *siguen una ley* $N(0,1)$.

Demostración. Por la propiedad 1 se sabe la forma de la función de densidad del vector aleatorio $(X,Y)'$ para obtener la función de densidad de X, se integra entonces A.1 sobre el dominio de y

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{+\infty} \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right) dy \\
&= \int_{-\infty}^{+\infty} \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2 + \rho^2x^2 - \rho^2x^2)\right) dy \\
&= \int_{-\infty}^{+\infty} \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2(1-\rho^2) + y^2 - 2xy\rho + \rho^2x^2)\right) dy \\
&= \exp\left(-\frac{x^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{(2\pi)\sqrt{(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) dy \\
&= \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{1/2}\sqrt{(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) dy
\end{aligned}$$

Haciendo $u = y - \rho x$ se tiene que

$$f_X(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{1/2}\sqrt{(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) du$$

pero

$$\frac{1}{(2\pi)^{1/2}\sqrt{(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right)$$

es la función de distribución de una variable U que sigue una ley $N(0,(1-\rho^2))$, de modo que

$$\int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{1/2}\sqrt{(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) du = 1$$

en consecuencia

$$f_X(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right)$$

de donde se sigue que $X \sim N(0,1)$. El procedimiento para la obtención de la ley marginal de Y es totalmente análogo al que se ha seguido para la variable aleatoria X. □

Propiedad 3. *Sea* $(X,Y) \sim N_2(0,\Sigma)$ *con* $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ $y\ z \in \mathrm{R}$ *se sigue entonces:*

i. $E(X \,|Y \geqslant z) = \rho \dfrac{\phi(z)}{\Phi(-z)}$

ii. $E(XY\,|Y \geqslant z\,) = \rho\left(z\dfrac{\phi(z)}{\Phi(-z)} + 1\right)$

iii. $E(X\,|X \geqslant z\,) = \dfrac{\phi(z)}{\Phi(-z)}$

iv. $E(X^2\,|Y \geqslant z\,) = (1-\rho^2) + \rho^2\left(z\dfrac{\phi(z)}{\Phi(-z)} + 1\right)$

v. $E(X^2\,|X \geqslant z\,) = z\dfrac{\phi(z)}{\Phi(-z)} + 1$

DEMOSTRACIÓN. Se prueba cada una de las propiedades en el orden establecido:

i. Se calcula la función de distribución de la variable aleatoria $X\,|Y \geqslant z$

$$\begin{aligned} F_{X|Y\geqslant z}(x) &= P(X \leq x\,|Y \geqslant z\,) \\ &= \frac{P(X \leq x, Y \geq z)}{P(Y \geq z)} \\ &= \frac{1}{P(Y \geq z)}\int_{-\infty}^{x}\int_{z}^{+\infty} f_{(X,Y)}(x,y)dydx \end{aligned}$$

Siendo $f(x,y)$ la función de densidad del vector (X,Y), pero $(X,Y) \sim N_2(0,\Sigma)$, por la propiedad 1

$$f_{(X,Y)}(x,y) = \frac{1}{(2\pi)\sqrt{(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right)$$

A más de ello, por la propiedad 2, se tiene que $Y \sim N(0,1)$ y debido a la simetría de la ley de Y

$$P(Y \geq z) = P(Y \leq -z)$$

y utilizando la notación usual para la función de distribución normal estándar

(A.6)
$$F_{X|Y\geqslant z}(x) = \frac{1}{\Phi(-z)}\int_{-\infty}^{x}\int_{z}^{+\infty}\frac{1}{(2\pi)\sqrt{(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right)dydx$$

Derivando esta expresión con respecto a x se tiene entonces que

(A.7)
$$f_{X|Y\geqslant z}(x) = \frac{1}{\Phi(-z)(2\pi)\sqrt{(1-\rho^2)}}\int_{z}^{+\infty}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right)dy$$

De modo que

$$\begin{aligned} E(X\,|Y \geq z\,) &= \int_{-\infty}^{+\infty} x f_{X|Y\geqslant z}(x)dx \\ &= \frac{1}{\Phi(-z)(2\pi)\sqrt{(1-\rho^2)}}\int_{-\infty}^{+\infty}\int_{z}^{+\infty} x\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2xy\rho + y^2)\right)dydx \\ &= \frac{1}{\Phi(-z)(2\pi)\sqrt{(1-\rho^2)}}\int_{-\infty}^{+\infty}\int_{z}^{+\infty} x\exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)} - \frac{y^2}{2}\right)dydx \end{aligned}$$

Ahora dado que $f(x\,|Y \geqslant z)$ es continua en $\mathrm{R} \times (z, +\infty)$, aplicando el Teorema de Fubini

$$\begin{aligned} E(X\,|Y \geq z) &= \frac{1}{\Phi(-z)(2\pi)\sqrt{(1-\rho^2)}} \int_z^{+\infty} \int_{-\infty}^{+\infty} x \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)} - \frac{y^2}{2}\right) dxdy \\ &= \frac{((2\pi)(1-\rho^2))^{-1}}{\Phi(-z)} \int_z^{+\infty} \exp\left(-\frac{y^2}{2}\right) \int_{-\infty}^{+\infty} x \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dxdy \\ &= \frac{((2\pi)(1-\rho^2))^{-1}}{\Phi(-z)} \int_z^{+\infty} \exp\left(-\frac{y^2}{2}\right) \left[\int_{-\infty}^{+\infty} (x-\rho y) \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \right. \\ &+ \left. \rho y \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned}$$

Haciendo $u = x - \rho y$

$$\begin{aligned} E(X\,|Y \geq z) = \frac{((2\pi)(1-\rho^2))^{-1/2}}{\Phi(-z)(2\pi)^{1/2}} \int_z^{+\infty} \exp\left(-\frac{y^2}{2}\right) \left[\int_{-\infty}^{+\infty} u \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \right. \\ \left. + \rho y \int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned} \tag{A.8}$$

Pero

$$f_U(u) = \frac{1}{(\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right)$$

es la función de densidad de una varibale $U \sim N(0, (1-\rho^2))$, en consecuencia

$$\begin{aligned} \int_{-\infty}^{+\infty} u \frac{1}{(\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) &= E(U) \\ &= 0 \end{aligned}$$

y

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) = 1$$

De modo que, utilizando estos resultados en A.8

$$E(X\,|Y \geq z) = \frac{\rho}{\Phi(-z)} \int_z^{+\infty} \frac{y}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

Haciendo un cambio de variable con $w = \frac{y^2}{2}$

$$\begin{aligned} E(X\,|Y \geq z) &= \frac{\rho}{\Phi(-z)} \int_{\frac{z^2}{2}}^{+\infty} \sqrt{2\pi} \exp(-w)\, dw \\ &= \frac{\rho}{\Phi(-z)\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \end{aligned}$$

