Leonela Tacure P., Víctor Villacorta P., Oscar Rafael T.

Posgrado Ciencia de Datos, Universidad Ricardo Palma, Lima

[Leonela.tp@gmail.com](mailto:Leonela.tp@gmail.com), [v.villacorta.unmsm@gmai.com](mailto:v.villacorta.unmsm@gmai.com), [orafaelp@alicorp.com](mailto:orafaelp@alicorp.com)

[[1]](#footnote-1)

Minería de datos para determinar incumplimiento de pagos en créditos para Banco Colombia

(Mayo de 2021)

Resumen –. El sistema financiero (SF) en Colombia, requiere implementar las nuevas tendencia en minería de datos para tener una ventaja competitiva frente a los demás Bancos, utilizando sus conocimientos y datos como principal activo. El objetivo de esta investigación es determinar el incumplimiento de pagos en los créditos brindados por Banco Colombia. En este sentido, se realizó una prueba escrita sobre la personalidad a los clientes con créditos pendientes de pago para ver como estas influyen en el comportamiento del pago crediticio.

A partir de esto, se realiza un análisis de los datos, como los de las pruebas influyen en el incumplimiento de pago. Posteriormente, se plantean la comparación entre 2 modelos: Regresión Logística y Enlaces Asimétricos; y contrastarlos con los indicadores de Gini, Divergencia y Curva ROC.

**Índice de Términos – Data Mining, Machine Learning, Default, Credit Scoring.**

# introducción

Debido a la crisis financiera de 1998, la Superintendencia Financiera de Colombia (SFC), solicitó a todas las entidades bancarias emplear modelos estadísticos que permitan determinar que un cliente no cumpla con el pago de sus obligaciones. Según la Circular Externa 011 de 2002 de la [Superintendencia Financiera de Colombia (SFC)](https://www.redalyc.org/jatsRepo/2821/282154593002/html/index.html#redalyc_282154593002_ref28), se define “la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que sus deudores fallen en el cumplimiento oportuno o cumplan imperfectamente los contratos de crédito”.

En este sentido, el objetivo general es proponer una metodología para estimar el riesgo de incumplimiento de pago que tienen obligaciones con Banco Colombi. El desarrollo del proyecto de investigación está enfocado hacia tres objetivos: primero identificar las variables que son determinantes en la estimación del incumplimiento de pago de los clientes, puesto que estas variables pueden operar como indicadores de alerta temprana ante cambios en la situación financiera. Como segundo objetivo, estimar el incumplimiento de pago de los clientes del banco de acuerdo a las variables de personalidad, y tercero, encontrar el mejor modelo que clasifique mejor el incumplimiento de pago.

De acuerdo con Atiya (2001) la predicción del incumplimiento de pago ha sido un fenómeno ampliamente estudiado debido a que las entidades necesitan predecir y provisionar sus recursos con la finalidad de evitar caer en bancarrota y maximizar utilidades.

# MARCO TEÓRICO

Los modelos analíticos para el manejo del riesgo de crédito han sido extensamente estudiados y utilizados por un sin fin de áreas como la medicina, sociología y por su puesto por instituciones financieras desde hace más de 50 años, mostrando resultados exitosos en todo el mundo ([Thomas 2004](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B20)).

El identificar aquellos clientes consolidados o potenciales que no generan un negocio a la entidad, requiere buscar el adoptar mejores lineamientos en su tratamiento y reservar base monetaria en caso de incumplimiento y/o migración de buckets (impagos) tolerables a buckets catalogados como vencidos, tal es el caso con las investigaciones de [Hsia (1978)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B8) que describe el aumento del método Scoring en importancia, mientras que otros cálculos son sugeridos por [Reichert, Cho y Wagner (1993)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103" \l "B16), [Joanes (1993)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103" \l "B9), [Hand y Henley (1997)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B7); que detallan el estudio de relaciones particulares entre las distribuciones de clientes buenos y malos que cuentan con la posibilidad de aceptar y rechazar población con características similares.

