X-Pedientes Económicos, Vol. 4 (8), Enero – Abril 2020, pp. 49-62

Análisis de los factores determinantes de la asignación de riesgo financiero y crediticio de la pequeña banca privada del Ecuador

Analysis of the determining factors of the allocation of financial and credit risk of small private banks in Ecuador

Jorge García Regalado[1]†, Octavio Rugel [2]

Fecha de recepción: 11/03/2020, Fecha de aceptación: 30/04/2020

RESUMEN

El propósito de la clasificación de los clientes crediticios en una institución financiera, puede convertirse en un problema de maximización de ganancia o minimización de la pérdida. Sin embargo, para solucionar el problema, además de la experiencia del especialista en crédito, es necesario disponer de herramientas que faciliten una aproximación a las condiciones reales del cliente al momento de otorgar un crédito. Los Modelos de Puntaje o "Scoring" crediticio son una de las herramientas más utilizadas en la actualidad.

Palabras clave: Modelos Logit; riesgo crediticio; puntuación de crédito.

ABSTRACT

The purpose of the classification of credit clients in a financial institution can become a problem of profit maximization or loss minimization. However, to solve the problem, in addition to the experience of the credit specialist, it is necessary to have tools that facilitate an approximation to the real conditions of the client when granting a loan. The Credit Scoring or Scoring Models are one of the most used tools today.

Keywords: Logit models; credit risk; credit score

[1] Universidad Agraria del Ecuador, Contacto: jgarcia@uagraria.edu.ec
[2] Universidad Agraria del Ecuador, Contacto: drugel@uagraria.edu.ec
† *Autor de correspondencia*

I. INTRODUCCIÓN

Los Modelos de Puntaje o "Scoring" crediticio son una de las herramientas más utilizadas en la actualidad. El Scoring es una metodología estadística-econométrica que asigna en rangos, la probabilidad de un resultado desconocido al otorgar puntajes a variables conocidas. El resultado final es un puntaje "score", que permite categorizar a los solicitantes de crédito en términos de su riesgo, como apoyo para determinar la aprobación crediticia.

Un modelo adecuado se proporcionará en el presente trabajo; se presentan los resultados del proceso de modelación para otorgamiento de crédito de altos puntajes a créditos de buen desempeño y bajos puntajes a créditos de alto riesgo de incumplimiento en el de consumo general. En la primera fase, se analizará las características de la población de clientes mediante análisis descriptivo, para luego, en la segunda fase, desarrollar y probar el modelo de otorgamiento crediticio.

Antecedentes Metodológicos del Modelo Estadístico

Los scorecards son desarrollados bajo el supuesto de que el comportamiento futuro está reflejado en el comportamiento pasado. En otras palabras, es posible predecir con un grado adecuado de efectividad el comportamiento futuro de las personas que reciben un crédito en base a su comportamiento pasado y sus características socio-económicas y demográficas al momento de la concesión del crédito.

El principio básico del modelo es simular el proceso de entrega del crédito, en el sentido de incorporar la información con que contaría el evaluador del banco al momento de recibir la solicitud y analizar la idoneidad del cliente. Por esta razón, no deberían incorporarse en el modelo variables que el evaluador desconocería al momento de empezar el proceso de decisión de la concesión del crédito.

Así, se decidió usar para el desarrollo del modelo una muestra información crediticia de 30 mil personas que hayan obtenido un crédito entre agosto de 2012 y febrero 2018. El periodo fue elegido tomando en cuenta que solo desde agosto de 2012 se cuenta con información socio-económica y demográfica completa. La fecha límite fue elegida de tal forma que se cuente con un periodo de al menos un año para observar el comportamiento de estos clientes, al tiempo que se deja un tiempo prudencial para generar muestras de validación recientes.

En este estudio se utilizó el modelo logístico para la obtención de la probabilidad de incumplimiento. Por cuanto estos modelos son los de mayor utilización dentro de la industria financiera, lo cual permite utilizar estas experiencias como puntos de partida en nuestras decisiones metodológicas y realizar comparaciones, especialmente en cuanto a la calidad de las predicciones. Los modelos logit tienen además la ventaja de fácil interpretación de los coeficientes obtenidos, acompañados de una fácil aplicación de metodologías de post estimación que permiten apreciar las bondades del modelo elaborado.