Donde

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

es la función de densidad de una variable $Z \sim N(0,1)$, de modo que utilizando la notación estándar para la función de densidad de esta ley

$$E(X\,|Y \geq z) = \rho \frac{\phi(z)}{\Phi(-z)}$$

ii. Se procede a calcular la función de distribución conjunta del vector $(X, Y)\,|Y \geq z$

$$\begin{aligned} F_{(X,Y)|Y\geq z}(x,y) &= P(X \leq x, Y \leq y\,|Y \geq z) \\ &= \frac{P(X \leq x, Y \leq y, Y \geq z)}{P(Y \geq z)} \\ &= \frac{P(X \leq x, z \leq Y \leq y)}{P(Y \geq z)} \\ &= \frac{1}{P(Y \geq z)} \int_{-\infty}^{x} \int_{z}^{y} f(x,y)dydx \end{aligned}$$

Derivando esta expresión parcialmente tanto para x como para y (aplicando la Observación 1 del Anexo B), se obtiene la función de densidad del vector aleatorio $X, Y\,|Y \geq z$, dada por la expresión

$$f_{(X,Y)|Y\geq z}(x,y) = \frac{f(x,y)}{P(Y \geq z)} \text{ si } y > z$$

donde $f(x,y)$ es la función de densidad de la ley normal bivariante de parámetros $(0, \Sigma)$, dado que $(X, Y) \sim N_2(0, \Sigma)$. Se calcula entonces $E(XY\,|Y \geq z)$ como sigue

$$\begin{aligned} E(XY\,|Y \geq z) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{(X,Y)|Y\geq z}(x,y)dydx \\ &= \int_{-\infty}^{+\infty} \int_{z}^{+\infty} xy \frac{f(x,y)}{P(Y \geq z)}dydx \end{aligned}$$

Luedo dado que $f(x,y)$ es continua sobre $\mathrm{R} \times [z, +\infty)$, aplicando el Teorema de Fubini (ver Anexo B) se tiene

$$\begin{aligned} E(XY\,|Y \geq z) &= \int_{z}^{+\infty} \int_{-\infty}^{+\infty} xy \frac{f(x,y)}{P(Y \geq z)}dxdy \\ &= \frac{1}{P(Y \geq z)} \int_{z}^{+\infty} y \int_{-\infty}^{+\infty} \frac{(x - \rho y + \rho y)}{2\pi\sqrt{(1-\rho^2)}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)} - \frac{y^2}{2}\right) dxdy \\ &= \frac{(2\pi(1-\rho^2))^{-1/2}}{(2\pi)^{1/2} P(Y \geq z)} \int_{z}^{+\infty} y \exp\left(-\frac{y^2}{2}\right) \left[\int_{-\infty}^{+\infty} (x - \rho y) \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \right. \\ &+ \left. \rho y \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned}$$

Haciendo $u = x - \rho y$

$$\begin{aligned} E(XY\,|Y \geq z) = \frac{((2\pi)(1-\rho^2))^{-1/2}}{P(Y \geq z)(2\pi)^{1/2}} \int_{z}^{+\infty} y \exp\left(-\frac{y^2}{2}\right) \left[\int_{-\infty}^{+\infty} u \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \right. \\ \left. + \rho y \int_{-\infty}^{+\infty} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned} \tag{A.9}$$

Pero

$$f_U(u) = \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right)$$

es la función de densidad de una varibale $U \sim N(0,(1-\rho^2))$, en consecuencia

$$\begin{aligned} E(U) &= \int_{-\infty}^{+\infty} u \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) \\ &= 0 \end{aligned}$$

y

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) = 1$$

De modo que, utilizando estos resultados en A.9

$$E(XY\,|Y \geq z) = \frac{\rho}{P(Y \geq z)} \int_z^{+\infty} \frac{y^2}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

Integrando por partes con $u = y$ y $v = -\exp(-y^2/2)$, se tiene

$$E(XY\,|Y \geq z) = \frac{\rho}{P(Y \geq z)} \left[\frac{z}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right) + \int_z^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy\right]$$

Luego usando la notación usual para representar tanto la función de densidad como de distribución de la ley normal estándar se sigue

$$E(XY\,|Y \geq z) = \rho\left(z\frac{\phi(z)}{\Phi(-z)} + 1\right)$$

iii. Se calcula función de distribución para la variable aleatoria $X\,|X \geq z$

$$\begin{aligned} F_{X|X\geq z}(x) &= \frac{P(X \leq x, X \leq z)}{P(X \leq z)} \\ &= \frac{P(z \leq X \leq x)}{P(X \geq z)} \text{ si } x > z \\ &= \frac{P(z \leq X \leq x)}{P(X \geq z)} \\ &= \frac{F(x) - F(z)}{P(X \geq z)} \end{aligned}$$

Derivando esta expresión con respecto a x, para obtener la ley de la variable aleatoria en cuestión se tiene

$$f_{X|X\geq z}(x) = \frac{f(x)}{P(X \geq z)}, \text{ si x>z} \tag{A.10}$$

cabe recalcar que se impone la condición $x > z$ para que la función de distribución no se anule. Luego se calcula el valor esperado de la variable $X\,|X \geq z$ como sigue

$$\begin{aligned} E(X\,|X \geq z) &= \int_z^{+\infty} x f_{X|X\geq z}(x) dx \\ &= \int_z^{+\infty} x \frac{f(x)}{P(X \geq z)} dx \end{aligned}$$

Ahora por la propiedad 1 se tiene que $X \sim N(0,1)$, de modo que

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

En consecuencia

$$E(X\,|X \geq z) \;=\; \frac{1}{P(X \geq z)\sqrt{2\pi}} \int_z^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx$$

Haciendo $u = \frac{x^2}{2}$ se tiene

$$\begin{aligned} E(X\,|X \geq z) &= \frac{1}{P(X \geq z)\sqrt{2\pi}} \int_{\frac{z^2}{2}}^{+\infty} \exp(-u)\, du \\ &= \frac{1}{P(X \geq z)\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \end{aligned}$$

Usando la notación estándar para representar tanto la función de densidad como la función de distribución acumulada de la ley normal

$$E(X\,|X \geq z) \;=\; \frac{\phi(z)}{\Phi(-z)}$$

iv. Usando (A.7) y la continuidad de $f(x,y)$ en $\mathrm{R} \times (z, +\infty)$ se tiene entonces

$$\begin{aligned} E(X^2\,|X \geq z) \quad &= \quad \frac{(2\pi(1-\rho^2))^{-1/2}}{(2\pi)^{1/2}\Phi(-z)} \int_z^{+\infty} \exp\left(-\frac{y^2}{2}\right) \int_{-\infty}^{+\infty} x^2 \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx dy \\ &= \quad \frac{(2\pi(1-\rho^2))^{-1/2}}{(2\pi)^{1/2}\Phi(-z)} \int_z^{+\infty} \left[\exp\left(-\frac{y^2}{2}\right)\right. \\ &\int_{-\infty}^{+\infty} \left.(x-\rho y+\rho y)^2 \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx\right] dy \\ &= \quad \frac{1}{(2\pi)^{1/2}\Phi(-z)} \int_z^{+\infty} \left[\exp\left(-\frac{y^2}{2}\right)\right. \\ &\int_{-\infty}^{+\infty} \frac{(x-\rho y)^2}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\ &+ \quad \int_{-\infty}^{+\infty} 2\rho y \frac{(x-\rho y)}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\ &+ \quad (\rho y)^2 \left.\int_{-\infty}^{+\infty} \frac{1}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned}$$

