Anuario Facultad de Ciencias Económicas y Empresariales
ISSN 2218-3639
Recibido: febrero 2020/Aceptado: julio 2020

La regresión logística binaria como instrumento para la predicción del impago

Binary logistic regression as an instrument for the prediction of default

Dr. C. David Expósito Martínez, davide@uo.edu.cu, https://orcid.org/0000-0001-7372-8017

Universidad de Oriente, Santiago de Cuba, Cuba

Resumen

El objetivo de este artículo es el desarrollo de un modelo de predicción del impago del sector cuentapropista en el Banco Popular de Ahorro (BPA) de la provincia Santiago de Cuba. Se empleó el método de análisis y síntesis, y la regresión logística binaria (RLB) para el tratamiento de los datos. El análisis de las principales metodologías de clasificación de clientes empleadas en la actividad bancaria permitió identificar a la RLB como instrumento de pronóstico. Se consideró como variable dependiente el impago y como variables independientes la Capacidad de Pago, Historial Crediticio, Evaluación Cualitativa, Historial Tributario y Experiencia, lo que conllevó a estimar la probabilidad de impago de los nuevos solicitantes de financiamientos.

Palabras clave: probabilidad de impago, puntaje crediticio, regresión logística binaria, Trabajo por Cuenta Propia.

Abstract

The objective of this article is the development of a prediction model of the non-payment of the self-employed sector in the Popular Savings Bank (BPA) of Santiago de Cuba province. It was used the analysis and synthesis method, and the binary logistic regression (BLR) for the data treatment. The analysis of the main methodologies of clients' classification used in the banking activity allowed identifying the BLR as a prognostic instrument. Default was considered a dependent variable and the Capacity to Pay, Credit History, Qualitative Evaluation, Tax History and Experience were considered independent variables, which led to an estimate of the probability of default of new applicants for financing.

Key words: probability of default, credit score, binary logistic regression, self-employment.

Introducción

En la actividad bancaria resulta determinante el cómo se gestionan y miden los riesgos. El amplio y acelerado desarrollo de las metodologías para su medición y control ha estado fuertemente influenciado por las transformaciones coyunturales en el sistema financiero internacional. En la actualidad, existe un vasto y acelerado desarrollo del uso de estas herramientas, que cuantifican el riesgo de crédito como una probabilidad, lo que contribuye a la mejor administración del mismo en las instituciones bancarias.

En Cuba, a partir de los Lineamientos de la Política Económica y Social del Partido y la Revolución, aprobados por el VII Congreso del Partido Comunista de Cuba, se ha propiciado la existencia de negocios con forma de gestión no estatal, llamados trabajadores por cuenta propia.

Los financiamientos dirigidos a este segmento, que crece como tendencia, deben someterse a estrictas reglas de análisis de riesgo que permitan su rápida recuperación.

En ese sentido, la gestión del riesgo de crédito ha centrado su atención en su administración con el objetivo de minimizar los riesgos en el proceso de otorgamiento de financiamientos al sector de TCP:

> "El otorgamiento de los financiamientos se realiza a partir de estrictos análisis de riesgo, dando especial atención a la valoración de la recuperación del financiamiento, a partir de la utilización de herramientas de medición del riesgo de crédito que tributen a a la toma de decisiones" (BCC, 2016, p.3)

El presente artículo tiene como objetivo estimar el impago, en términos de probabilidades, a partir del empleo de un modelo de regresión logística binaria. El estudio consideró una serie de datos concernientes a las características de los financiamientos y su relación con los negocios y sus administradores, en el período del 2017-2019.

Se exponen los principales modelos de puntuación crediticia cuyo objetivo fundamental es la estimación del impago, en el segmento de los pequeños negocios, donde resalta el empleo de herramientas estadísticas y econométricas de pronóstico.

Para el empleo de esta herramienta, se expuso una serie de consideraciones de orden técnico, necesaria para su correcta implementación. Se enfatiza en la interpretación de los principales parámetros del modelo de regresión logística, así como las ventajas que ofrece su uso.

Al realizar el pronóstico del impago, se utilizan como variables independientes la Capacidad de Pago, Historial Crediticio, Evaluación Cualitativa, Historial Tributario y Experiencia; todas identificadas en investigaciones enfocadas a la gestión del riesgo crediticio en el segmento cuentapropista, realizadas en sucursales bancarias de la provincia. La estimación se realizó para tres escenarios diseñados a partir del resultado de los estadígrafos calculados para cada variable del estudio.