Caracterización de modelos logit

Los modelos de logit siguen la siguiente forma funcional:

$$P(y = 1|\boldsymbol{x}) = P(y = 1|x_1, x_2, \dots, x_k) \quad (1)$$

En donde x representa un vector de características del individuo que pueden ser edad, años de escolaridad, área de hábitat, entre otras. Dentro de este tipo de modelos el interés de análisis se encuentra en la respuesta de probabilidad. La diferenciación de estos modelos se encuentra en la función de respuesta binaria que siguen:

$$P(y = 1|\boldsymbol{x}) = G(\beta_0 + \beta_1 x_1 + .. + \beta_k x_k) = G(\beta_0 + \boldsymbol{x\beta}) \quad (2)$$

En donde la función $G2$ adquiere valores estrictamente entre cero y uno:

$0 < G(z) < 1$, para todos los números reales. Esta función asegura que las respuestas probabilísticas estimadas se encuentran estrictamente en el rango mencionado. Varias funciones no lineales han sido sugeridas para que la función G tome los valores estrictamente dentro del rango en análisis.

Para el modelo logit la función logística sigue la siguiente forma:

$$G(z) = exp(z)/[1 + exp(z)] = \Lambda(z) \quad (3)$$

Esta función toma valores entre cero y uno para todos los números reales z. Esta es la función de distribución acumulada para una variable aleatoria logística.

La selección de la función [ecuación (3)] asegura que (2) toma valores estrictos entre cero y uno para todos los valores de los parámetros en el vector x. Las función (3) es creciente y crece más rápidamente cuando $z = 0, G(z) \to 0$ cuando $z \to -\infty$, y, $G(z) \to 1$ cuando $z \to \infty$.

Estimación de máximo verosimilitud

La estimación de los modelos logit se la realiza mediante la estimación de máximo verosimilitud (MVS), la misma sigue el mismo principio que la metodología de mínimos cuadrados ordinarios. Este tipo de estimación se da porque el valor esperado de y dado x,

$E(y|\boldsymbol{x})$, no es lineal. Para la estimación de este tipo de modelos, la estimación MVS es indispensable, dado a que la estimación MVS se encuentra basada en la distribución de y dado $\boldsymbol{x}$, la heterocedasticidad en $Var(y|\boldsymbol{x})$ es automáticamente calculada.

Asumiendo que se encuentra en análisis una muestra aleatoria de tamaño n, para obtener el estimador MVS, condicional en las variables independientes, se necesita obtener la densidad de y_i dado $\boldsymbol{x_i}$, se determina como:

$$f(y|\boldsymbol{x}_i; \boldsymbol{\beta}) = [G(\boldsymbol{x}_i\boldsymbol{\beta})]^y[1 - G(\boldsymbol{x}_i\boldsymbol{\beta})]^{1-y}, y = 0,1 \quad (4)$$

En donde por simplicidad, se absorbe el intercepto de la regresión dentro del vector$x$. Fácilmente se puede apreciar que cuando $y = 1$ se obtiene $G(x_i\beta)$ y cuando $y = 0$ se obtiene $1 - G(x_i\beta)$. La función de logarítmica de verosimilitud para cada observación i es una función de los parámetros y de los datos (x_i, y_i) y se la obtiene tomando el logaritmo de (4):

$$\ell_i(\boldsymbol{\beta}) = y_i \log[G(\boldsymbol{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\boldsymbol{x}_i\boldsymbol{\beta})] \quad (5)$$

Dado que $G(\cdot)$ toma valores estrictamente entre cero y uno para la función logística y probabilística, $\ell_i(\beta)$ se encuentra bien definido para todos los valores de β. La log-verosimilitud para una muestra de tamaño n se la obtiene sumando (9) entre todas las observaciones $\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell_i(\beta)$. El estimador de máximo verosimilitud de β, denotado por $\hat{\beta}$, maximiza el logaritmo de la verosimilitud. Si $G(\cdot)$ toma sigue la función logística entonces $\hat{\beta}$ es el estimador logit.

En el caso de no cumplimiento del supuesto de homocedasticidad, la estimación de varianza y error estándar no es eficiente e insesgada. Cuando los errores sufren de heterocedasticidad, $Var(\varepsilon_i|x_i) = \sigma_i^2$, por lo que el estimador de mínimos cuadrados ordinarios se lo escribe como:

[2] La función G corresponde a la función logística

$$\widehat{\beta_\iota} = \beta_i + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

En donde:

$$Var(\widehat{\beta_\iota}) = \beta_i + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sigma_i{}^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \qquad (6)$$

Dado que el error estándar del coeficiente se basa en la estimación de la varianza del mismo, se debe estimar la ecuación (12) ante la existencia de heterocedasticidad. White (1980) demostró cómo hacer dicha estimación ante la existencia de heterocedasticidad, utilizando los residuos como un estimador válido de $\mathrm{Var}(\widehat{\beta_1})$, el mismo que se detalla:

$$\widehat{\beta_\iota} = \beta_i + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{\varepsilon}_i{}^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \qquad (7)$$