Haciendo $u = x - \rho y$

$$\begin{aligned} E(X^2\,|Y \geq z) = \frac{1}{(2\pi)^{1/2}\Phi(-z)} \int_z^{+\infty} &\left[\exp\left(-\frac{y^2}{2}\right)\right. \\ &\int_{-\infty}^{+\infty} \frac{u^2}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \\ &+ 2\rho y \int_{-\infty}^{+\infty} \frac{u}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \\ &+ (\rho y)^2 \left.\int_{-\infty}^{+\infty} \frac{1}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx\right] dy \end{aligned} \tag{A.11}$$

Pero

$$f_U(u) = \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right)$$

es la función de densidad de una varibale $U \sim N(0,(1-\rho^2))$, de manera que

$$\begin{aligned} E(U) &= \int_{-\infty}^{+\infty} \frac{u}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \\ &= 0 \\ E(U^2) &= \int_{-\infty}^{+\infty} \frac{u^2}{(2\pi(1-\rho^2))^{1/2}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) dx \\ &= 1-\rho^2 \end{aligned}$$

y teniendo en consideración

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{(2\pi)(1-\rho^2)}} \exp\left(-\frac{u^2}{2(1-\rho^2)}\right) = 1$$

Se tiene que (A.11) se transforma en

$$E(X^2\,|Y \geq z) = (1-\rho^2) + \frac{\rho^2}{\Phi(-z)} \int_z^{+\infty} y^2 \exp\left(-\frac{y^2}{2}\right) dy$$

Integrando por partes con $u = y$ y $v = -\exp\left(-\frac{y^2}{2}\right)$ se tiene

$$E(X^2\,|Y \geq z) = (1-\rho^2) + \frac{\rho^2}{\Phi(-z)} \left[\frac{z}{\sqrt{2\pi}} + \Phi(-z)\right] \tag{A.12}$$

de donde

$$E(X^2\,|X \geqslant z) = z\frac{\phi(z)}{\Phi(-z)} + 1$$

v. Se usa entonces la función de distribución calculada en el numeral iii (A.10), en consecuencia

$$\begin{aligned} E(X^2\,|X \geq z) &= \int_z^{+\infty} x^2 f_{X|X\geq z}(x)dx \\ &= \frac{1}{\sqrt{2\pi}P(X \geq z)} \int_z^{+\infty} x^2 \exp\left(-\frac{x^2}{2}\right) dx \end{aligned}$$

Integrando por partes con $u = x$ y $v = -\exp\left(-\frac{x^2}{2}\right)$ se tiene

$$E(X^2\,|X \geq z) = \frac{1}{\sqrt{2\pi}P(X \geq z)} \left[z\exp\left(-\frac{x^2}{2}\right) + \int_z^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx\right]$$

Luego usando la notación usual para representar la función de densidad y distribución para la ley normal estándar, y sumado a esto el resultado de $X \sim N(0,1)$, se tiene que

$$E(X^2\,|X \geq z) = z\frac{\phi(z)}{\Phi(-z)} + 1$$

□

ANEXO B

Teoremas Complementarios

Teorema 1 (Teorema de Fubini)**.** *Sea* $A = [a,b] \times [c,d]$*,* $f : A \longrightarrow \mathrm{R}$ *una función integrable en* R^2*, tal que las funciones* $f_x : [c,d] \longrightarrow \mathrm{R}$*,* $f_y : [a,b] \longrightarrow \mathrm{R}$*, definidas por* $f_x(y') = f(x,y')$*,* $f_y(x') = f(x',y)$ *son integrables en* $[c,d]$ *y en* $[a,b]$ *respectivamente, para todo* $(x,y) \in A$*. En consecuencia,*

$$\int_A f = \int_a^b \int_c^d f_x(y)dydx = \int_c^d \int_a^b f_y(x)dxdy$$

lo que es equivalente a

$$\int_a^b \int_c^d f(x,y)dydx = \int_c^d \int_a^b f(x,y)dxdy$$

Observación 1. *Si* f *es continua entonces la funciones* f*,* f_x*,* f_y *(con* $(x,y) \in [a,b] \times [c,d]$*) son todas integrables, y entonces, por el teorema anterior, se tiene que*

$$\int_a^b \int_c^d f(x,y)dydx = \int_c^d \int_a^b f(x,y)dxdy$$

Teorema 2. *Sea* f *continua en* $A = [a,b] \times [c,d]$*, y se define*

$$F(x,y) = \int_a^x du \int_c^y f(u,v)dv$$

Entonces, para $(x,y) \in A$

$$\frac{\partial^2 F}{\partial x \partial y} = \frac{\partial^2 F}{\partial y \partial x} = f(x,y)$$

ANEXO C

Árboles de Decisión

Este anexo contiene los árboles de decisión usados para la construcción de las 27 variables indicadoras en el modelo clásico tratado en el capítulo 2, así como los que fueron utilizados para la creación de las variables indicadoras, con mayor poder discriminatorio en la variable dependiente del modelo probit, usado en el esquema de estimación de la formulación propuesta por Greene en el capítulo 3.

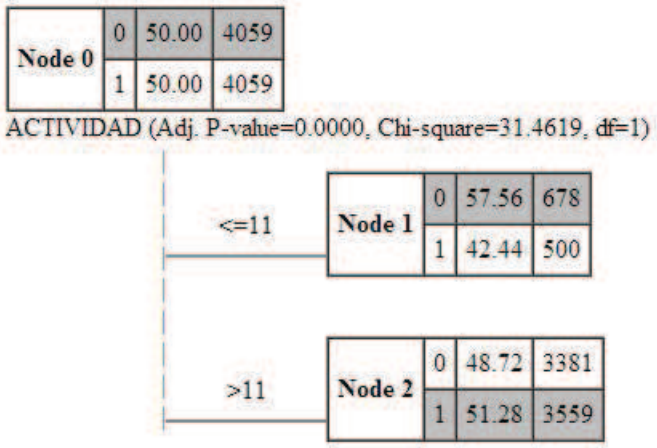

Figura 1. Actividad.

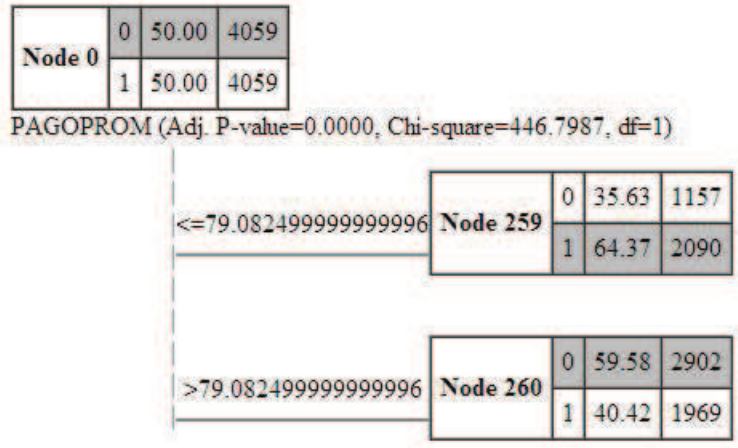

Figura 2. Pago Promedio.