Fundamentación teórica

Los modelos basados en puntajes o estadísticos de corte, a los que es habitual llamar métodos de *scoring*, son metodologías ampliamente utilizadas en la práctica comercial y bancaria, que sirven para discernir entre los clientes a los cuales se les otorga o no el crédito (Leal, Aranguiz, & Gallegos, 2018; Montalván, 2019; Narváez, 2019).

Los *credit scoring* son métodos estadísticos utilizados para clasificar a los solicitantes de crédito, o incluso a quienes ya son clientes de la entidad evaluadora, discriminando entre "buenos" y "malos" riesgos (Demma, 2017; Chopra & Bhilare, 2018). De esta manera, estos sistemas conocidos además como clasificación del riesgo de insolvencia, morosidad o impago, también se pueden concebir como un sistema que, mediante predicciones, califica un crédito y mide el riesgo inherente al mismo.

La probabilidad de impago, desempeña un papel de extraordinaria importancia en los sistemas de calificación del riesgo de crédito, puesto que es la base para calcular la pérdida esperada en orden al cómputo del capital económico.

En la construcción de un modelo de este tipo es importante el uso de herramientas estadísticas para clasificar los créditos, pues con ellas se conseguirá un mejor análisis de toda información relacionada con el financiamiento otorgado y su relación con la morosidad, así como la relación entre el riesgo y la rentabilidad, y la agilización general de procesos de análisis que permite la reducción del costo en la concesión de un crédito.

Una revisión de los trabajos destacados en materia de pronóstico del impago en la banca comercial, destaca el uso de modelos pertenecientes tanto a técnicas paramétricas como no paramétricas. Aunque no en la misma medida del caso del financiamiento al sector de empresas, el análisis y evaluación de los financiamientos a los pequeños negocios también ha evolucionado con la aplicación de estas herramientas.

Una de las primeras aplicaciones para este tipo de segmento de clientes, empleó el análisis discriminante multivariante para la construcción de un modelo de predicción estadísticas (Viganò, 1993). Por su parte, Sharma & Zeller (1997) y Zeller (1998) utilizaron la herramienta paramétrica no lineal *Tobit*, lo que le permitió identificar algunos factores que incidían en el impago.

Vogelgesang (2003) desarrolló una aplicación cuyo objetivo fue predecir el riesgo de impago con el empleo de la regresión logística multinomial, apoyado por la estimación con modelos probit bivariados. Las variables explicativas empleadas en el estudio fueron agrupadas en variables personales del cliente, variables del negocio, variables del préstamo y variables del entorno.

Diallo (2006), empleó la regresión logística en la explicación de los factores que incidían en el impago, comparando los resultados obtenidos con un análisis discriminante. Su estudio demostró que el modelo de regresión logística mejoraba la capacidad predictiva del análisis discriminante, en el cual los valores de la sensibilidad y especificidad son bastante cercanos.

Van Gool, Baesens, Sercu, & Verbeke (2009) para estimar la probabilidad de impago emplearon una regresión logística binaria, obteniendo altos porcentajes de clasificación global de sus clientes.

Otro estudio que aplica esta herramienta para el pronóstico del impago en clientes de la banca comercial es el Expósito y Rodríguez (2020), donde construyen un modelo de puntaje crediticio para el segmento cuentapropista cubano.

El presente artículo propone pronosticar el impago, a partir del empleo de la regresión logística binaria, e incluye en el estudio un conjunto de variables predictoras identificadas en el estudio de Expósito, Díaz y Rodríguez (2018) como factores determinantes en el impago para este segmento: Capacidad de Pago, Historial Crediticio, Evaluación Cualitativa, Historial Tributario y Experiencia.

El modelo empleado para pronosticar el impago (denotado con la letra p) es de tipo Logit. A continuación se exponen consideraciones de esta herramienta (Franco, 2010; Aguayo, 2016; Aguayo y Lora, 2016),

La naturaleza no lineal de la regresión logística no permite el empleo del método de los mínimos cuadrados, las estimaciones se hacen mediante máxima verosimilitud.

En el modelo de elección binaria:

$$P(Y = 1/X_1, X_2, \dots, X_k) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Donde

β_0: es una constante

$\beta_1, \beta_2 \dots \beta_k$: son los coeficientes logísticos correspondientes a cada variable predictora

$X_1, X_2 \dots X_k$ son variables predictoras

G es una función que toma estrictamente valores entre 0 y 1, o sea, $0 \leq G(z) \leq 1$ para todos los números reales z.