Para la aplicación en modelos de Scoring (los mismos que son modelos de regresión múltiple), el estimador válido $\mathrm{Var}(\widehat{\beta_1})$

$$\widehat{Var(\beta_J)} = \beta_i + \frac{\sum_{i=1}^{n}\hat{r}_{ij}^2\hat{\varepsilon}_{ij}^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \qquad (8)$$

En donde $\widehat{r_{iJ}}$corresponde al i-ésimo residuo de regresar x_j sobre todas las demás variables independientes. Es decir el residuo de la regresión múltiple es $x_j = \delta_0 + \sum_{h\neq j}^{n}\delta_h x_h + \xi$. La raíz cuadrada de (8) corresponde al error estándar robusto a la heterocedasticidad para su respectivo coeficiente, magnitud con la cual se cumple el supuesto de homocedasticidad.

Transformación de datos e Implementación del Modelo Logit

Si bien este sistema supera en confiabilidad, flexibilidad, y muchas veces en poder predictivo a los típicos métodos automatizados de selección de variables (step wise selection methods), poseen, por otro lado, la desventaja de que es más difícil de sistematizar.

Esto último, en el sentido de que es más complejo documentar (y peor aún reportar) la innumerable cantidad de combinaciones y *trials and errors* que se implementaron. Esta falta de sistematización es vista con sospechas por algunos investigadores que quieren saber exactamente cómo se llegó a ese modelo y cómo podemos estar seguros de que es el mejor.

Finalmente, es importante subrayar que este reporte no presentará determinación de puntos de corte por cuanto esto es una decisión que corresponderá a las entidades financieras en base los objetivos y características de sus negocios. No obstante que más adelante se presentaran distribuciones de scores que contaran con información valiosa para ese tipo de decisiones.

II. METODOLOGÍA

Definición e identificación de casos especiales

Nuestra metodología incluye la división de la población bajo estudio en dos segmentos, que dependen de si el número de meses durante el cual el cliente ha tenido una relación transaccional con la institución (tiempo IFI) es mayor o menor a seis meses. Como se verá más adelante, las variables empleadas para cada segmento no

son exactamente las mismas, producto principalmente de que el menor número de observaciones en el segmento "reciente" hace que muchas de ellas resulten no significativas.

En el caso extremo de que una persona no haya tenido ninguna experiencia crediticia anterior, sus únicos predictores serían las variables demográficas con que cuente. Así también, se crearon dos variables adicionales que identifican si la persona tuvo experiencias crediticias o de uso de tarjetas de crédito. Se imputo valor cero en las variables que estas personas no presentaban (al no tener historias).

Análisis bivariado

En este acápite se reportan una serie de gráficos que resumen la relación entre algunas de las variables más importantes usadas en nuestro modelo predictivo y la variable dependiente (malos clientes). Allí vemos por ejemplo que:

- Los hombres tienen en promedio un punto porcentual adicional de riesgo de default.
- El riesgo se reduce considerablemente con la edad.
- Las mayores diferencias entre hombres y mujeres se encuentran en el tramo de edades menores a 31 años.
- La probabilidad de no pago es sustancialmente más alta entre los que reportan no tener estudios. Esta probabilidad disminuye en la medida que el nivel de estudios es más alto, excepto para la categoría universitarios que es más alta que la de los técnicos y similar al de los que reportan estudios secundarios.
- Cuando introducimos las diferencias por género al análisis por niveles educativo, es notorio que las mayores diferencias están entre aquellos sin educación y los que reportan educación técnica.
- Los solteros tienen más del doble de riesgo de default que los casados.
- El riesgo disminuye con el número de cargas familiares y el tiempo como cliente IFI.
- Los trabajadores independientes y empleados privados presentan el mayor riesgo de default, caso contrario al de las amas de casa y quienes reciben remesas del exterior.

Los datos indican que tanto los hombres como mujeres tienen un grado de ser clientes que se consideran malos. Sin embargo, son las personas con género masculino quienes están más propensos a cancelar sus obligaciones posteriores a las fechas de pago.