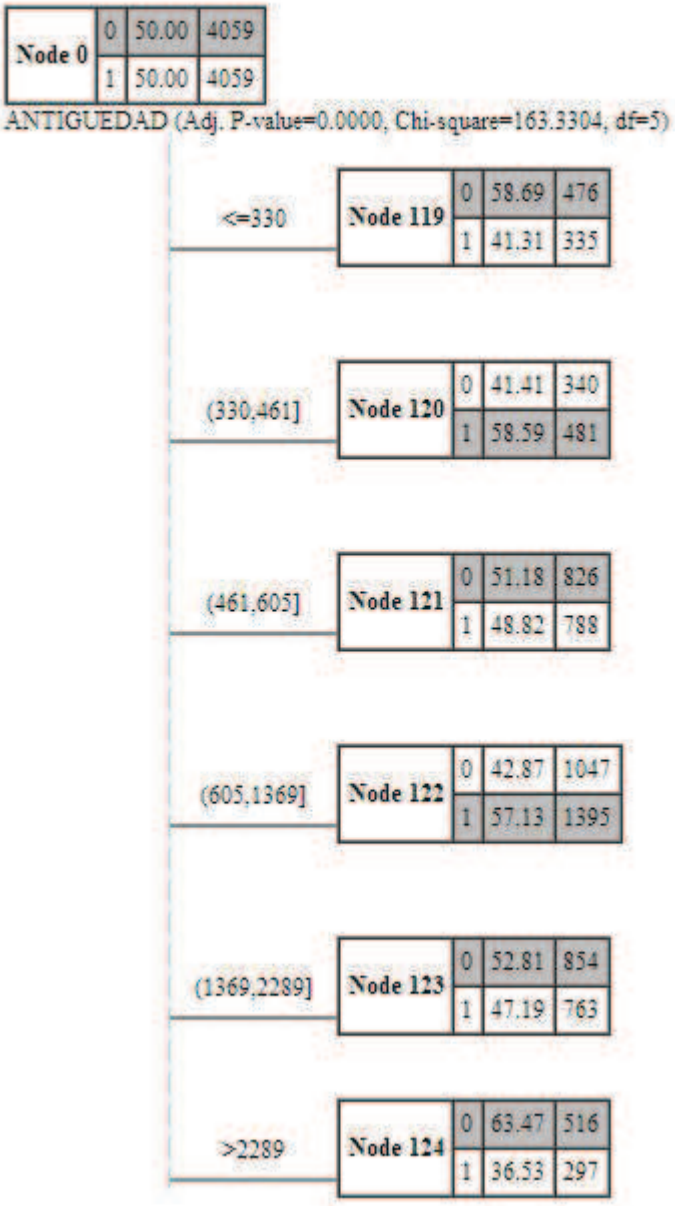

Figura 3. Antiguedad.

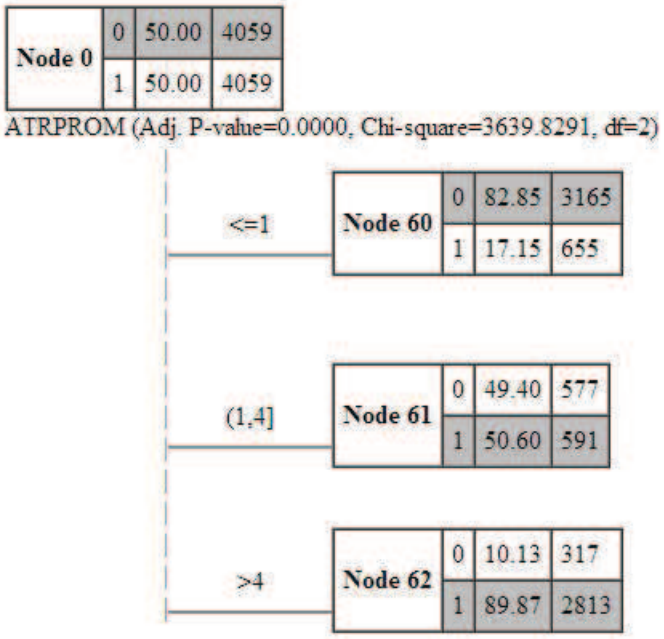

Figura 4. Atraso Promedio.

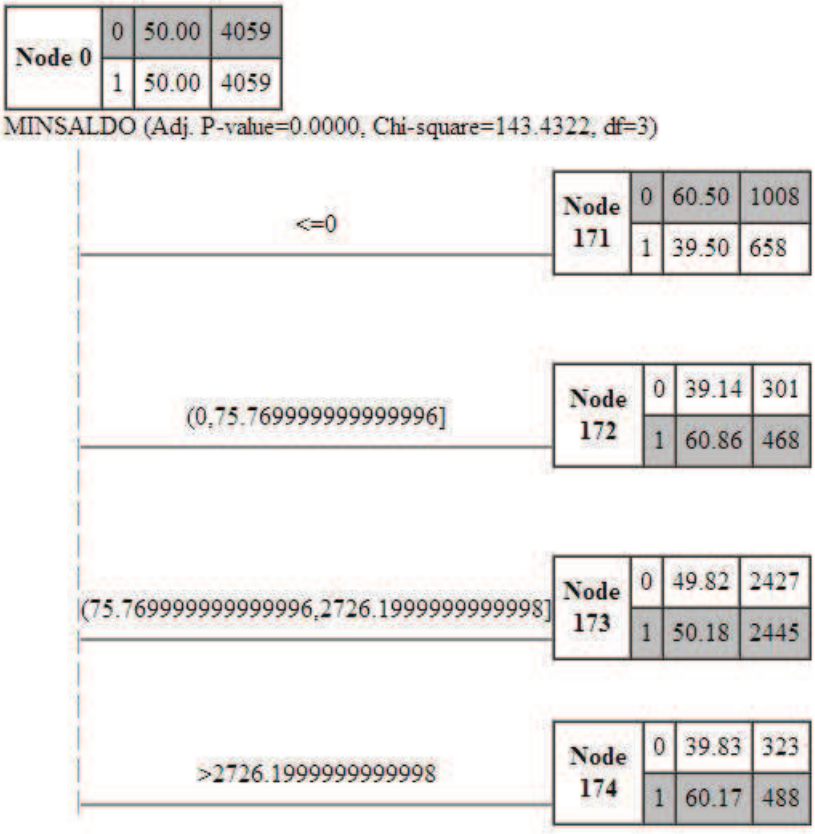

Figura 5. Saldo mínimo.