La historia de los modelos de Scoring se remonta cuando [Fisher (1936)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B5) introduce la idea de discriminar diferentes grupos dentro de una población específica. Esta idea fue desarrollada más ampliamente por [Durand (1941)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B4), aplicando contexto financiero para discriminar entre un “Buen” y un “Mal” pagador, ver de igual manera a [Thomas (2004)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B20). En relación a la creación de proveedores (vendors) de modelos scorging, Bill Fair y Earl Isaac a finales de los cincuentas empezaron con el desarrollo de un sistema analítico en relación al análisis de riesgo de créditos. En los años 60 con la creación de nuevos instrumentos financieros, como, por ejemplo: las tarjetas de crédito, los modelos de Scoring realzaron y mostraron su real importancia y utilidad. Según [Myers (1963)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B15) este tipo de modelos es superior como predictor que cualquier juicio experto cualitativo. Otro hito importante en este contexto fue el desarrollo del Z-Score propuesto por [Altman (1968)](http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-53462016000100103#B1) que ha sido aplicado en muchas empresas del sector financiero.

## Incumplimiento de Pago

## Metodología

# Procedimiento DESARROLLO DEL TRABAJO

## Recolección de fuentes de información

1. **Fuente de información:** Se cuenta con datos de 80 mil clientes del Banco Colombia que cuenta con prestamos crediticios pendientes.

TABLA I

Fuentes de información

|  |  |
| --- | --- |
| Nombre | Descripción |
| score\_train | Información crediticia de 63’901 clientes con prueba escrita de personalidad. |
| score\_test | Información crediticia de 16’116 clientes con prueba escrita de personalidad. |

Fuente: Elaboración Propia

**2. Operacionalización de variables: describimos a las variables por cada fuente de información**

TABLA II

DESCRIPCIÓN BASE SCORE\_TRAIN

|  |  |
| --- | --- |
| Nombre | Descripción |
| Empleado | 0=Desempleado, 1=Empleado actualmente |
| Tiempo\_empleado | Número de años que lleva en el empleo actual (si aplica) |
| Saldo\_cuenta | Saldo en miles de pesos de la cuenta principal con el banco |
| Valor\_prestamo | Valor actual que adeuda en préstamos en el sistema financiero |
| Cuentas\_otros | Número de cuentas que tiene en otros bancos |
| Autocontrol | Mediante una prueba escrita se mide el nivel de autocontrol de la persona, en puntaje de 0 a 100 |
| Impulsividad | Mediante una prueba escrita se mide el nivel de impulsividad de la persona, en puntaje de 0 a 100 |
| Confianza | Mediante una prueba escrita se mide el nivel de confianza en sí misma de la persona, en puntaje de 0 a 100 |
| Incumplimiento | 0=Cumplimiento; 1=Incumplimiento |

Fuente: Elaboración Propia

Las variables presentan las siguientes características:

Dicotómicas: (1) Empleado

Numéricas discretas: (5) Tiempo\_empleo, Cuentas\_otros, Autocontrol, Impulsividad, Confianza.

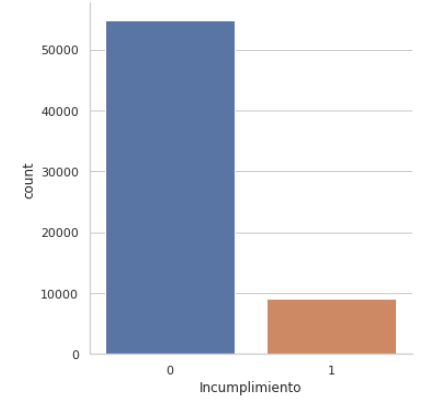
Numéricas continuas: (2) Saldo\_cuenta, Valor\_prestamo

Target: (1) Incumplimiento.

## Análisis exploratorio de los datos

1. **Estadística descriptiva de variables:**

**Target:** clasificación del incumplimiento de pago, está etiquetada como 0 (Cumplimiento) y 1 (Incumplimiento), con un 85.9% y 14.1% respectivamente.



