

Diseño de un modelo matemático para la calificación de clientes morosos en una entidad comercial mediante las metodologías de árboles de decisión, análisis discriminante y regresión logística.

Design of a mathematical model for the qualification of delinquent clients in a commercial entity through the methodologies of decision trees, discriminant analysis and logistic regression.

Hugo Renán Ruiz López

Universidad Internacional del Ecuador, Ecuador

Autor por correspondencia: huruizlo@uide.edu.ec

Fecha de recepción: 30 de Abril de 2017 - Fecha de aceptación: 30 de Junio de 2017

Resumen

En el presente trabajo se diseña un modelo matemático para la calificación de clientes morosos con la finalidad optimizar la gestión de cobranzas en una empresa comercial. Para ello desarrollamos dos modelos paramétricos, Análisis Discriminante y Regresión Logística y un modelo no paramétrico, los Árboles de Decisión. Se determina la estabilidad de cada modelo, su poder de discriminación y por último establecer cuál de las metodologías permite optimizar mejor la gestión de cobro.

Palabras claves: riesgo de crédito; árboles de decisión; análisis discriminante; regresión logística; morosidad

Abstract

In the present work, a mathematical model is designed for the qualification of delinquent clients in order to optimize the management of collections in a commercial enterprise. To do this we developed two parametric models, Discriminant Analysis and Logistic Regression and a non-parametric model, the Decision Trees. It determines the stability of each model, its power of discrimination and finally establish which of the methodologies allows better optimization of collection management.

Key words: credit risk; decision trees; discriminant analysis; logistic regression; delinquency

Introducción

La situación economía en el Ecuador en dos últimos años ha tenido un impacto importante en los clientes con crédito a la hora de cancelar sus cuentas. Es así que ciertas entidades comerciales implementaron planes de refinanciamientos masivos para mantener sus niveles de morosidad dentro de parámetros estables y manejables.

La elección de un modelo matemático que permita priorizar la gestión de cobro en edades tempranas de morosidad es de vital importancia en las empresas para lograr la recuperación de créditos, de manera que los activos exigibles o deudas se conviertan en activos líquidos de manera eficiente y rápida (Acción InSight, 2008). Lo adecuado es tratar de mantener una buena relación con el cliente para no romper el vínculo Cliente-Empresa para futuras relaciones comerciales.

Este trabajo está centrado en tres metodologías (de Árboles de Decisión, Análisis Discriminante y Regresión Logística) que actualmente se utilizan con mucho éxito en el diagnóstico del riesgo crediticio en las instituciones financieras y en la mayoría de empresas comerciales, que en el crédito tienen su mejor arma de crecimiento, sin embargo debe ser utilizado con mucho cuidado ya que la venta concluye cuando el cliente termina de pagar su obligación crediticia.

Gestión de Cobranzas

La gestión de cobranzas es un conjunto de acciones aplicadas oportunamente a los clientes para lograr la recuperación eficiente de los créditos vencidos, cuyo objetivo principal es reactivar la relación comercial entre Cliente-Empresa (Acción InSight, 2008).

En la gestión de cobranzas pueden tomarse acciones preventivas y curativas.

Las acciones preventivas tienen el objetivo de gestionar clientes con niveles de alto riesgo de morosidad (Añez, 2004), cuyo propósito es la reducción de la probabilidad que el cliente entre en la cartera morosa.

En el caso de las acciones curativas tienen como finalidad la recuperación de la deuda vencida (Añez, 2004), gestión que puede iniciarse desde el primer día del incumplimiento del pago. El propósito es iniciar la gestión con clientes de niveles de riesgo más alto de morosidad.

La gestión de cobranzas puede utilizar diferentes medios como son:

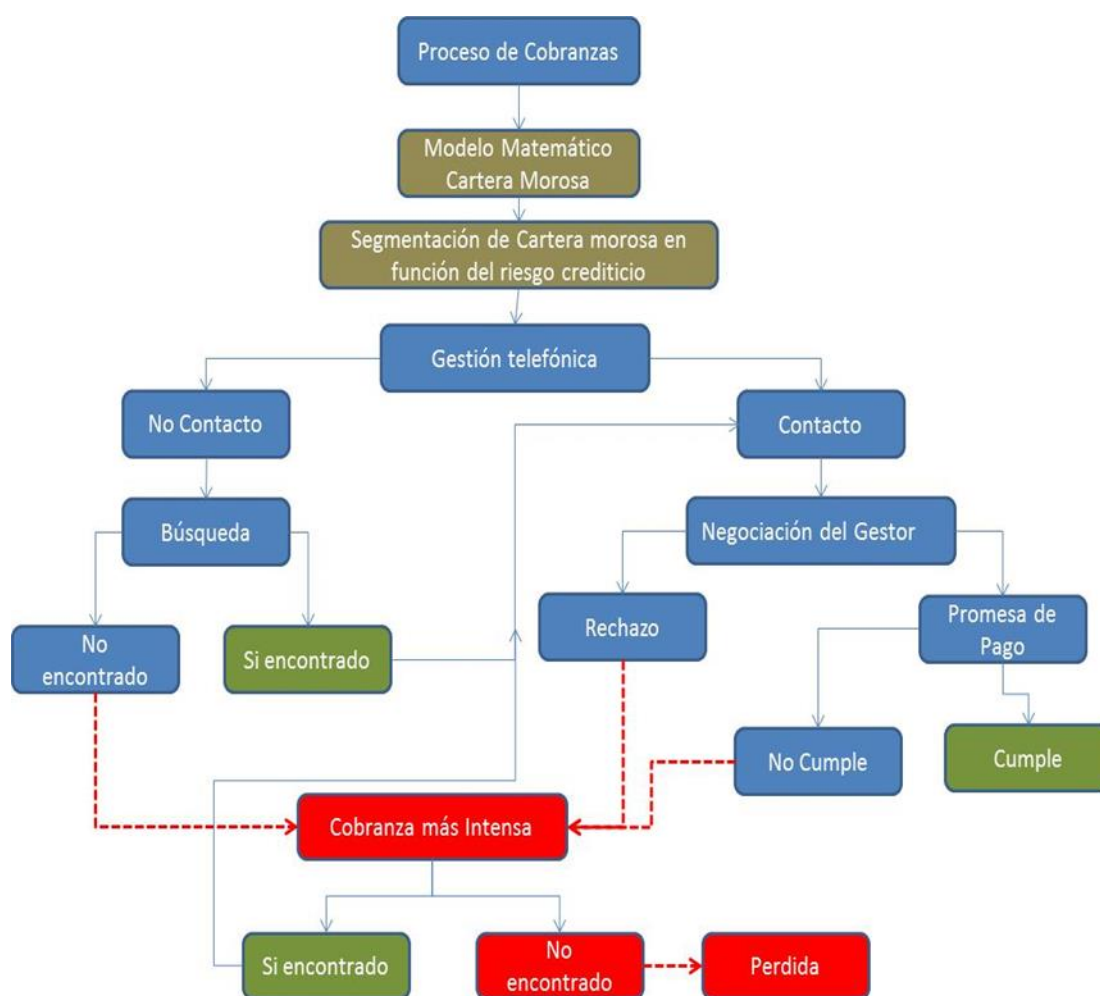
- Gestión telefónica
- Cartas
- Mensajes de texto
- Correos Electrónicos
- Visitas de gestores de campo a domicilio
- Mensajes Whatsapp

La gestión de cobranzas es un proceso que tiene interacción constante con el cliente, que como se mencionó inicia con el análisis de riesgo, para realizar un oportuno contacto.

Una política clara de cada empresa es que la recuperación de la cartera se realice sin afectar la relación con el cliente, el personal que realiza la gestión de cobranza debe tener mucho cuidado de no ser demasiado agresivo con el cliente (Acción InSight, 2008).

Toda gestión de cobranzas puede incluir un proceso de control de calidad que permita evaluar, retroalimentar y mejorar las prácticas empleadas. Esta implementación permite a la compañía identificar las necesidades de los clientes y hacia donde encaminar las alternativas de solución, además ayuda a conocer si el personal de cobro actúa bajo los lineamientos de la compañía.

Proceso de cobranza de cartera vencida que incluye el proceso de calificación de clientes en función de su riesgo antes de realizar la gestión. (Gráfica 1)



Gráfica 1: Proceso de Cobranzas

Presentada la importancia de realizar una gestión de cobro focalizada y los riesgos que implica no realizarla, el presente trabajo se complementa con un estudio empírico, cuyo objetivo principal es encontrar segmentos de clientes con base a su comportamiento crediticio, que nos indique el nivel de riesgo de cumplir o no las obligaciones pendientes.

Lo importante es tratar de determinar los factores que nos ayuden a entender el comportamiento de pagos de los clientes en las empresas comerciales. El objetivo se orienta en explicar el comportamiento de una variable binomial categórica que será: cancela o no cancela el valor vencido.