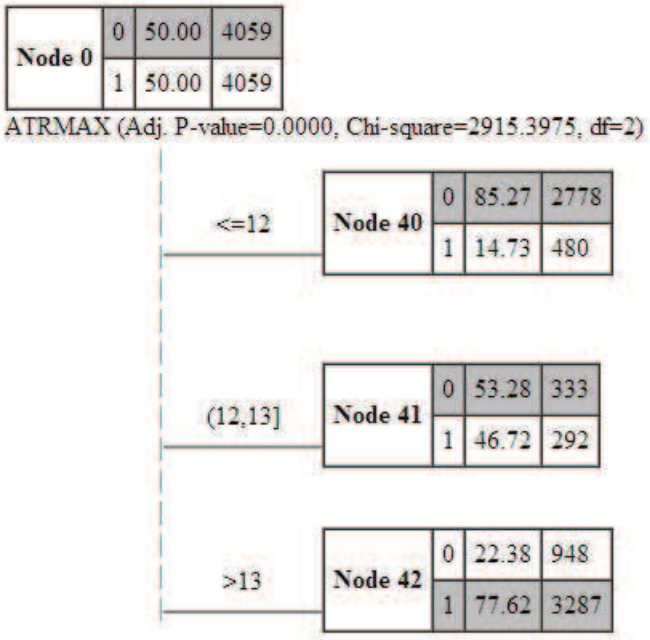

Figura 6. Atraso máximo.

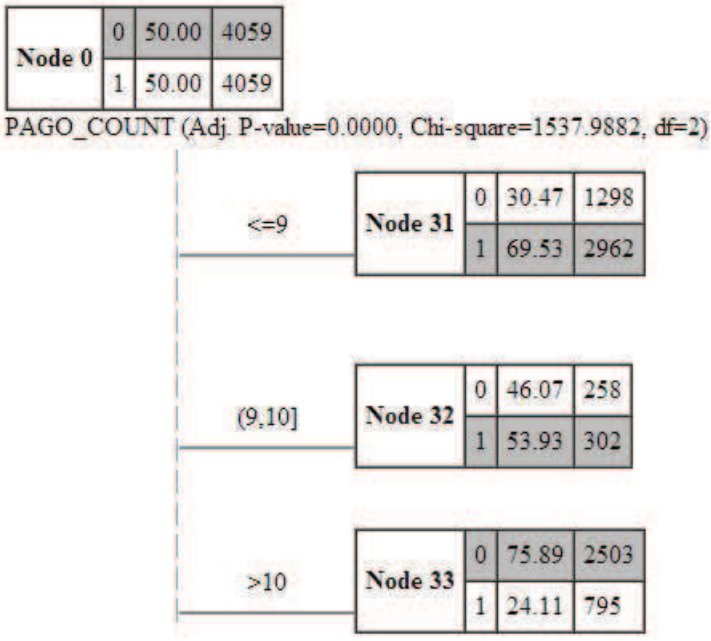

Figura 7. Número de veces que el campo pago fue distinto de cero en el período 1.

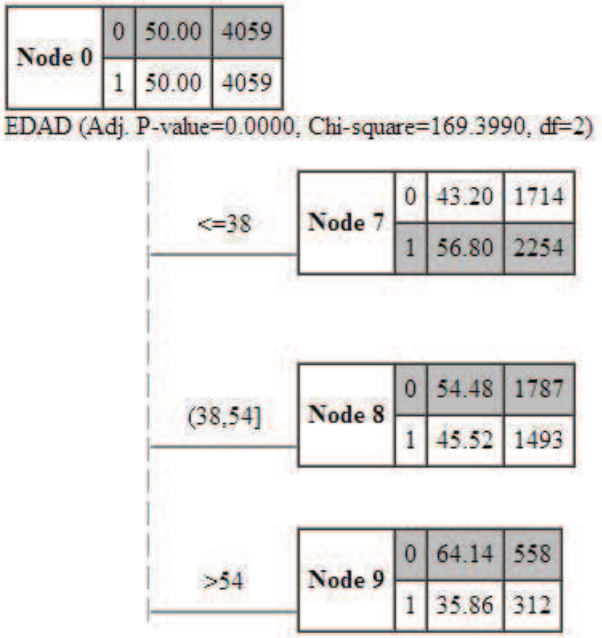

Figura 8. Edad.

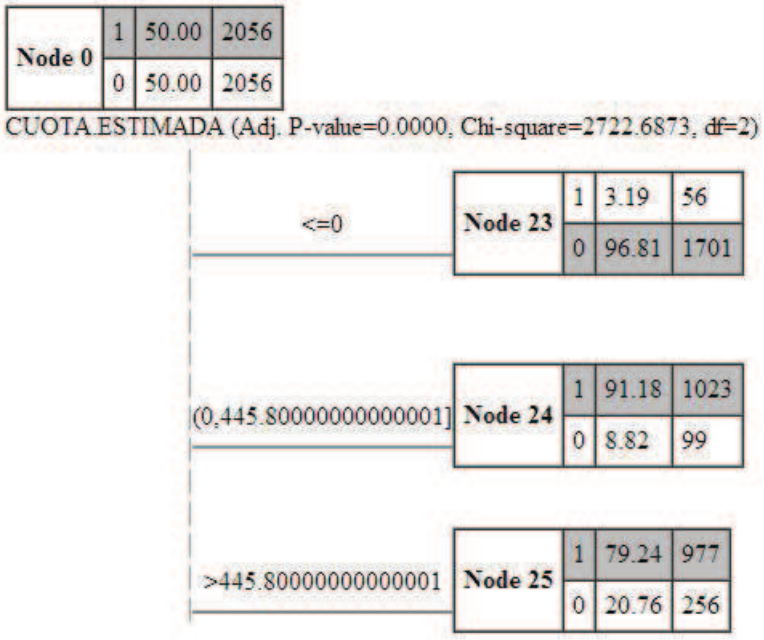

Figura 9. Cuota estimada.

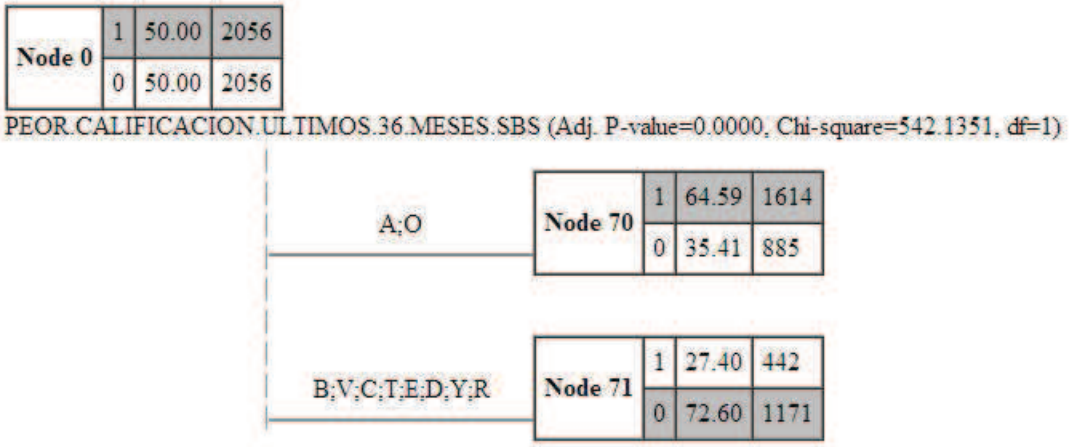

Figura 10. Peor Calificación en los últimos 36 meses SBS.