Si $G(z) = \frac{e^{-z}}{1+e^{-z}}$ estamos ante el modelo logit cuya expresión será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

Determinación de los Coeficientes y su significación

Los coeficientes estimados ($\beta_0, \beta_1, \beta_2 \dots \beta_k$) son las medidas de los cambios en el ratio de probabilidades, están expresados en logaritmos y se denominan odds ratio (OR):

$$\frac{\text{Prob}_{(\text{evento})}}{\text{Prob}_{(\text{noevento})}} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}$$

Si el coeficiente estimado es positivo, su transformación antilogarítmica será mayor a 1 y el odds ratio aumentará y, por tanto, el modelo tendrá una alta probabilidad de ocurrencia; lo contrario sucede cuando toma valores negativos.

La interpretación de este parámetro cuando es igual a 1, indica que no existe factor de riesgo, pues existe equiprobabilidad en ambas categorías de la variable. En cambio, si su valor es mayor que indica que aumenta la probabilidad de ocurrencia del evento, y un valor menor de 1 implica disminución en esa probabilidad (Uanhoro, Wang, & O'Connell, 2019).

Por otro lado, la OR tiene muy buenas propiedades matemáticas:

- OR toma valores entre cero e infinito. Una transformación logarítmica de la odds proporciona una importante medida para el análisis de datos categóricos, denominada transformación logit que varía entre - ∞ y + ∞ y se define sobre una categoría de la variable dependiente.
- El modelo logístico de regresión puede emplearse para determinar intervalos de confianza para la OR: si dichos intervalos contienen al valor OR=1, no puede

rechazarse que el factor de riesgo no sea tal. En otro caso, se dice que aumenta o disminuye la oportunidad del evento en función de que el intervalo de confianza sea de valores mayores o menores que uno respectivamente.

En la ecuación es fácil verificar que p no está linealmente relacionado con z (es decir con Xi), lo que crea un problema de estimación, porque p es no lineal no solamente en X, sino también en los β. Esto significa que no puede emplearse el procedimiento de los Mínimos Cuadrados Ordinarios para estimar los parámetros. Para la estimación del modelo se emplea el método de estimación por Máxima Verosimilitud que no establece restricción alguna respecto a las características de las variables predictoras.

En el procedimiento de máxima verosimilitud se seleccionan las estimaciones de los parámetros que hagan posible que los resultados observados sean lo más verosímiles posible. A la probabilidad de los resultados observados, dadas las estimaciones de los parámetros, se le denomina verosimilitud.

Los coeficientes β son las medidas de los cambios en la razón de probabilidad denominado OR y están expresados en logaritmos, por lo que deben ser transformados para ser interpretados. Un coeficiente positivo aumenta la probabilidad de ocurrencia y un coeficiente negativo la disminuye.

Para contrastar la hipótesis nula de que los coeficientes son iguales a cero se utiliza el estadístico W de Wald, que es igual al cuadrado de la razón entre un coeficiente de regresión y su error típico (Bangdiwala, 2018; Xiao, *et al*., 2018). El estadístico W sigue una distribución Chi cuadrado, con un grado de libertad, lo que es apropiado para su uso con datos categóricos y desempeñan el mismo papel que el estadístico t en la regresión lineal múltiple, para las variables incluidas en la ecuación.

Verificación de la Bondad del ajuste

La regresión logística maximiza la verosimilitud de que un suceso tenga lugar; la utilización de esa técnica de estimación alternativa requiere evaluar el modelo de una forma diferente, tal y como se presenta a continuación:

- La medida global del ajuste del modelo, similar al coeficiente de determinación R^2, viene dada por el valor de la verosimilitud, como su valor es pequeño, se utiliza menos dos veces el logaritmo del valor de verosimilitud y se representa por -2LL. Un buen modelo con un buen ajuste tendrá un valor pequeño para -2LL. El

valor mínimo para -2LL es cero. Un ajuste perfecto tiene una verosimilitud de 1 y -2LL igual a cero.

- El contraste Chi cuadrado para la reducción en el logaritmo del valor de verosimilitud proporciona una medida de mejora debido a la introducción de variables independientes. Se comparan las diferencias entre -2LL; el punto de partida para la comparación lo proporciona un modelo nulo, en el cual no se incluyen variables predictoras y se verifican las desviaciones en el modelo al incluir una o más variables predictoras. Chi cuadrado contrasta la hipótesis nula, que los coeficientes de todos los términos excepto la constante son cero. Los grados de libertad en este caso están dados por la diferencia entre el número de los parámetros de los dos modelos.
- Existen, además, varias medidas similares al R^2 como medidas globales del ajuste.
- Otro contraste para la bondad del ajuste es el desarrollado por Hosmer y Lemeshow, que plantea como hipótesis nula que el modelo estimado es el adecuado y constituye un homólogo del R^2 del modelo de regresión lineal clásico; por tanto, lo que se desea es que su significación sea tan grande como sea posible. Este contraste proporciona una medida global de capacidad predictiva, que no se basa en el valor de la verosimilitud, sino en la predicción real de la variable dependiente.