**Variables score\_train:** Las variables del dataset son todas variables numéricas, analizando la distribución:

* Las variables de prueba de personalidad presentan una distribución normal.
* Las variables de tiempo de empleo y saldo en cuenta, tienen una distribución asimétrica negativa.
* La variable valor préstamo, aparentemente tiene una distribución normal, la cual se debe comprobar con la prueba de normalidad.
* La variable de cuentas en otros bancos no se puede definir si tiene distribución normal o no.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

## Supuestos

1. **Normalidad**

Mediante el Test de D’Agostino obtenemos que laa pruebas de personalidad: Impulsividad, Confianza y Autocontrol, evidencian una distribución normal. Rechazando la normalidad de las demás variables.

1. **Homocedasticidad**

Para las variables con distribución Normal se aplica el Test de Bartlet, que nos dice que las pruebas de personalidad: Impulsividad, Confianza y Autocontrol, tienen una varianza constante (homogénea).

Para las variables con distribución No Normal se aplica el Terst de Levene, nos dice que las variables tiene una varianza constante (homogénea).

## Archivos electrónicos de I

## magen (Opcional)

Usted tendrá mayor control sobre la apariencia de sus figuras si usted puede preparar los archivos electrónicos de imagen. Si usted no tiene las habilidades de computación requeridas, sólo envíe las impresiones de papel como se describió anteriormente y salte esta sección.

*1) la Manera más fácil:* Si usted tiene un escáner, la mejor manera y más rápida de preparar los archivos de la figura sin color es imprimir sus tablas y figuras en el papel exactamente como usted quiere que ellas aparezcan, explórelas (con el scanner), y luego guárdelas en un archivo en PostScript (PS) o PostScript encapsulado (EPS). Use un archivo separado para cada imagen. Los nombres de los archivos deben tener el formato "fig1.ps" o "fig2.eps."

2) *La manera un poco más difícil*: Usando un scanner como se describe en el numeral 1, guarde las imágenes en formato TIFF. Alto-contraste línea de línea de figuras y tablas deben prepararse con resolución de 600 dpi y salvadas sin compresión. 1 bit por pixel (monocromo), los nombres de los archivos de la forma "fig3.tif" o "table1.tif.”,). Para obtener un 3,45-en la figura (una-columna de ancho)a 600 dpi la figura requiere un tamaño horizontal de 2070 píxeles. El tamaño típico de los ficheros será del orden de 0,5 MB.

Fotografías y figuras en escala de grises deben prepararse con 220 dpi de resolución y salvadas sin compresión, 8 bits por pixel (escala de grises). ). Para obtener un 3,45-en la figura (una-columna de ancho) a 220 dpi, la figura debería tener un tamaño horizontal de 759 píxeles.

El color de las figuras se deben preparar con 400 dpi de resolución y guardado sin compresión, 8 bits por píxel (paleta ó 256 color). Para obtener un 3,45-en la figura (una-columna de ancho) a 400 dpi, la figura debería tener un tamaño horizontal de 1380 píxeles.   
Para obtener más información sobre los archivos TIFF, por favor vaya a http://www.ieee.org/organizations/pubs/transactions/information.htm y haga clic en el enlace “Guidelines for Author Supplied Electronic Text and Graphics.”

3) Algo mas dificil: Si no dispone de un escáner, puede crear noncolor PostScript figuras por "impresión" a los archivos. En primer lugar, descargar un controlador de impresora PostScript de http://www.adobe.com/support/downloads/pdrvwin.htm (para Windows) o desde http://www.adobe.com/support/downloads/ pdrvmac.htm (por Macintosh) e instalar el "Generic PostScript para imprimir" definición. En Word, pegue la imagen en un nuevo documento. Envia a Imprimir el archivo utilizando el controlador de impresora PostScript. El Nombre de los archivos debe tener el formato "fig5.ps." Utilize fuentes Adobe Tipo 1 al crear su figuras, de ser posible.

4) Otra forma: Usuarios avanzados de la computadora puede convertir figuras y tablas de su formato original al formato TIFF. Algunas convertidores de imágenes útiles son Adobe Photoshop, Corel Draw, y Microsoft Photo Editor, una aplicación que forma parte de Microsoft Office 97 y Office 2000 (busque C: \ Program Files \ Common Files \ Microsoft Shared \ PhotoEd \ PHOTOED.EXE. (Usted tiene la opcion de instalar el photo editor de su disco original de Oficce)



Fig. 1. función de Magnetización. Tenga en cuenta que "Fig." Está abreviado. Hay un espacio después del numero de figura, seguido por dos espacios. Es una buena práctica para explicar la importancia de la figura en el subtítulo.