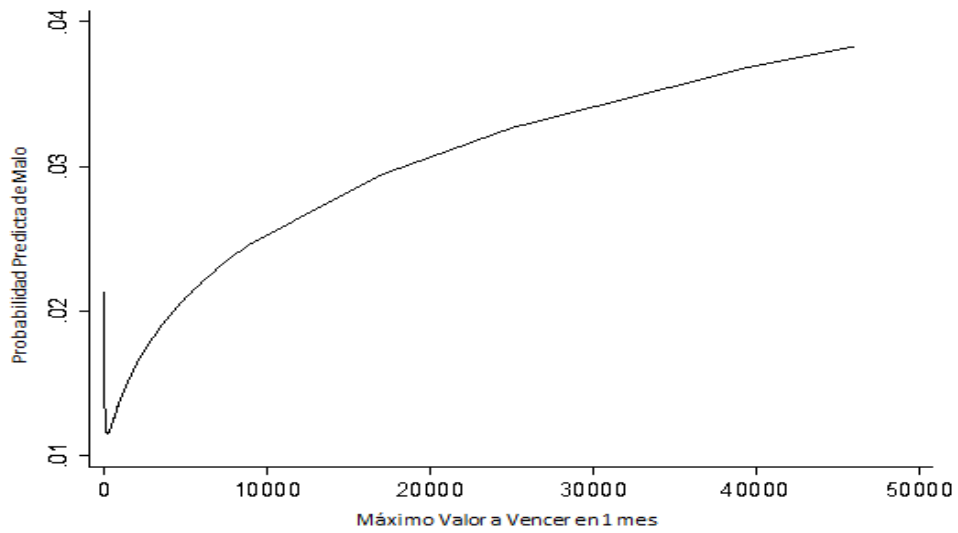

Figura 1: Mora por valor a vencer en un mes
Fuente: Base de datos Superintendencia de Bancos

Existe alto grado de probabilidad de ser un mal cliente cuando a un mes de vencimiento del crédito, el valor a vencer es mayor. En otras palabras, mientras más valor a vencer se enfrente el consumidor del crédito en un mes, es más probable que sea considerado como un mal cliente. Dicha situación se explica por el hecho de que los altos valores de créditos exigen liquidez suficiente para cumplir con dichas obligaciones, tal como se muestra en la Figura 1, donde se aprecia una curva que indica el comportamiento de la probabilidad de ser mal consumidor a medida que el valor a vencer a un mes sea superior

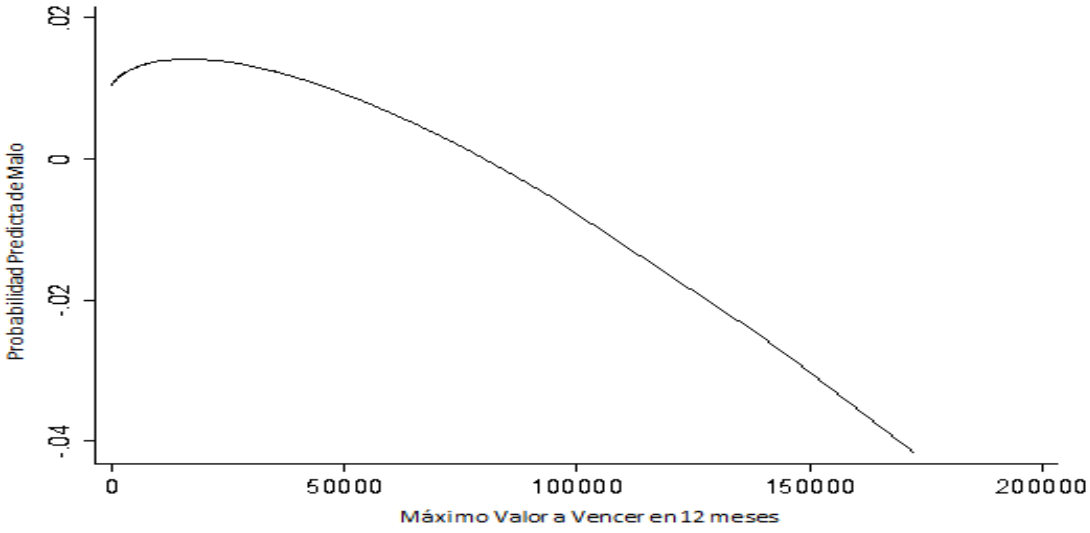

Figura 2: Mora por valor a vencer en doce meses
Fuente: Base de datos Superintendencia de Bancos

Por el contrario, existe un bajo grado de probabilidad de ser un mal cliente cuando a doce meses de vencimiento del crédito, el valor a vencer es mayor. En otras palabras, mientras más valor a vencer se enfrente el consumidor del crédito en doce meses, es más probable que sea considerado como un buen cliente. Sin

embargo, dentro de la muestra existen sujetos bajo estudio que, a pesar de tener tiempo para cumplir con sus deudas, no logran cumplirlas dentro del plazo previamente pactado. En la Figura 2 se aprecia una curva que indica el comportamiento de la probabilidad de ser mal consumidor a medida que el valor a vencer a un mes sea superior.