  Una restricción en su uso es que se necesita contar con una muestra grande que asegure por lo menos cinco observaciones en cada grupo. Por otro lado, Chi cuadrado es sensible al tamaño muestral y se puede encontrar significación estadística en diferencias pequeñas al aumentar el tamaño de la muestra.

Requisitos y limitaciones del modelo

Al analizar el modelo *logit* se pueden plantear algunos requisitos y limitaciones para su empleo, entre ellas se destacan:

- Los parámetros del modelo se calculan usando una estimación de máxima verosimilitud. Estas solo son válidas cuando para cada combinación de variables independientes se cuenta con un número suficientemente alto de observaciones. Si los parámetros estimados en el modelo son anormalmente grandes, posiblemente esta condición sea violada y se puede solucionar agrupando categorías (donde tenga sentido).
- No introducir variables innecesarias.

- Ninguna variable relevante debe ser excluida. Si se identifican variables confusoras el estudio puede ser estratificado en submuestras.
- La colinealidad es un problema como ocurría en la regresión lineal múltiple. Si los errores típicos en la estimación de los coeficientes, o los intervalos de confianza son anormalmente grandes, es posible que esta situación se esté dando.

El problema de la introducción de variables innecesarias puede ser resuelto a través de la regresión de combinaciones de las diferentes variables hasta lograr la más eficiente; es decir, la estimación de más de un modelo de Regresión Logística; teniendo en cuenta que la apreciación de un único modelo que incluya el conjunto de todas las variables explicativas, puede arrojar peores resultados que una valoración individual.

Para analizar el inconveniente de la colinealidad entre las variables regresoras, es necesario plantear primero, que colinealidad o multicolinealidad se refiere a la existencia de una relación entre algunas o todas las variables explicativas de un modelo de regresión, es decir que se encuentran correlacionadas. La multicolinealidad es un problema de grado y no de clase. La distinción importante no es entre la presencia y la ausencia de esta, sino entre sus diferentes grados, para ello existen algunos procedimientos que permiten medir si es grave o no, entre los más notorios:

- Una medida global de bondad del ajuste elevada pero pocas razones w significativas. Si las pruebas realizadas para la bondad de ajuste (R^2 de Nagelkerke y Hosmer y Lemeschow) muestran resultados altos, esto es por encima de 0,8, se puede considerar un ajuste bueno; pero las pruebas w individuales mostrarán que ninguno o muy pocos coeficientes son estadísticamente distintos de cero. En este caso la multicolinealidad sólo se considera dañina cuando no se pueden separar las influencias de las variables independientes sobre Y.
- Altas correlaciones entre parejas de variables regresoras. Si el coeficiente de correlación de orden cero o entre dos regresoras es superior a 0,8, entonces la multicolinealidad es un problema grave.

En consecuencia, el proceso para lograr un modelo que supere los inconvenientes enunciados será:

1. Lograr una regresión eficiente que incluya el número preciso de regresores y que revele los factores de riesgo existentes en el estudio.
2. Comprobar la existencia o no, de la multicolinealidad a través del estadístico Chi-Cuadrado que permite contrastar la hipótesis de que las variables son

independientes, sin indicar la magnitud o dirección de la relación. Si es significativa la prueba quiere decir que existe dependencia entre las variables explicativas y, por tanto, colinealidad, teniendo en cuenta que esta prueba no indica la gravedad del fenómeno.

3. Establecer la gravedad de la violación del requisito antes mencionado y determinar qué elección hacer frente al problema, si seguir la corriente de pensadores que plantean no hacer nada o aplicar algunas reglas prácticas (transformación de variables, aumentar el tamaño de la muestra, entre otras); todo dependerá de la intensidad de la violación.
4. Verificar la Bondad del Ajuste.

El modelo de Regresión Logística empleado permite, dada una o más variables independientes ya sean cuantitativas y/o cualitativas, obtener una función lineal de las variables independientes, que permita clasificar a los solicitantes en uno de los dos grupos establecidos por los dos valores de variable dependiente, el grupo de los clientes morosos y no morosos.