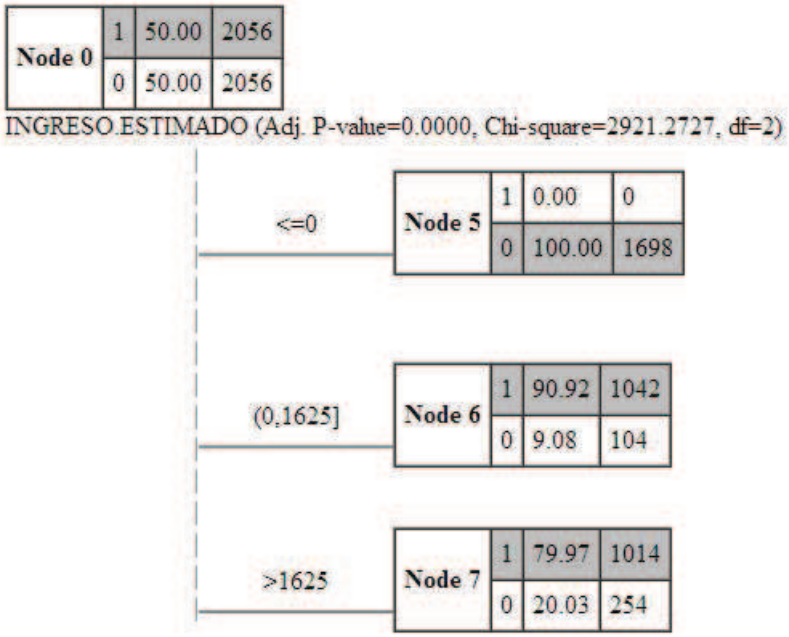

Figura 11. Ingreso Estimado.

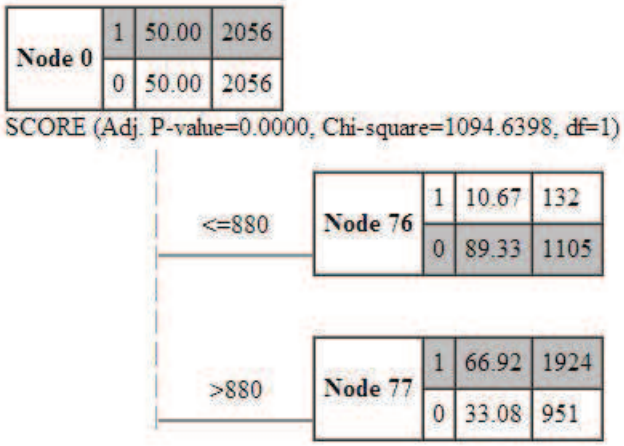

Figura 12. Puntaje dado por el buró de crédito.

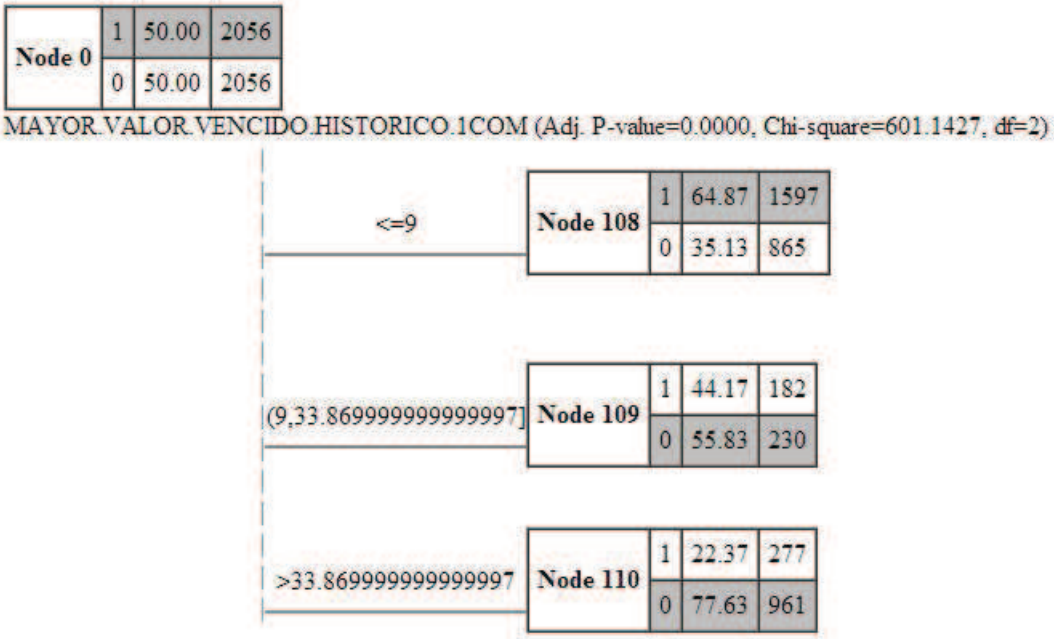

Figura 13. Mayor valor vencido histórico en el sistema comercial.

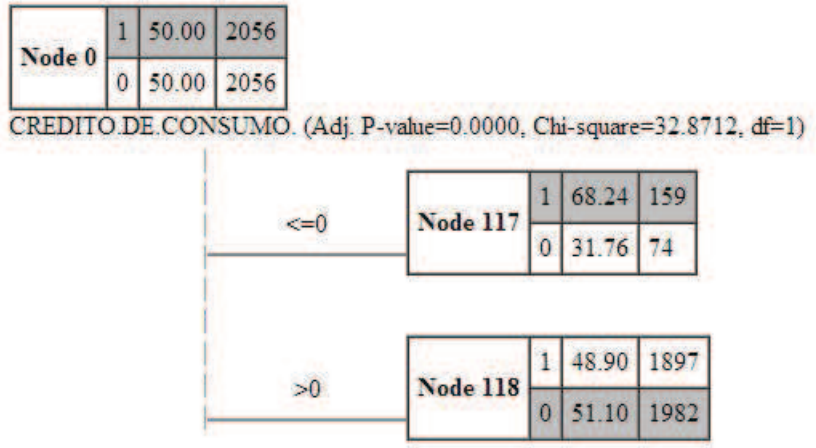

Figura 14. Crédito de consumo.