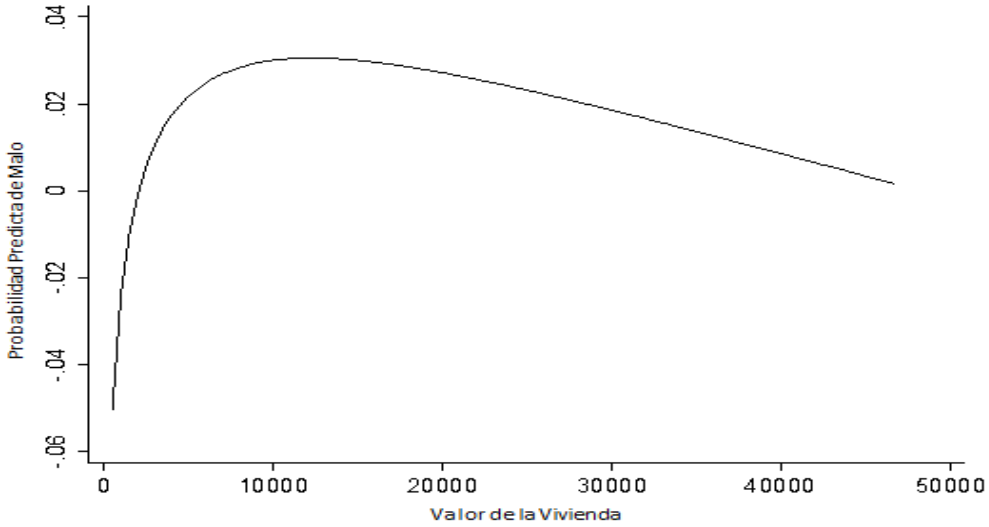

Figura 3. Comportamiento de clientes según el valor de su vivienda
Fuente: Base de datos Superintendencia de Bancos

Los datos indican que aquellos clientes que se consideraron malos corresponden en su mayoría a personas con empleos privados, amas de casas, estudiantes. Existe un bajo nivel del impago de los créditos dentro del plazo determinado que caracteriza al buen cliente en las personas del sector público, trabajadores autónomos, rentistas, jubilados y quienes reciben remesas.

Los datos indican que tanto los hombres como mujeres tienen un grado de ser clientes que se consideran malos. Sin embargo, para efectos del presente estudio de personas con una relación transaccional menor a 6 meses con una IFI, son las personas con género femenino quienes están más propensos a cancelar sus obligaciones posteriores a las fechas de pago. Por otro lado, podemos observar que, según el origen del ingreso, existe una alta frecuencia de ser consideradas como buenos clientes aquellas personas que han mantenido una relación transaccional mayor a los 6 meses con una IFI (Ver Figura 3).

III. RESULTADOS

En el ámbito de la modelación estadística existen dos tipos fundamentales de análisis. En el primero, se trata de explicar el resultado y sus determinantes, por lo cual juega un rol trascendental el valor de los parámetros estimados (típicamente coeficientes en una regresión) y sus características inferenciales. Es el caso de modelos económicos canónicos como el de Mincer y su retorno a la educación, o el famoso beta financiero utilizado en valoración de activos.

En otros casos, se da menos preponderancia al marco teórico detrás de los modelos estadísticos y lo que mayormente importa es la calidad o confiabilidad de las predicciones; sean estos, segmentos que comparten características comunes (clústeres), o valores de variables que distinguen de forma óptima distintos tipos de entidades (árboles de clasificación, por ejemplo).

En nuestro caso, si bien se trató en lo posible de incorporar variables y explicaciones que tengan sentido desde la óptica financiera/económica, se dio especial preponderancia a las cualidades predictivas del modelo en conjunto. Así, el principio rector detrás de las modelaciones que vamos a presentar, es el de entregar información relevante para que los usuarios del sistema puedan efectivamente separar malos de buenos clientes. De esta forma se mejora la eficiencia en la asignación de créditos y se disminuye al mismo tiempo el riesgo global del sistema financiero, con lo cual se benefician no solo dichas instituciones sino el país como un todo.