Con el objetivo de adaptar el modelo a las condiciones de funcionamiento de la actividad bancaria, se partió de los criterios de selección de las variables dependientes e independientes; quedó representado en la siguiente fórmula:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}$$

Donde

p: es la probabilidad de observar la categoría o evento a predecir, la probabilidad de impago.

β_i: parámetros a estimar.

x_i: variables independientes o presuntos factores de riesgo a incluir en el análisis. Para codificar las variables se parte del criterio que existe una variable dependiente convencionalmente denominada Y, que puede ser dicotómica y una o más variables independientes, denominadas X, que pueden ser de cualquier naturaleza bien sean cualitativas o cuantitativas. La variable dependiente tomará valor 1 cuando el cliente sea moroso y valor 0 en el caso de que no lo sea.

Métodos utilizados

En el artículo se empleó el método de análisis y síntesis, que permitió estudiar las diferentes herramientas empleadas en el pronóstico del impago en el segmento de particulares, por parte de instituciones bancarias. Al analizar el objeto de estudio, se identificaron las variables relacionadas con el mismo, así como su incidencia. Se empleó la estadística descriptiva para determinar estadígrafos de las variables predictoras empleadas en el estudio. Finalmente se empleó la regresión logística binaria con el fin de pronosticar el impago en dos sucursales bancarias de la provincia, para así arribar a conclusiones ajustadas al marco teórico y a los hechos empíricos detectados.

Discusión y resultados

Las variables empleadas en el pronóstico fueron identificadas a partir del análisis de una base de datos constituida por la información correspondiente a los financiamientos y las características del negocio y su administrador. Se analizó una muestra de 148 clientes de dos de las principales sucursales del BPA de la provincia Santiago de Cuba, en el periodo del 2017 - 2019.

A partir de estos resultados, se procedió a estimar la probabilidad de impago para tres escenarios que representan tres posibles características de una solicitud de financiamiento. Para la construcción de los escenarios bajo los cuales se pronosticó el impago, se determinaron los valores máximos, medios y mínimos de cada uno de los elementos a considerar en la evaluación de un financiamiento a la actividad cuentapropista, mediante el uso de estadígrafos.

Definición de variables

El estudio realizado estableció como variable dependiente la probabilidad de impago (PD), definida como variable dicotómica en la que el valor cero (0) corresponde a clientes que pagan todas las cuotas y uno (1) a clientes atrasados.

Las variables independientes son la Capacidad de Pago, Historial Crediticio, Evaluación Cualitativa, Historial Tributario y Experiencia, clasificadas de la siguiente manera:

Capacidad de pago (X1), variable ordinal que representa la capacidad que tiene el solicitante de amortizar el financiamiento, que adquiere los siguientes valores: (0) Insatisfactoria, (1) Regular, (2) Buena, (3) Satisfactoria, (4) Muy buena. El criterio para establecer esta clasificación es el estipulado en la Instrucción número 4 (BCC, 2016).

Historial crediticio (X2), variable ordinal que representa la clasificación crediticia a partir del comportamiento de los pagos de amortización, que adquiere los siguientes valores: (0) Insatisfactorio, (1) Irregular, (2) Bueno, (3) Muy bueno. El criterio para establecer esta clasificación es el estipulado en la Instrucción número 4 de 2016 del BCC.

Es necesario destacar que la clasificación para los clientes sin información en los historiales de crédito, o sea, nuevos clientes, sea "Muy bueno" (valor 3), siempre que el criterio del especialista encargado de la evaluación e inspección económica y financiera del correspondiente negocio no sea otro.

Evaluación cualitativa (X3): valoración dada por los especialistas a partir de indicadores cualitativos. Variable ordinal: (0) Muy mala, (1) Mala, (2) Regular, (3) Buena, (4) Muy buena.

Historial tributario (X4), variable dicotómica que representa la clasificación tributaria a partir del comportamiento del pago de los impuestos. Adquiere los valores siguientes: (0) "Malo" y (1) "Bueno".

La siguiente variable, Experiencia (X5), que representa la cantidad de años que tiene el solicitante ejerciendo la actividad cuentapropista, es clasificada como variable dicotómica; toma el valor 0 cuando "no tiene experiencia" y 1 cuando "sí tiene experiencia".

El criterio para establecer esta clasificación es: cuando el solicitante tiene dos años o menos ejerciendo su negocio o de estar vinculado directamente con la actividad que desarrolla se le considera inexperto (valor 0) o con insuficiente experiencia. Si el tiempo es superior entonces es experimentado (valor 1).