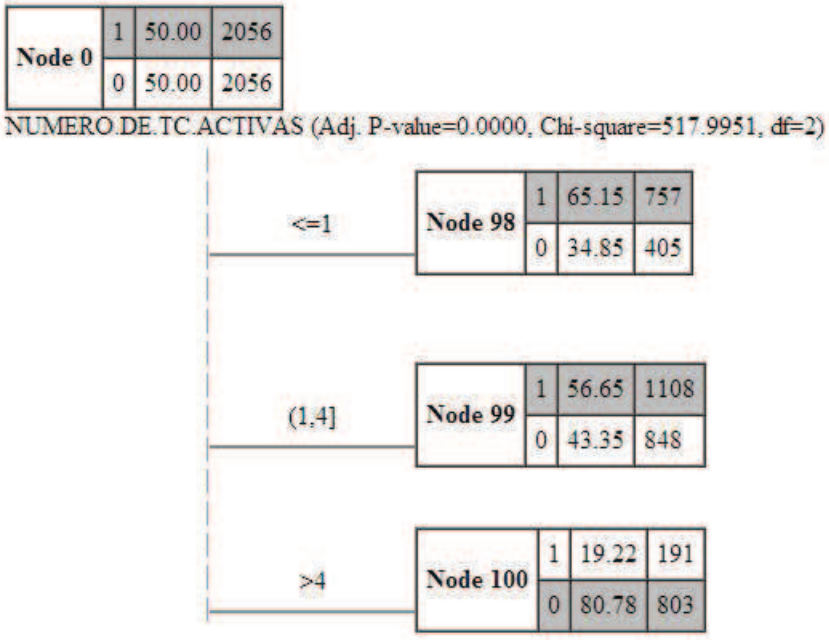

Figura 15. Número de tarjetas activas.

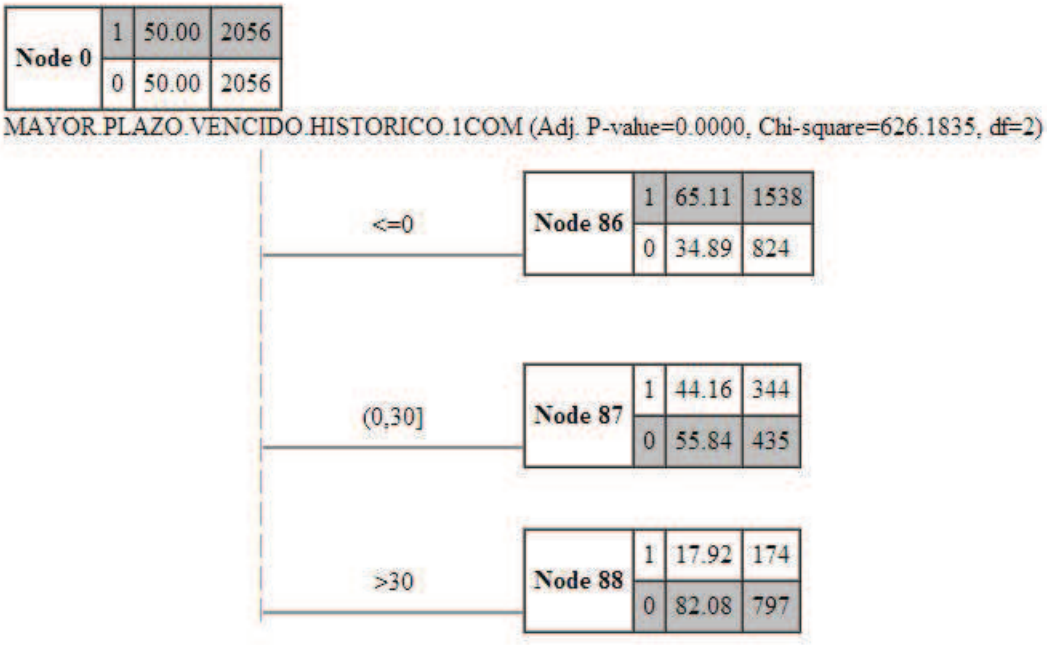

Figura 16. Mayor plazo vencido histórico en el sistema comercial.

ANEXO D

Complementos

D.1. Algoritmo de Construcción de la Variable Atraso

Para la construcción de la variable atraso se partió de las definiciones de mora, pre-Mora y pagos vencidos, variables utilizadas por la institución para calificar el comportamiento de pago de sus tarjetahabientes:

- **Pre-Mora:** Número de días impago contados desde fecha tope de pago hasta la última fecha del mes de referencia.
- **Mora:** Número de días impago contados desde la última fecha del mes anterior hasta la última fecha del mes de referencia del cálculo, sumado el número de días de Pre-Mora y Mora del mes anterior.
- **Pagos Vencidos:** Número de cuotas no pagadas que el cliente mantiene.

Todas estas definiciones aplican siempre y cuando la fecha de pago sea mayor a última fecha del mes que se esté tomando como referencia para el cálculo, caso contrario las variables toman el valor de cero; por otro lado, se tomó en cuenta las reestricciones para las variables descritas:

- Si el campo pre-mora es distinto de cero entonces la mora debe ser cero.
- Si el campo mora es distinto de cero, entonces la pre-mora debe ser cero.

Además se estudió el comportamiento de estas variables mes a mes, durante un año, observando lo siguiente:

- Los campos mora, pre-mora y pagos vencidos toman el valor de cero cuando la tarjeta ha sido cancelada o bloqueada, es decir si no se toma en cuenta el estado de la tarjeta y únicamente se hace referencia a la mora o pre-mora se puede calificar a un cliente con Atraso=mora=0, lo que sería un error.
- Existen clientes con Mora y Pre-Mora iguales a cero y pagos vencidos distintos de cero, de modo que los pagos vencidos deben ser tomados en cuenta para calcular la variable Atraso.

Tomando en cuenta entonces lo mencionado anteriormente, se define el algoritmo para el cálculo de la variable *Atraso*, condensando así la mora, pre-mora, pagos vencidos y el estado de las tarjetas, como sigue:

Algoritmo 1 Algoritmo variable Atraso

Si (Mora = 0 & PagosVencidos = 0 & PreMora != 0) **entonces**
Atraso = PreMora.
Si (Mora!= 0) **entonces**
Atraso = Mora.
Si ((Mora = 0 & Pagos Vencidos = 0 & PreMora = 0) & (EstadoTarjeta = Activa o EstadoTarjeta = Sobregirada Activa)) **entonces**
Atraso = 0 **caso contrario**
Atraso = NULL.
Si (PagosVencidos != 0 & Mora = 0 & PreMora = 0) **entonces**
Atraso = PagosVencidos*30.

D.2. Prueba CUSUM

La prueba CUSUM parte del modelo de regresión lineal

$$y_i = \beta' \mathbf{x}_i + u_i \ \ (i = 1, \ldots, n) \tag{D.1}$$

donde $x_i = (1, x_{1i}, \ldots, x_{ki})$ y $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$.Esta prueba se basa en el cálculo de la suma acumulada de residuos estandarizados $(\tilde{u}_i)$ asociados a este modelo, la cual está dada por la siguiente expresión:

$$W_n(t) = \frac{1}{\hat{\sigma}\sqrt{\eta}} \sum_{i=k+1}^{k+\lfloor t\eta \rfloor} \tilde{u}_i \ \ (0 \leq t \leq 1) \tag{D.2}$$

donde, $\eta = n - k$ es el número de residuos recursivos, $\lfloor t\eta \rfloor$ es la parte entera de $t\eta$ y

$$\tilde{u}_i = \frac{y_i - \hat{\beta}^{(0,i-1)'} \mathbf{x}_i}{\sqrt{1 + \mathrm{x}_i'(\mathbf{X}^{(0,i-1)'}\mathbf{X}^{(0,i-1)})\mathbf{x}_i}} \ \ (i = k+1, \ldots, n) \tag{D.3}$$

siendo $\hat{\beta}^{(i,j)}$ el estimador de mínimos cuadrados ordinarios de los coeficientes de D.1 basados en las observaciones $i+1, \ldots, i+j$, $\mathbf{X}^{(i,j)}$ la matriz de regresores desde $i+1$ hasta $i+j$ y $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=k+1}^{n}(\tilde{u}_i - \bar{\tilde{u}}_i)$.