Algunas ventajas en cuanto a generación de ingresos y, sobretodo, potenciales gastos. Es una edad de consolidación en el mercado laboral y donde aún no se tiene demasiadas responsabilidades familiares, consecuentemente las barreras para poder cumplir con los pagos se reducen (al menos en relación al segmento más joven). Así también, se obtuvo que personas con educación superior presentan un mejor perfil de riesgo. Esto probablemente tiene que ver con su mayor capacidad de generación de ingresos, menor probabilidad de caer en desempleo, e incluso mejores capacidades organizacionales para evitar o resolver eventuales retrasos en los pagos.

En el caso del estado civil la interpretación es más compleja por las interacciones con el lugar de residencia. El resultado más claro, considerando lo presentado anteriormente en el análisis bivariado, es que casado y viudo son estados civiles asociados con menor riesgo de default, especialmente si esas personas viven en provincias diferentes a las ya mencionadas.

Demográficas	**Odds-ratios**
Edad	
18 - 29	(base)
30 – 33	0.543
	[0.104]***
>34	0.784
	[0.104]*
Universitario	0.593
	[0.068]***
Casado / Viudo	3.078
	[1.288]***
No Guayas & Casado/Viudo	0.470
	[0.105]***
Guayas & No Casado/Viudo	1.978
	[0.272]***
No Manabí & Casado/Viudo	0.453
	[0.164]**
Manabí & No Casado/Viudo	1.776

	[0.416]**
Socioeconómicas	
Ama de Casa / Estudiantes	2.282
	[1.098]*
Entretenimiento / S. Privado	1.231
	[0.150]*
Tiempo en Vivienda	
1 año	(base)
<1 año	0.251
	[0.087]***
2 -6 años	0.627
	[0.133]**
> 6 años	0.360
	[0.065]***
Vivienda Propia	0.486
	[0.092]***
Valor Vivienda>50.000	0.125
	[0.127]**
Financieras	
log(IFI) ≥ 2.75	0.628
	[0.105]***
log(IFI) < 2.75	0.720
	[0.065]***
log(crédito x vencer <1 mes)	0.849
	[0.022]***
cred_v_ven3_monto[a]	9.455
	[2.832]***
tarjeta_v_ven3_monto[b]	3.716
	[0.879]***
Sin Historia en Tarjetas Crédito	6.373
	[4.582]**
Tarjeta en Institución "x"	34.090

	[24.814]***
R2	0.32
N	14198

Cuadro 1: Probabilidad de Default en créditos de consumo y microcréditos, IFI>6
Nota: *, **, ***: Estadísticamente significativo al 10%, 5%, y 1% respectivamente

En cuanto a las variables socioeconómicas que resultaron predictoras del riesgo de default se encontró que quienes declararon como origen de ingresos ser amas de casa o estudiantes tienen más del doble de chances de caer en default que quienes declararon otro origen de ingresos. Así también quienes declararon como actividad económica "entretenimiento" o "sector privado" tienen un 20% más de riesgo de caer en default.

Finalmente, las variables relativas a la vivienda del cliente resultaron ser altamente predictivas. Por ejemplo, quienes declaran tener un año en esa vivienda presentan un peor perfil, mientras que quienes cuentan con vivienda propia tienen una probabilidad relativa mucho menor de caer en default, especialmente si esa vivienda está avaluada en más de 50 mil dólares.

Dentro del conjunto de variables financieras, lo primero que se reporta es el efecto del logaritmo de IFI separado en dos tramos (obtenidos en el análisis descriptivo). No obstante, en ambos tramos el efecto de un mayor tiempo como cliente reduce el riesgo de caer en default especialmente para personas con más de 16 meses como cliente de la institución donde obtuvo el crédito. Así también, a medida que el monto por vencer en los siguientes 30 días se incrementa se reducen las chances de caer en default.

Las siguientes dos variables (cred_v_ven3_monto y tarjeta_v_ven3_monto) fueron construidas para reflejar el efecto del ratio: valores vencidos promedio en los últimos 3 meses en relación al valor de la operación (para créditos regulares y tarjetas, respectivamente). Esta es una variable dicotómica que toma el valor 1 si este ratio está por encima de 0.66% (lo que es equivalente a un retraso de 2% en tres meses).

Estas variables tienen un alto valor predictivo y quienes caen en estos atrasos presentan un riesgo relativo elevado. Situación similar ocurre con aquellos que no usaban tarjeta de crédito (sin historia). Finalmente, una característica que en principio elevaría de forma trascendental las chances de caer en default es el uso de la tarjeta de crédito de una institución en particular.

En el ámbito de las variables financieras, la variable crédito por vencer a treinta días o menos, en interacción con las variables relativas a vivienda, resulto ser un predictor adecuado. Finalmente, se encontró que las personas que tenían un crédito vigente en una institución en particular estaban asociadas con un alto nivel de default[3].