Adicionalmente realizaremos un análisis comparativo de las metodologías utilizadas, con la finalidad de determinar cuál de estas técnicas aplicadas se ajusta mejor al problema de recuperación de cartera.

Construcción de modelos mediante las metodologías matemáticas

El presente trabajo compara las metodologías de Árboles de Decisión, Análisis Discriminante y Regresión Logística, con la finalidad de determinar cuál de dichas técnicas, nos permite comprender y entender el proceso de morosidad en las instituciones comerciales, específicamente en edad temprana, tomando como referencia una muestra de clientes vencidos a 60 días.

A continuación, presentamos los conceptos básicos para la construcción y validación de los modelos estadísticos:

Definición de Cliente Bueno y Cliente Malo: Es la variable binaria, donde considerará cliente bueno aquel que durante el mes de gestión realiza un pago, caso contrario se lo considera malo.

Variables de Comportamiento: Se generó un total de 5 variables de comportamiento crediticio de los clientes dentro de la institución durante el último año. A continuación, se presenta el código y descripción de cada variable.

- N_PAGOS_6M: Número de pagos en los últimos 6 meses.
- MMORMX12: Mora máxima alcanzada en el último año.
- MSAL_CUP: Saldo actual dividido para el cupo.
- PAGO_1M: Monto cancelado en el último mes.
- MNM030_6: Número de moras en 30 días en los últimos 6 meses.

Variables Descriptivas: Se consideraron para el estudio a parte de las variables de comportamiento, las variables descriptivas como sexo, estado civil y antigüedad del cliente en la institución.

Muestra: Se tomó dos muestras de clientes vencidos, una para la construcción del modelo matemático de 3.000 mil clientes y otra independiente de 2.000 clientes para la comprobación del mismo. El porcentaje de clientes buenos (cliente que realizan pagos) fue del 63% en ambas muestras.

Prueba Kolmogorov-Smirnov (K-S): Es una prueba que se basa en medir la separación de las distribuciones acumuladas de los buenos y malos para cada rango percentil del puntaje. A

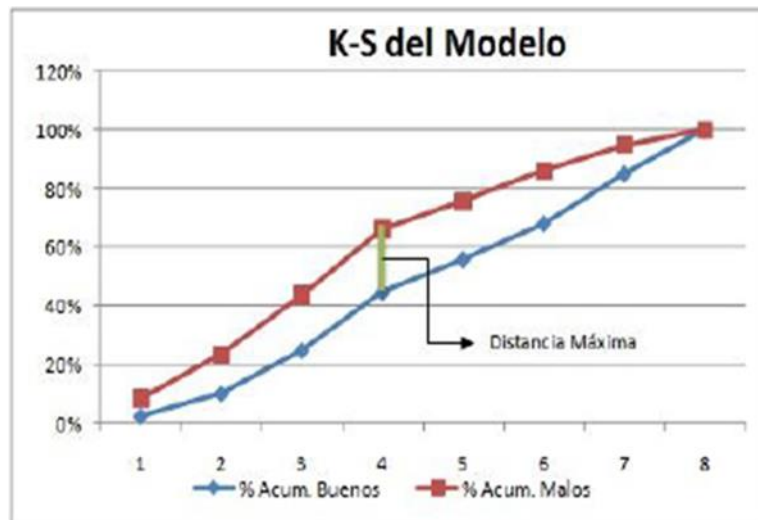
mayor K-S, el modelo matemático permitirá una mejor discriminación entre los grupos de clientes buenos y malos. El estadístico K-S está dado por siguiente ecuación:

$$K - S = \text{MAX}|F_b x - F_m x|$$

Donde

$F_b x$: es la distribución acumulada de los clientes buenos

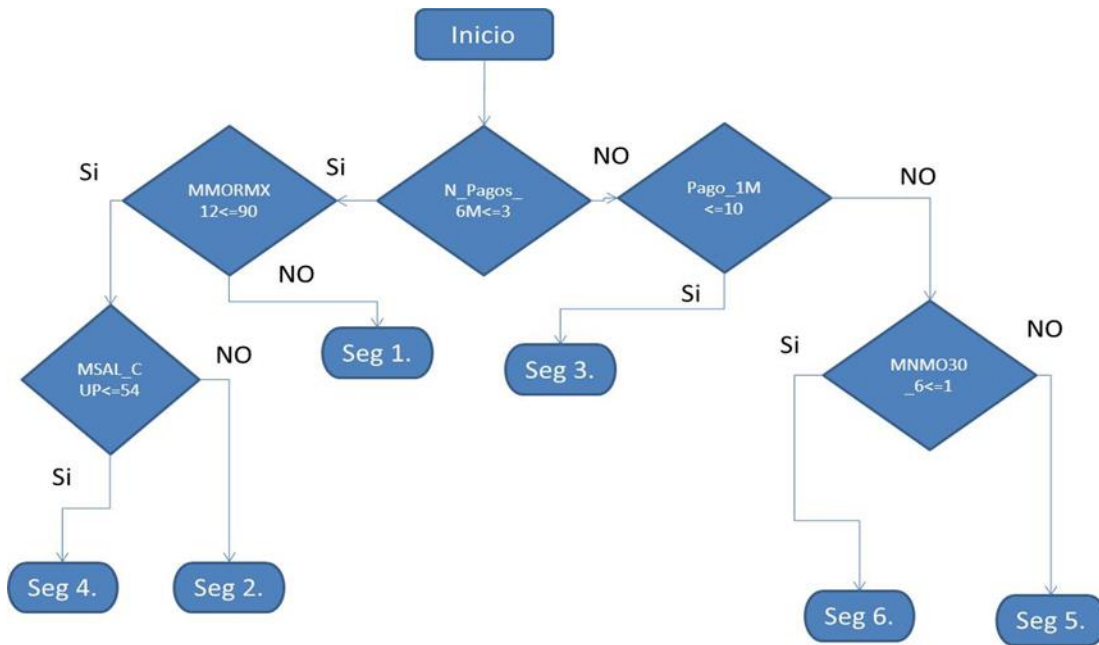
$F_m x$: es la distribución acumulada de los clientes malos



Gráfica 2: K-S gráficamente

Modelo de Árbol de Decisión

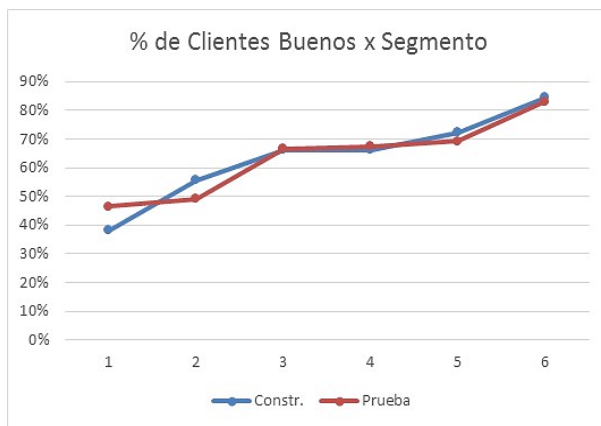
Los árboles de decisión binarios son métodos no paramétricos que tienen la ventaja de no requerir supuestos distribucionales, inventado en el año 500 por lógicos sirios (Ruiz, 2013). Tiene como su principio básico generar particiones recursivas de la muestra hasta llegar a una clasificación final. El programa estadístico SPSS se utilizó para determinar el árbol de decisión, el cual compuesto por 6 segmentos, los que se describe a continuación en el siguiente gráfico:



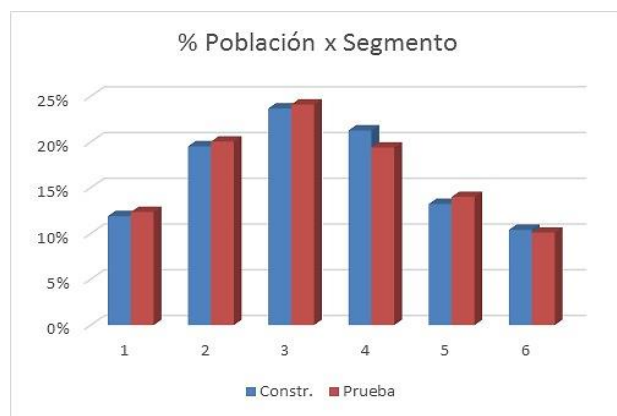
Gráfica 3: Modelo Árbol de Decisión

Los dos primeros segmentos contienen una tasa de clientes buenos inferior al promedio general que es del 63%, la característica principal de estos dos segmentos es que número de pagos en los últimos 6 meses en menor igual a 3, es decir que son clientes que en los últimos 6 meses se han encontrado vencidos. Mientras que los clientes de los dos últimos segmentos contienen una tasa de buenos muy superior al promedio general del 63%, la característica de estos segmentos es que a pesar de estar vencidos ha realizado un abono a su deuda en el último mes mayor a \$ 10 dólares.