Para la estimación del impago, en términos de probabilidad, fue empleado el software *Statistical Package for the Social Science* (SPSS 22.0).

Tablas de validación del modelo

Se parte de conocer los parámetros de las variables incluidas en el estudio como predictoras del impago. Como se observa en la tabla 1, los coeficientes del modelo pueden interpretarse de acuerdo a su signo. Los valores negativos aumentan las posibilidades de que la variable dependiente tome valor 0 (que pague); en cambio, si el signo es positivo, la mora entonces aumentará.

Tabla 1. Parámetros de las variables del modelo

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 5[e]	X1	-,544	,202	7,258	1	,017	,580
	X2	-,004	,001	17,690	1	,000	,996
	X3	-,005	,002	12,304	1	,000	,995
	X4	-,006	,000	9,974	1	,015	,994
	X5	-,343	,170	24,094	1	,000	,710
	Constante	,177	,693	,065	1	,799	1,193

Fuente: Visor de resultados del SPSS 22.0

El valor de todos los coeficientes de las variables independientes del modelo es negativo, lo cual indica que la relación entre estas variables y la variable dependiente es inversa. Mientras el solicitante del crédito posea buena calificación (mayor valor) de Capacidad de pago (X1), Historial crediticio (X2), Historial tributario (X3), Valoración cualitativa (X4) y Experiencia (X5), tendrá mayores probabilidades de pagar el financiamiento recibido (variable dependiente cercana a 0). Los resultados esperados para este caso se corresponden con lo expresado por el modelo.

El parámetro constante no es un elemento muy relevante en este tipo de regresión; no obstante, su signo positivo indica que cuando se estén dando a la vez todas las codificaciones de los regresores que indican factores de riesgo, las posibilidades de que el demandante de fondo no pague aumentan.

Otro elemento para interpretar el modelo es la evaluación de las exponeciales β (odds ratios) que en este caso son menores que uno, lo que denota una baja probabilidad de que el evento ocurra (impago) con la presencia de las cinco variables predictoras.

De manera individual, las odds ratios pueden analizarse en términos de oportunidades. Se observa que un solicitante de fondos con alta capacidad de pago tiene una posibilidad 0,58 veces mayor (o sea, disminuye la probabilidad del impago en un 42 %) de no pagar que si esta no fuera alta. Para el caso de la variable experiencia (dicotómica), la oportunidad de que un solicitante experimentado en su actividad no pague es 0,71 veces mayor (o sea, disminuye la probabilidad de impago en un 29 %) a que si no lo fuera.

La tabla de clasificación (tabla 2) muestra que la especificidad del modelo es alta (95,5%), lo que posibilita clasificar correctamente a un cliente que cumple con el pago de sus deudas. En cambio, la sensibilidad es inferior, al clasificar, de forma correcta, a un cliente moroso en un 78,4 %.

Tabla 2. Tabla de Clasificación

Observado			Pronosticado		
			MORA		Corrección de porcentaje
			NO MOROSO	MOROSO	
Paso 1	MORA	NO MOROSO	101	10	91,0

		MOROSO	9	28	75,7
	Porcentaje global				87,2
Paso 2	MORA	NO MOROSO	105	6	94,6
		MOROSO	7	30	81,1
	Porcentaje global				91,2
Paso 3	MORA	NO MOROSO	106	5	95,5
		MOROSO	8	29	78,4
	Porcentaje global				91,2
Paso 4	MORA	NO MOROSO	105	6	94,6
		MOROSO	6	31	83,8
	Porcentaje global				91,9
Paso 5	MORA	NO MOROSO	106	5	95,5
		MOROSO	8	29	78,4
	Porcentaje global				91,2
a. El valor de corte es ,500					

Fuente: Tabla de resultados del SPSS 22.0.

No obstante, al analizar de forma conjunta el porcentaje de casos correctamente clasificados (91,2 %) se puede aseverar que la información aportada por estas variables es muy significativa, reafirmando que el modelo de regresión logística es válido para lograr el objetivo propuesto.

Para evaluar la bondad del ajuste del modelo, deben observarse los indicadores de la tabla 3, los que evidencian que luego del quinto paso se explica el 53 % de la variabilidad de los datos (R^2 de Cox y Snell). En cambio, con este modelo se ha logrado explicar el 78,6 % de la variabilidad (R^2 de Nagelkerke) de los datos recogidos sobre el número de morosos y no morosos de las sucursales seleccionadas en el estudio.

La disminución de -2LL (-2 log de la verosimilitud) indica que con cada paso la verosimilitud es mayor y, por tanto, mejor el ajuste del modelo.