Demográficas	**Odds-ratios**
Edad	
18 - 29	2.056
	[0.575]***
30 – 33	2.460
	[0.774]***

[3] Tanto en el efecto de las interacciones como de esta institución, tenemos las mismas aprehensiones que para el caso de la tarjeta "x" en el segmento IFI>6 (ver pie de página 16).

>34	(base)
Sin Estudios	1.649
	[0.391]**
Casado / Viudo	0.534
	[0.139]**
Cargas Familiares ≤1	0.613
	[0.153]**
Guayas / Manabi	1.586
	[0.383]*
Socioeconómicas	
Origen Ingresos: Empleado Privado	2.117
	[0.589]***
Sector Comercio	2.393
	[0.589]***
Vivienda Propia	0.181
	[0.091]***
Valor Vivienda>0 &<50.000	5.348
	[3.072]***
Financieras	
log(crédito x vencer <1 mes)	0.662
	[0.096]***
log(credxvencer<1mes)*Viv. Propia	1.754
	[0.326]***
log(credxvencer<1mes)*V.V.>0 &<50m	0.650
	[0.132]**
Crédito en Institución "x"	6.556
	[4.372]***
R2	0.11
N	4435

Cuadro 2: Probabilidad de Default en créditos de consumo y microcréditos, IFI ≤ 6
Nota: *, **, ***: Estadísticamente significativo al 10%, 5%, y 1% respectivamente

Pruebas de idoneidad de las predicciones

Curva ROC (Receiver Operating Characteristic)

Es una línea que relaciona la tasa de verdaderos positivos (sensibilidad) con la tasa de falsos positivos y determina en qué nivel el modelo caracteriza bien a los individuos bajo la variable de análisis (Hilbe 2009). El área debajo de la curva determina cuan bien el modelo distingue entre los dos grupos que están bajo análisis.

Prueba de Hosmer-Lemeshow

La prueba evalúa si las tasas de eventos observados coinciden con las tasas de eventos esperados en subgrupos de la población modelo. La prueba de Hosmer-Lemeshow específicamente identifica subgrupos como deciles de riesgo de los valores ajustados. Modelos para los cuales los valores observados y esperados en los subgrupos son similares estarían bien calibrados.

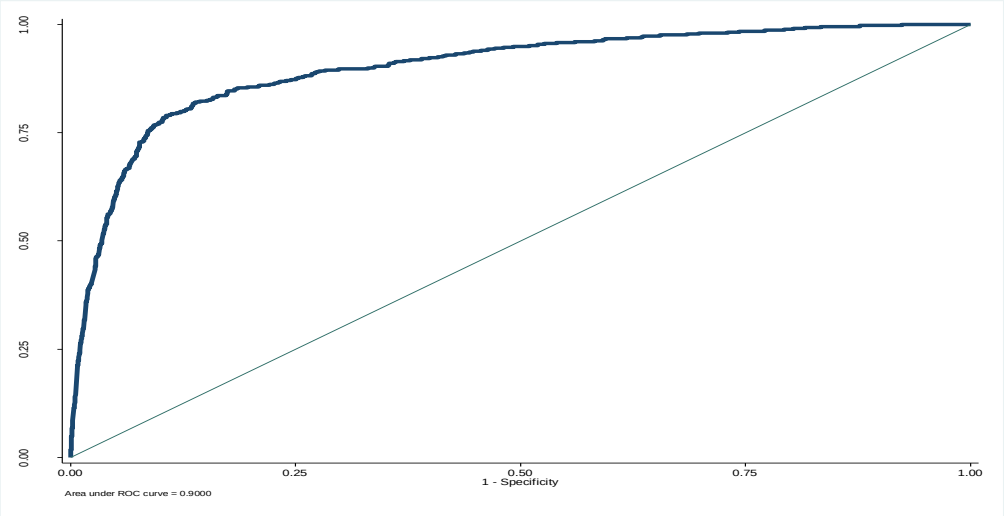

Figura 4: Curva ROC. IFI ≤ 6

Ese es el caso de nuestros resultados, donde el p-value>0.05 nos indica que no se rechaza la hipótesis nula de que los valores esperados y observados a través de los deciles son estadísticamente no distintos. En otras palabras, lo observado y lo predicho tienen una elevada coincidencia, lo cual daría cuenta de las buenas propiedades predictivas de estos modelos.