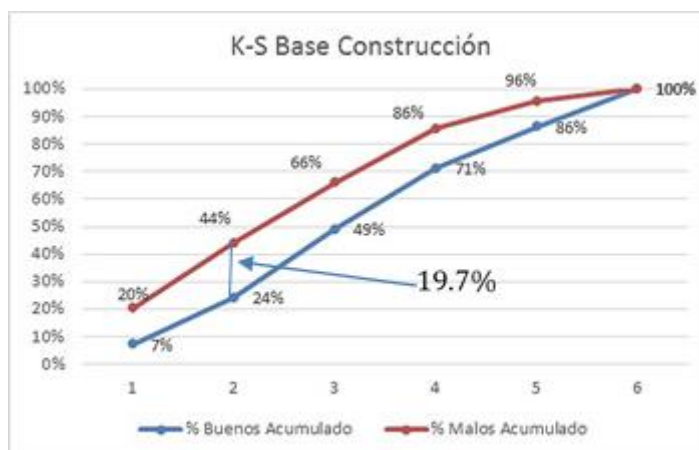
A continuación, se presenta los gráficos de tasa de bueno (clientes que han cancelado por segmento) y participación en el porcentaje de población, adicionalmente se puede apreciar que el modelo mantiene el mismo comportamiento en la base de construcción y de prueba.



Gráfica 4: Tasa de Clientes Buenos por Segmento



Gráfica 5: Porcentaje de Clientes por Segmento



Gráfica 6: K-S Modelo Árbol de decisión

Modelo Análisis Discriminante

El análisis discriminante es una técnica estadística (Bonilla, Omeda, & Puertas, 2003), en el cual la variable independiente es una variable no numérica, cuya finalidad es analizar y determinar cuál o cuáles variables contribuyen a discriminar entre dos o más grupos de estudio.

El programa utilizado para el estudio fue SPSS, y el criterio de clasificación empleado fue de minimización de lambda de Wilks para la presentación de resultados. Los coeficientes de la función discriminante se dan a continuación:

Tabla 1: Coeficientes Análisis Discriminante

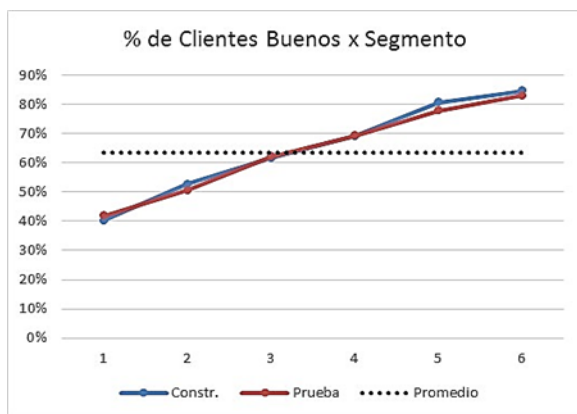
Variables	Coeficientes F. Discriminantes	Coeficientes Estandarizado	Lambda de Wilks	F exacta	Sig.
MNMO30_6	0,258	0,282	0,994	31,300	0,000
N_Pago_1M	-0,325	-0,213	0,983	85,811	0,000
N_Pagos_6M	-0,260	0,600	0,968	164,165	0,000
MSAL_CUP	0,001	0,392	0,986	68,438	0,000
MMORMX12	0,010	0,403	0,975	127,433	0,000
(Constante)	-0,945				

Al utilizar los coeficientes de la función discriminante se obtiene la puntuación para cada cliente, con base a esta puntuación los clasificamos en seis grupos de acuerdo a la proporción obtenida con el modelo de Árboles de Decisión para realizar una mejor comparación entre modelos.

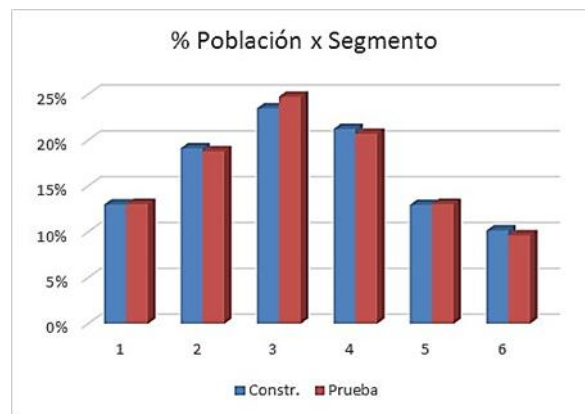
Tabla 2: Clasificación de Segmentos Análisis

Segmento	Puntaje
1	>1,04
2	>0.38 y <=1,04
3	>-0.14 y <=0.38
4	>-0.65 y <=-0.14
5	>1.16 y <=-0.65
6	<=- 1.16

Como podemos apreciar en los gráficos siguientes la tasa de buenos y los porcentajes de población por segmento tanto en la base de construcción como en la de prueba se mantienen estables.