A través de $W_n(t)$ se puede contrastar $H_0 : \beta = \beta_0$ contra $H_1 : \beta \neq \beta_0$; es decir, si los coeficientes de D.1 se mantienen constantes a lo largo del tiempo (no existe cambio de estructura). Para ello se define el estadístico

$$S = \sup_{0 \leq t \leq 1} \left| \frac{W_n(t)}{1 + 2t} \right| \tag{D.4}$$

Si $S > \lambda$, donde λ es un parámetro que depende del nivel de significación de la prueba, se rechaza H_0 (para mayor información sobre los valores de λ ver **[12]**). S se conoce como el estadístico asociado a la prueba CUSUM.

Cabe recalcar que bajo la hipótesis nula el proceso $W_n(t)$ tiene media cero; es decir, si existe un cambio estructural en el período t_0, el proceso tendría media cero hasta ese punto y por lo tanto, a partir de ahí debería de alejarse de cero conforme $t > t_0$, de modo que debería de rechazarse H_0 si así fuera. Bajo este supuesto se puede graficar el proceso $W_n(t)$ y ver el instante donde el fenómeno tiene lugar, siendo éste un criterio gráfico para detectar un cambio estructural. Paquetes estadísticos como

R presentan tanto el criterio gráfico, así como el valor p asociado al estadístico de prueba.

D.3. Pruebas de Bondad de Ajuste

D.3.1. Prueba de Kolmogorov-Smirnov. Esta prueba se basa en una comparación entre las funciones de distribución acumulativa que se observan en la muestra ordenada y la distribución propuesta bajo la hipótesis nula de que ambas son iguales. Si esta comparación revela una diferencia suficientemente grande entre las funciones de distribución muestral y propuesta, entonces la hipótesis nula se rechaza.

Sea $S_n(x)$ la proporción del número de valores en la muestra que son menores o iguales a x y $F_0(x)$ la proporción del número de valores en la población que son menores o iguales a x. Ya que $F_0(x)$ se encuentra completamente especificada, es posible evaluar a $F_0(x)$ para algún valor deseado de x, y entonces comparar este último con el valor correspondiente de $S_n(x)$. Si la hipótesis nula es verdadera, entonces se espera que la diferencia sea relativamente pequeña. El estadístico asociado a esta prueba se define como

$$D_n = \text{máx}\,|S_n(x) - F_0(x)|$$

Para un nivel de confianza α y un tamaño de muestra n se tiene que la probabilidad crítica o valor p es

$$p = P\left(D_n > \frac{c}{\sqrt{n}}\right) = \alpha$$

donde la hipótesis nula se rechaza si para algún valor x observado, el valor de D_n se encuentra dentro de la región crítica de tamaño α.

D.3.2. Prueba Ji-cuadrado. A diferencia de la prueba anterior se supondrá que las observaciones de la muestra están agrupadas en k clases, siendo n_i la cantidad de observaciones en cada clase $i = 1, \ldots, k$. Con el modelo especificado $f_0(x)$ se puede calcular la probabilidad p_i que una observación cualquiera pertenezca a una clase i. Con este valor de probabilidad se puede encontrar la frecuencia esperada e_i para la clase i, es decir, la cantidad de datos que según el modelo especificado deberían estar incluidos en la clase i:

$$e_i = n_i p_i$$

Donde, n_i es la frecuencia observada (correspondiente a los datos de la muestra) y e_i es la frecuencia esperada (correspondiente al modelo propuesto). Al igual que la prueba anterior, diremos que si la hipótesis nula es verdadera la diferencia entre estas frecuencias será pequeña. El estadístico para la prueba Chi-cuadrado se define como

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

El estadístico χ^2, bajo el supuesto de la hipótesis nula, sigue asintóticamente una distribución ji-cuadrado con $k-1$ grados de libertad (χ_{k-1}), donde se rechaza ésta hipótesis cuando $\chi^2 > \chi^2_\alpha$, siendo χ^2_α el fractil de nivel α de la ley χ_{k-1}.

D.4. Prueba F

La prueba F mide que tan significante es el modelo de regresión lineal

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_{ik} \text{ con } i = 1, \ldots, n \tag{D.5}$$

Cuando se estiman los coeficientes β_i con el método de mínimos cuadrados ordinarios (MCO). La prueba contrasta las hipótesis:

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = \cdots = \beta_k = 0 \\ H_1 &: \exists \beta_j, j \geq 2, \beta_j \neq 0 \end{aligned}$$

Es decir y_i no depende linealmente de sus regresores (H_0) frente a y_i depende linealmente de al menos un regresor (H_1). El estadístico que se usa para el contraste de éstas dos hipótesis se denomina razón F y viene dado por la siguiente expresión:

$$F = \left(\frac{n-k}{k-1}\right)\frac{SEC}{SRC}$$

Donde SEC es la suma de estimaciones al cuadrado y SRC es la suma de residuos al cuadrado, que vienen dadas por las expresiones

$$\begin{aligned} SEC &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \\ SRC &= \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \end{aligned}$$

Siendo $\hat{y}_i$ las estimaciones del modelo lineal utilizando MCO y $\overline{y}$ es el promedio de las observaciones y_i. Se rechaza entonces la hipótesis nula si

$$F > F_{(k-1,n-k)}(\alpha)$$

Donde $F_{(k-1,n-k)}(\alpha)$ el fractil de la ley de Fisher de nivel α con $(k-1, n-k)$ grados de libertad. Los resultados anteriores se suelen escribir en un cuadro denominado ANOVA (análisis de varianza), que tiene la estructura del cuadro 1, donde

$$STC = SEC + SRC$$

Además S.C. es la suma de cuadrados, M.S.C. la media de la suma de cuadrados y el Valor-p o probabilidad crítica es otro criterio para tomar una decisión, donde se rechaza la hipótesis nula cuando este valor es menor que el nivel α fijado, generalmente 0.01, 0.05 ó 0.1.

Fuente	**G. L.**	**S.C.**	**M.S.C.**	**F**	**Valor−p**
Regresión	$k-1$	SEC	$MSE = \frac{SEC}{k-1}$	$\frac{MSEC}{MSRC}$	p
Error	$n-k$	SRC	$MSE = \frac{SRC}{n-k}$		
Total	$n-1$	STC			

Cuadro 1. Tabla ANOVA