TABLE I

Units for Magnetic Properties

|  |  |  |
| --- | --- | --- |
| Symbol | Quantity | Conversion from Gaussian and  CGS EMU to SI a |
| Φ | magnetic flux | 1 Mx → 10−8 Wb = 10−8 V·s |
| *B* | magnetic flux density,  magnetic induction | 1 G → 10−4 T = 10−4 Wb/m2 |
| *H* | magnetic field strength | 1 Oe → 103/(4π) A/m |
| *m* | magnetic moment | 1 erg/G = 1 emu  → 10−3 A·m2 = 10−3 J/T |
| *M* | magnetization | 1 erg/(G·cm3) = 1 emu/cm3  → 103 A/m |
| 4π*M* | magnetization | 1 G → 103/(4π) A/m |
| σ | specific magnetization | 1 erg/(G·g) = 1 emu/g → 1 A·m2/kg |
| *j* | magnetic dipole  moment | 1 erg/G = 1 emu  → 4π × 10−10 Wb·m |
| *J* | magnetic polarization | 1 erg/(G·cm3) = 1 emu/cm3  → 4π × 10−4 T |
| χ*,* κ | susceptibility | 1 → 4π |
| χρ | mass susceptibility | 1 cm3/g → 4π × 10−3 m3/kg |
| μ | permeability | 1 → 4π × 10−7 H/m  = 4π × 10−7 Wb/(A·m) |
| μr | relative permeability | μ → μr |
| *w, W* | energy density | 1 erg/cm3 → 10−1 J/m3 |
| *N, D* | demagnetizing factor | 1 → 1/(4π) |

No vertical lines in table. Statements that serve as captions for the entire table do not need footnote letters.

aGaussian units are the same as cgs emu for magnetostatics; Mx = maxwell, G = gauss, Oe = oersted; Wb = weber, V = volt, s = second, T = tesla, m = meter, A = ampere, J = joule, kg = kilogram, H = henry.

## Formulario de Copyright

Un formulario de IEEE de derechos de autor debe acompañar su presentación final. Usted puede obtener una versión. Pdf, Html o. Doc en http://www.ieee.org/copyright. Los autores son responsables de la obtención de cualquier autorización.

# La matemática

Si usted está usando *Word*, use el Editor de Ecuaciones de Microsoft o el complemento *MathType* (http://www.mathtype.com) para las ecuaciones en su documento (Insertar | Objeto | Crear Nuevo | Editor de Ecuaciones de Microsoft o Ecuación MathType).

.

## Abreviaciones y Siglas

Defina las abreviaciones y siglas la primera vez que sean usadas en el texto, incluso después de que se hayan definido en la teoría. Las abreviaciones como ACM, IEEE, SI, ac, y dc no tienen que ser definidas. Las abreviaciones que llevan puntos incorporados no deben tener espacios: escriba “C.N.R.S.,” no “C. N. R. S.” *No use las abreviaciones en el título* a menos que ellas sean inevitables (por ejemplo, “IEEE” en el título de este artículo).

## Ecuaciones

Numere las ecuaciones consecutivamente con los números de la ecuación en paréntesis contra el margen derecho, como en (1). Primero use el editor de ecuaciones para crear la ecuación. Luego seleccione estilo de “Ecuación”. Presione la tecla tab y escriba el número de la ecuación en los paréntesis. Para hacer sus ecuaciones más compactas, usted puede usar (/), la función exp, o exponentes apropiados. Use los paréntesis para evitar las ambigüedades en los denominadores. Puntúe las ecuaciones cuando sean parte de una frase, como en

 (1)

Asegúrese de que los símbolos en su ecuación han estado definidos antes de aparecer la ecuación o inmediatamente enseguida. Ponga en cursiva los símbolos (*T* podría referirse a la temperatura, pero T es la unidad tesla). Refiérase a “(1),” no a “Eq. (1)” o “la ecuación (1),” excepto al principio de una oración: “la Ecuación (1) es....”