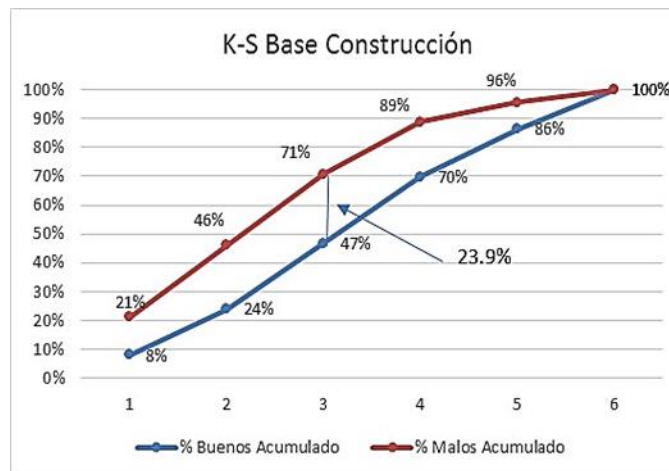


Gráfica 7: K-S Modelo Análisis Discriminante



Gráfica 8: Porcentaje de Clientes por Segmento

El K-S del modelo en la base construcción fue del 23,9%, mientras que en la base de prueba fue del 22,4%.



Gráfica 9: K-S Modelo Análisis Discriminante

Modelo Regresión Logística

La regresión logística permite determinar una probabilidad de que el cliente realice el pago o no, es un tipo de regresión especial que se utiliza para explicar y predecir una variable categórica (Mures, García, & Vallejo, 2005), en función de varias variables independientes que pueden ser cualitativas o cuantitativas.

El programa utilizado para el estudio fue SPSS, y el criterio de clasificación empleado fue mediante el estadístico de Wald para la presentación de resultados. Los coeficientes de la función se dan a continuación:

Tabla 3: Coeficientes Regresión Logística

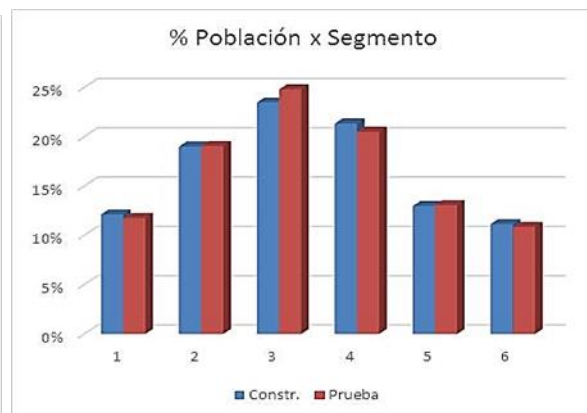
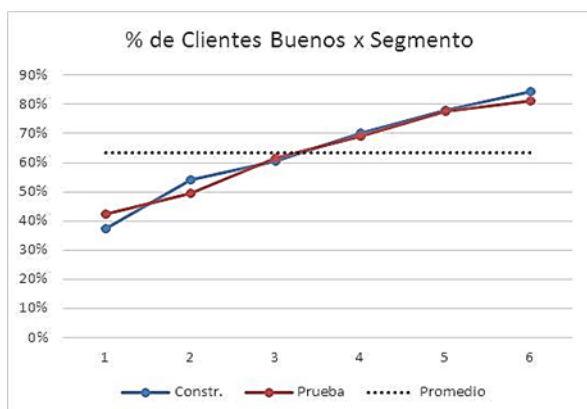
Variables	B	Wald	Sig.
MNMO30_6	-0,158	24,400	0,000
N_Pago_1M	0,236	12,945	0,000
N_Pagos_6M	0,175	79,455	0,000
MSAL_CUP	-0,001	47,119	0,000
MMORMX12	-0,005	36,184	0,000
(Constante)	0,972	88,378	0,000

Al igual que el análisis discriminante se tomó como referencia el número de segmentos y su composición en población del modelo de árboles de decisión para clasificar a los clientes en seis grupos, de acuerdo a su probabilidad obtenida en la función logística.