Tabla 3. Resumen del modelo

Escalón	- 2 Logaritmo de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	102,651[a]	,350	,519
2	75,142[b]	,460	,682
3	68,176[b]	,485	,719
4	59,761[c]	,514	,761
5	54,556[c]	,530	,786

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.
b. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.
Fuente: Tabla de resultados del SPSS 22.0.

El contraste para la bondad del ajuste de Hosmer y Lemeshow (tabla 4), devuelve resultados positivos con una significación de 0,999 en el último escalón.

Tabla 4. Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	27,033	8	,001
2	5,346	8	,720
3	5,479	8	,705
4	3,907	8	,865
5	,811	8	,999

Fuente: Tabla de resultados del SPSS 22.0.

La prueba Ómnibus (tabla 5) realizada a todos los coeficientes del modelo presenta significaciones por debajo del 0,05. Esto expresa que, al menos una de las variables independientes, pueda explicar el comportamiento de la dependiente.

Tabla 5. Pruebas ómnibus de coeficiente de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	63,800	1	,000
	Bloque	63,800	1	,000
	Modelo	63,800	1	,000
Paso 2	Escalón	27,509	1	,000
	Bloque	91,309	2	,000
	Modelo	91,309	2	,000
Paso 3	Escalón	6,966	1	,008
	Bloque	98,275	3	,000
	Modelo	98,275	3	,000
Paso 4	Escalón	8,415	1	,004
	Bloque	106,690	4	,000
	Modelo	106,690	4	,000
Paso 5	Escalón	5,205	1	,023
	Bloque	111,895	5	,000
	Modelo	111,895	5	,000

Fuente: Tabla de resultados del SPSS 22.0.

Todo lo anterior permitió comprobar que la ecuación logística estimada cumple con las pruebas estadísticas necesarias, lo que la hace útil para obtener pronósticos de la probabilidad de que un determinando cliente pague un financiamiento recibido. Su expresión quedaría de la siguiente manera:

$$p = \frac{e^{2,469-0,043x_1-0,734x_3-0,301x_5-0,045x_{18}-1,747x_{16}}}{1+e^{2,469-0,043x_1-0,734x_3-0,301x_5-0,045x_{18}-1,747x_{16}}} \qquad (1)$$

En lo adelante, se pronosticó la probabilidad del impago, a partir de los valores arrojados por cada uno de los escenarios definidos. Se tienen las características de tres cuentapropistas que desarrollan la misma actividad y solicitan el mismo monto ($ 30 000).

Escenario 1: solicitante con las peores características.

Escenario 2: solicitante con características medias.

Escenario 3: solicitante con las mejores características.

La tabla 6 muestra los resultados de los estadísticos necesarios empleados para cada una de las variables incluidas en el sistema de puntaje crediticio.

Tabla 6. Estadísticos de las variables predictoras

	X1	X2	X3	X4	X5
Total de casos	148	148	148	148	148
Media	2,922	2,129	,521	2,917	,779
Mínimo	,0	,0	,0	2	,0
Máximo	4,0	4,0	1,0	4,0	1,0

Fuente: Elaboración propia

A partir de la expresión matemática 1, se pronosticó la probabilidad de impago para cada uno de los escenarios establecidos. De esta manera, se puede conocer el impago, en términos de probabilidad, y la institución financiadora podría clasificar el riesgo de crédito de manera más precisa. Los perfiles de riesgos mostrados son definidos en la Instrucción 4 del 2016 del BCC. La tabla 7 muestra los resultados del pronóstico para cada uno de los escenarios.

Tabla 7. Probabilidad de impago pronosticada por escenarios

Escenario	Probabilidad de impago (%)	Perfil de riesgo
1	54,41	MUY ALTA
2	15,37	MEDIA
3	8,47	BAJA

Fuente: Elaboración propia

Como se aprecia en la tabla anterior, ninguna de las operaciones fue calificada como de mínimo riesgo, mientras que el solicitante con las características pertenecientes al escenario 1 es calificado como muy riesgoso. Se observa, además, que el riesgo disminuye a medida que mejoran los valores de las variables predictoras.