# RESULTADOS DE LOS MODELOS DE MACHINE LEARNING

Presentación de un manuscrito no es necesaria para la participación en una conferencia. No envíe una versión de una nueva presentación de un documento que usted ha enviado o ha publicado en otra parte. No publique datos o resultados “preliminares”. El autor que remite es el único responsable para estar de acuerdo con todos los coautores y cualquier consentimiento requerido de los patrocinadores antes de enviar un documento (paper). La IEEE rechaza radicalmente la paternidad literaria de cortesía. Es obligación de los autores citar el trabajo previo pertinente.

# DISCUSIÓN

El contenido de TRANSACTIONS y JOURNALS es revisado y archivado por expertos.

Los autores deben considerar los siguientes puntos:

1) Los documentos técnicos enviados para publicación deben adelantar el estado de conocimiento y deben citar el trabajo previo pertinente.

2) La longitud de un documento enviado debe ser correspondiente con la importancia, o apropiado a la complejidad, del trabajo. Por ejemplo, una extensión obvia de trabajo previamente publicado no podría ser apropiada para la publicación o podría tratarse adecuadamente en sólo unas páginas.

3) Los autores deben convencer al Comité Editorial, por medio de su documento, del mérito académico, científico o técnico del documento; las normas de evaluación son más exigentes cuando se reportan resultados extraordinarios o inesperados.

### 

4) Debido a que la repetición se requiere para el progreso científico, los documentos enviados para publicación deben proporcionar información suficiente para permitirles a los lectores tener acceso a las referencias utilizadas, especialmente si estas son URLs, realizar experimentos similares o cálculos y usar los resultados informados. Aunque no todo necesita ser descubierto, un documento debe contener información nueva, usada y totalmente descubierta. Por ejemplo, la composición química de un espécimen necesita que no se informe si el propósito principal de un documento es introducir una nueva técnica de la medida. *Los autores deben esperar ser desafiados por críticos si los resultados no son soportados por los datos adecuados y los detalles críticos*.

5) Documentos que describen el trabajo en curso o muestran un reciente logro técnico, que sean adecuados para su presentación en una conferencia profesional, pueden no ser apropiadas para su publicación en un TRANSACTIONS or JOURNAL

IX. CONCLUSIÓN

Una sección de conclusión no es necesaria. Sin embargo esta puede repasar los puntos principales del artículo, no repita el resumen como conclusión. Una conclusión se elabora con base en la importancia del trabajo realizado o en las aplicaciones y extensiones sugeridas.

References

1. Atiya, A. F. (2001). Bankruptcy Prediction for Credit Risk Using Neural. IEEE TRANSACTIONS ON NEURAL NETWORKS, 929-935.
2. Thomas, L. C., Edelman, D. B., Crook, J. N., (2004), Readings in Credit *Scoring*, Oxford University Press, Oxford.
3. Hsia, D. C. (1978). Credit *Scoring* and the Equal Credit Opportunity act. Hast. Law. J., 30, pp. 371-448.
4. Reichert, A. K., Cho, C. and G. M. Wagner, (1983). An Examination of Conceptual Issues Involved in Developing Credit-*Scoring* Models. J. Bus. Econ. Stat., 1 (2), pp. 101-14.
5. Joanes, D. N. (1993). Reject Inference applied to Logistic Regression for Credit Scoring. IMA J. Math. Appl. Bus. Industry, 5, pp. 35-43.
6. Hand, D. J., and W. E. Henley (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), pp. 160: 523-541.
7. Durand, D. (1941). Risk Elements in Consumer Instalment Financing, Studies in Consumer Instalment Financing. New York: National Bureau of Economic Research.
8. Fisher R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7(2), pp. pp. 179-188.
9. Myers, J. H. and Forgy, E. W. (1963). The Development of Numerical Credit Evaluation Systems. J. Am. Statist. Ass., 58, pp. 799-806.
10. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), pp. 589-609.

1. [↑](#footnote-ref-1)