Tabla 4: Clasificación de Segmentos Regresión Logística

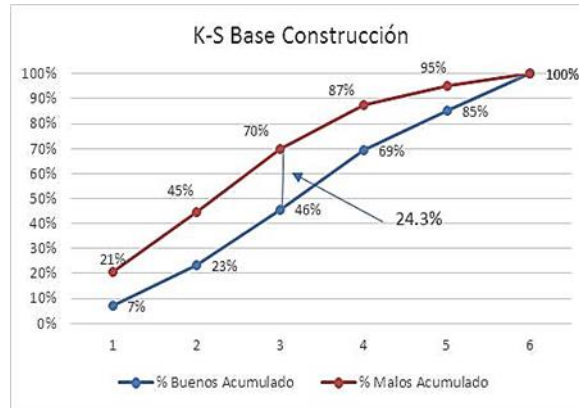
Segmento	Probabilidad
1	≤ 0.48
2	> 0.48 y ≤ 0.57
3	> 0.57 y ≤ 0.65
4	> 0.65 y ≤ 0.72
5	> 0.72 y ≤ 0.78
6	> 0.78

Los siguientes gráficos muestran la estabilidad del modelo tanto en la tasa de buenos y participación de población por segmentos.



Gráfica 10: K-S Modelo Regresión Logística Gráfica 11: Porcentaje de Clientes por Segmento

El K-S del modelo en la base construcción fue del 24,3%, mientras que en la base de prueba fue del 22,2%.

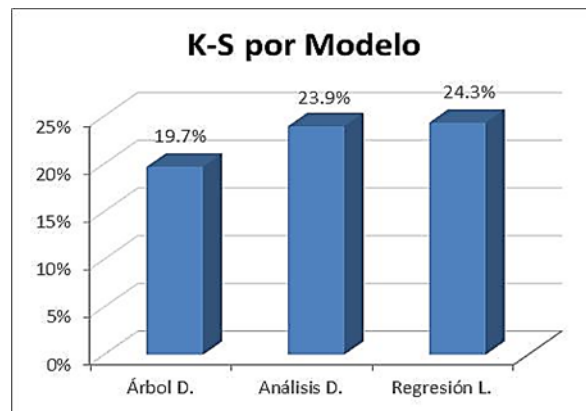


Gráfica 12: K-S Modelo Regresión Logística

Comparación de Resultados entre Modelos Matemáticos

Una vez presentados los modelos matemáticos y efectuada la validación de los mismos con la base prueba, se procede a realizar las comparaciones en tasa de buenos por segmento y en el poder de discriminación de los mismos.

K-S: Como se mencionó una medida para determinar el poder de discriminación de un modelo es el K-S, esta medida con la metodología de Regresión Logística registro el coeficiente más alto con el 24.3%, seguido por el Análisis Discriminante con el 23.9% y por último los Árboles de Decisión se obtiene un coeficiente del 19.7%.



Gráfica 13: K-S por Modelo Matemático

Tasa de Buenos por Segmentos: El porcentaje de clientes buenos (clientes que realizan pagos) por segmentos es prácticamente similar entre las metodologías de Análisis Discriminante y Regresión Logística, no así con el modelo de Árboles de Decisión que tienen diferencias en ciertos segmentos como en el tres y cinco.



Gráfica 14: K-S por Modelo Matemático

Porcentaje de Coincidencia entre segmentos: Los tres modelos están compuestos por 6 segmentos, para realizar una mejor comparación entre ellos se los agrupo en niveles de riesgo de acuerdo a su nivel de pagos.

Tabla 5: Clasificación del nivel de Riesgo

Segmento	Riesgo	% Pagos		
		Árbol. D.	Regresión L.	Análisis D.
1	Alto	49%	47%	47%
2				
3	Medio	67%	65%	65%
4				
5	Bajo	77%	81%	81%
6				

Análisis Discriminante vs Regresión Logística: El nivel de coincidencia en la asignación del riesgo a los clientes entre estos dos modelos es del 95%, es decir que 4.751 de los 5.000 clientes tienen asignado el mismo riesgo. Se puede concluir que ambos modelos tienen la misma capacidad predictiva.

Tabla 6: Comparación Regresión L. vs Análisis D.