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0019	1	0.6	473	473.4	474
2	0.0035	0	1.1	418	416.9	418
3	0.0057	2	1.9	438	438.1	440
4	0.0078	1	2.9	441	439.1	442
5	0.0114	4	4.2	440	439.8	444
6	0.0152	7	6.3	458	458.7	465
7	0.0229	10	7.7	412	414.3	422
8	0.0306	14	11.5	430	432.5	444
9	0.0454	17	16.2	433	433.8	450
10	0.2820	27	30.6	409	405.4	436

```
number of observations =      4435
      number of groups =        10
Hosmer-Lemeshow chi2(8) =      4.51
           Prob > chi2 =      0.8084
```

Figura 5: Prueba de Hosmer-Lemeshow IFI ≤ 6

Cálculo del score crediticio

El primer paso en la asignación de scores consiste en dividir la distribución de probabilidades de default en 5 segmentos (ventiles). Luego se establece un puntaje, normalizándolo a un valor de odds ratio elegido por el autor. En este caso, se escogió que a los 720 puntos se de una relación de buenos a malos de 50:1 (o 5000:100). El puntaje va de 0 a 1000.

En el Gráfico 12. se observa que, para el caso IFI>6, el porcentaje de clientes malos se reduce en la medida que la persona tenga un mejor puntaje, pasando de 14% en la peor categoría (211-688) a menos de 1% en la mejor (>929). Aquí también se reportan los odds ratios para cada tramo, los cuales se incrementan exponencialmente en la medida que el tramo del puntaje se incrementa. Por ejemplo, en la mejor categoría tenemos más de 70000 buenos por cada 100 malos, mientras que en la peor apenas 630 por cada 100 malos. Esto da cuenta de la idoneidad de este modelo (y los puntajes generados) para distinguir entre ambos tipos de clientes, lo cual da información valiosa al usuario.

Complementariamente, el valor del estadístico KS alcanza casi el 70%, lo cual cataloga a este modelo como de alto valor predictivo. Así también se reporta en el Gráfico 13. el porcentaje acumulado de clientes malos en el peor 20, 40 y 60%. Se observa que ya en el primer ventil el modelo captura al 85% de los clientes malos, quedando apenas un 3% de ellos en los dos tramos de puntajes superiores. Desde otra óptica, se puede decir también que si una institución financiera decide conceder créditos solo a quienes tienen puntajes por encima de los 843 puntos solo 0.27% resultarían malos lo que también se interpreta como buen síntoma de la capacidad discriminadora del modelo.

En el caso del segmento IFI≤ 6, los resultados son un poco más modestos empero de un nivel más que aceptable. Los odd ratios se incrementan exponencialmente; el KS es de 44%; y, un 95% de los malos clientes

son capturados en los tres peores tramos (dejando de nuevo poco más de 0.2% de malos - respecto del total de clientes- en los dos mejores tramos).

IV. CONCLUSIONES

Este documento presentó los resultados para el modelo nacional genérico de credit scoring (calificación crediticia). En esta versión, por restricciones relativas al acceso a datos, se obtuvieron y evaluaron modelos solo para créditos de consumo, microcréditos, y tarjetas de crédito, separados en dos segmentos: clientes con más de 6 meses de relación con la institución financiera que les otorgo el crédito (IFI) y su complemento. Se trabajó con una muestra pequeña aleatoriamente elegida de 30 mil personas, que después de aplicar exclusiones se redujo a cerca de la mitad. La ventana de análisis comprendió los meses de agosto de 2012 a febrero de 2013 y el seguimiento fue de un año.

Los modelos para el segmento IFI>6 presentan excelentes cualidades predictivas, ofreciendo información relevante para que los usuarios puedan efectivamente separar malos de buenos clientes y mejorar la eficiencia en la asignación de créditos, disminuyendo al mismo tiempo el riesgo global del sistema financiero. En el caso del segmento IFI ≤ 6, esperamos mejorar sus capacidades predictivas cuando se tenga acceso a mayor cantidad y calidad de datos. No obstante, la verdadera calidad de las predicciones será determinada en una siguiente versión cuando se implementen las validaciones respectivas.

Las correlaciones se utilizan únicamente entre las variables independientes y con el fin de evitar redundancia de variables, la correlación cercana a uno -o menos uno- indica alta dependencia lineal entre las dos variables, una correlación cercana a cero indica independencia lineal entre las variables. El detalle de la correlación entre las variables se indica en el anexo al final. (Anderson 2007)

REFERENCIAS

Anderson A. (2007). The Credit Scoring Toolkit Theory and Practice for Retail Credit Risk Management and Decision. *Automation*. New York.

Hilbe, J. M. (2009). *Logistic regression models*. CRC press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 817-838.