Conclusiones

1. ***En la actualidad existe un amplio y acelerado desarrollo del empleo de modelos de puntaje crediticio en el segmento de particulares, fundamentados en herramientas paramétricas y no paramétricas, que cuantifican el riesgo de crédito como una probabilidad, lo que contribuye a la mejor administración del mismo en las instituciones bancarias.***
2. ***La regresión logística binaria ofrece un conjunto de bondades y un alto nivel de pronóstico, lo que la convierte en una poderosa herramienta de estimación del impago en la gestión del riesgo crediticio.***
3. ***El empleo de la regresión logística binaria permite estandarizar el análisis de las solicitudes de financiamiento, y obtener un pronóstico del impago en términos de probabilidades.***

Referencias bibliográficas

1. Aguayo, M. (2016, febrero 18). *Fundación Andaluza Beturia para la investigación en salud.* Retrieved from fabis.org
2. Aguayo, M., & Lora, E. (2016, febrero 22). *Fundación Andaluza Beturia para la investigación en salud.* Retrieved from fabis.org
3. Bangdiwala, S. (2018). Regression: binary logistic. *International journal of injury control and safety promotion*, *3*(25), 336-338.
4. BCC. (2016). *Instrucción número 4. Normas para el otorgamiento, control y recuperación de los financiamientos a los trabajadores por cuenta propia y personas naturales autorizadas a ejercer otras formas de gestión no estatal.* La Habana: Banco Central de Cuba.
5. Chopra, A., & Bhilare, P. (2018). application of ensemble models in credit scoring models. *Business Perspectives and Research, 2*(6), 120-141.
6. Demma, C. (2017). Credit scoring and the quality of business credit during the crisis. *Economic Notes: Review of Banking, Finance and Monetary Economics*, *2*(46), 269-306.
7. Diallo, B. (2006, septiembre 21). *Un modele de "credit scoring" pour une institution de microfinance Africaine: le cas de Nyesigiso au Mali.* Retrieved from Mali: Mimeo: https://halshs.archives-ouvertes.fr/halshs-00069163/document
8. Expósito, D., & Rodríguez, S. (2020). Perfeccionamiento de la evaluación del riesgo de crédito al segmento cuentapropista en el Banco Popular de Ahorro. In F. Borrás, *Banca comercial cubana: propuestas de desarrollo* (pp. 176-197). La Habana: Félix Varela.

9. Expósito, D., Díaz, J., & Rodríguez, S. (2018). Factores determinantes del riesgo de crédito en el Banco Popular de Ahorro. *Anuario Facultad de Ciencias Económicas y Empresariales*(Anuario Especial), 137-153.

10. Franco, M. (2010). *Nuevo procedimiento para el análisis del riesgo de crédito en el Banco de Crédito y Comercio en Santiago de Cuba.* Santiago de Cuba: Tesis Doctoral.

11. Leal, A., Aranguiz, M., & Gallegos, J. (2018). Credit risk analysis, credit scoring model proposal. *Revista Facultad de Ciencias Económicas: Investigación y Reflexión, 1*(26), 181-207.

12. Montalván, C. (2019). *Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito.* Quito: Master's thesis, Universidad Andina Simón Bolívar.

13. Narváez, A. (2019). *Variables determinantes de la probabilidad de incumplimiento de los créditos comerciales en una institución financiera del Ecuador, aproximación bajo el modelo de regresión logística binaria.* Ambato: Master's thesis, Universidad Técnica de Ambato. Facultad de Contabilidad y Auditoría. Dirección de Posgrado.

14. Sharma, M., & Zeller, M. (1997). Repayment performance in group-based credit programs in Bangladesh: An empirical analysis. *World Development, 10*(25), 1731-1742.

15. Uanhoro, J., Wang, Y., & O'Connell, A. (2019). Problems With Using Odds Ratios as Effect Sizes in Binary Logistic Regression and Alternative Approaches. *The Journal of Experimental Education*,, 1-20.

16. Van Gool, J., Baesens, B., Sercu, P., & Verbeke, W. (2009). An Analysis of the Applicability of Credit Scoring for Microfinance. *Orlando: Academic and Business Research Institute Conference*.

17. Viganò, L. (1993). A credit-scoring model for development banks: An African case study. *Savings and Development, 17*(4), 441-482.

18. Vogelgesang, U. (2003). Microfinance in times of crisis: The effects of competition, rising indebtness, and economic crisis on repaymentbehaviour. *World Development, 12*(31), 2085-2114.

19. Xiao, R., Liu, Y., Huang, X., Shi, R., Yuw, & Zhang, T. (2018). Exploring the driving forces of farmland loss under rapidurbanization using binary logistic regression and spatial regression: A case study of Shanghai and Hangzhou Bay. *Ecological Indicators, 95*, 455-467.

20. Zeller, M. (1998). Determinants of repayment performance in credit groups: The role of program design, intragroup risk pooling, and social cohesion. *Economic Development and Cultural Change*, *3*(46), 599-620.