Regresión L.	Análisis D.			Total general
	Alto	Medio	Bajo	
Alto	1,504	45		1,549
Medio	97	2,126	27	2,250
Bajo		80	1,121	1,201
Total general	1,601	2,251	1,148	5,000

Regresión Logística – Análisis Discriminante vs Árbol de Decisión: El nivel de coincidencia en la asignación del riesgo a los clientes entre los dos modelos paramétricos (Regresión L. y Análisis D.) versus el modelo no paramétrico (Árboles D.) es del 68%. En las siguientes tablas se puede observar diferencias en la asignación del riesgo entre estos modelos matemáticos.

Tabla 7: Comparación Regresión L. vs Árboles D.

Regresión L.	Árbol D.			Total general
	Alto	Medio	Bajo	
Alto	1,141	358	50	1,549
Medio	444	1,474	332	2,250
Bajo	6	386	809	1,201
Total general	1,591	2,218	1,191	5,000

Tabla 8: Análisis D. vs Árboles D.

Análisis D.	Árbol D.			Total general
	Alto	Medio	Bajo	
Alto	1.150	367	84	1.601
Medio	433	1.456	362	2.251
Bajo	8	395	745	1.148
Total general	1.591	2.218	1.191	5.000

Conclusiones

En este trabajo presentamos tres metodologías diferentes para optimizar la prioridad de cobro de una cartera con mora temprana, los cuales resultaron muy significativos y estables, tal como lo muestran los estadísticos calculados y la respectiva validación con la muestra de prueba.

Este estudio demuestra la factibilidad de usar cualquier de estas tres metodologías estadísticas estudiadas para realizar el diagnóstico en instituciones o empresas, cuyo producto principal de ventas sea el crédito.

En términos de K-S la metodología de Regresión Logística registró un mayor poder de discriminación con un K-S del 24.3%, mientras que los Árboles de Decisión obtienen el menor K-S entre los tres modelos con el 19,7%.

Las metodologías paramétricas Regresión Logística y Análisis Discriminante alcanzan el 95% de coincidencia en calificar a los clientes por niveles de riesgo.

A pesar que los árboles de decisión obtienen menor poder de discriminación que los dos métodos paramétricos estudiados, estos tienen la ventaja que sus segmentos tienen una descripción explícita de sus perfiles, mientras que las otras técnicas asignan un puntaje o una probabilidad.

La experiencia obtenida en la construcción de estos modelos nos enseña que deben ser evaluados y monitoreados de forma continua su nivel de predicción, con la finalidad de detectar anomalías tempranas y tomar las medidas correctivas de forma oportuna.

Bibliografía

Acción InSight. (2008). Mejores Prácticas en Estrategias de Cobranza. Acción InSight.

Añez, A. (2004). Riesgo Crediticio.

- Boj del Val, E., Claramunt, M., Esteve, A., & Fortina, J. (2009). Criterios de selección de modelo en el credit scoring. www.researchgate.net.
- Bonilla, M., Omeda, I., & Puertas, R. (2003). Modelos paramétricos y no paramétricos en los problemas de crédito. *Revista Española de Financiación y Contabilidad*, 833-869.
- Cardona, A. (2004). Aplicación de árboles de decisión en modelos de riesgo de crédito. *Revista Colombiana de Estadística*. *Revista Colombiana de Estadística*, Volumen 27, pp. 139-151.
- Dupouy, C. (2014). Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile. *Economía y Negocios*, Universidad de Chile.
- García, J. C. (s.f.). Nuevas técnicas de medición del riesgo de crédito. Departamento de Metodologías de Riesgo Corporativo. BBVA.
- Llaugel, F., & Fernández, A. (2001). Evaluación del uso de modelos de regresión logística para el diagnóstico de instituciones financieras. *Revista Ciencias y Sociedad*.
- Mures, J., García, A., & Vallejo, E. (2005). Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras. *Pecvnia*, 175-199.
- Ríos, A., Arguedas, R., & Casals, J. (2010). *Gestión y control del riesgo de crédito con modelos avanzados*. Ediciones Académicas S.A Madrid, ISBN: 978-84-92477-35-7.
- Ruiz, H. (2013). Diseño de un modelo matemático para optimizar la gestión de un sistema de cobranzas en una empresa comercial. *Escuela Politécnica del Litoral*. Ecuador.
- Saavedra, M. L., & Saavedra, M. J. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuad. Adm. Bogotá*, Volumen 23, pp. 295-391.
- Viroto, M. N. (2014). Factores que influyen en el endeudamiento por tarjeta de crédito en casas comerciales y default. *Economía y Negocios*, Universidad de Chile.