Modelo de credit scoring para la empresa Grupo Factoring de Occidente S. A. S.

Carolina Santiago Sandoval
Alejandro Urán Vélez

Trabajo presentado como requisito parcial para optar al título de
Magíster en Administración Financiera

Asesor: Brayan Rojas

UNIVERSIDAD EAFIT
ESCUELA DE ECONOMÍA Y FINANZAS
MAESTRÍA EN ADMINISTRACIÓN FINANCIERA
SANTIAGO DE CALI
2021

Resumen

El riesgo de crédito consiste en el establecimiento de políticas y procedimientos que se adecúan a la reglamentación vigente y al perfil de riesgo de los accionistas e intenta minimizar la probabilidad de ocurrencia de situaciones que pongan en peligro los recursos de las organizaciones; este se da por medio de un adecuado porcentaje de provisión de cartera por incumplimiento de los pagadores. El presente trabajo tiene como propósito contribuir al desarrollo de un modelo de *credit scoring* que permita realizar seguimiento a los clientes de la compañía Grupo Factoring de Occidente S. A. S. con el fin de analizar el riesgo de incumplimiento usando las variables más relevantes y significativas, tanto de los clientes como del sector y del entorno mediante el uso de modelos econométricos ya probados anteriormente.

Palabras clave: credit scoring, modelo econométrico, prima de riesgo, logit, árbol de decisión.

Abstract

Credit risk consists of the establishment of policies and procedures, wish are appropiate to current regulations and to the shareholder’s risk profile and tend to minimize the probability of occurrence of situations that put the companie´s resources at risk through an adequate percentage of portfolio provision due to default of payers. The present research aims to contribute to the development of a credit scoring model that allows to do tracking to the company´s customers “Grupo Factoring de Occidente SAS” to analyze the risk of payment failures, using the most relevant and significant variables of both the clients, the sector and the environment, through the use of previously proven econometric models.

Keywords: credit scoring, econometric model, risk rating, logit, decision tree.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS|

1. INTRODUCCIÓN

La entrada en vigor de las Normas Internacionales de Información Financiera (NIIF) y la adopción de las recomendaciones hechas por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en el marco del Comité de Supervisión Bancaria de Basilea, por parte del gobierno colombiano y sus instituciones, han exigido que las instituciones financieras se pongan a la vanguardia en los temas referentes a las buenas prácticas desarrolladas en el mundo, en este caso, en cuanto al establecimiento de procedimientos y políticas que permitan de una manera fiable medir, controlar y mitigar el riesgo de crédito. Así se establece que las instituciones financieras deben medir su riesgo de crédito de cartera por medio de diferentes métodos de acuerdo con su nivel de sofisticación, entre ellos se proponen los métodos estándar, el método basado en calificaciones internas básico y el método basado en calificaciones internas avanzado, los cuales pretenden medir el requerimiento de capital para una operación de crédito tomando como base la probabilidad de impago. Si bien la empresa objeto de estudio no es considerada por la regulación bancaria colombiana como una institución financiera, se considera necesario empezar a adoptar la recomendación en cuanto a la medición del riesgo de crédito y la probabilidad de impago de los deudores de la entidad (Bank for Internacional Settlements, 2017).

En este documento se pretende desarrollar un modelo que le permita a la empresa Grupo Factoring de Occidente S. A. S. (de ahora en adelante GFO), basado en modelos econométricos, solucionar el interrogante: ¿cuál es la probabilidad de impago del cliente? esto con el fin de que la compañía pueda, mediante dicho modelo, minimizar las probabilidades de ocurrencia de eventos nocivos y estimar posibles pérdidas acordes al nivel de riesgo de la compañía, así como empezar con la implementación de sistemas de administración de riesgo crediticio, ya que si bien la compañía no se encuentra vigilada por la Superintendencia Financiera de Colombia, estas buenas prácticas en administración de riesgo aportan para ponerla en otro nivel.

En Colombia se han desarrollado trabajos valiosos sobre el tema, enfocados en instituciones financieras, como es el caso del artículo de González (2010), en donde se identifica cómo algunos indicadores financieros de las compañías y algunos indicadores macroeconómicos como crecimiento del PIB, inflación y desempleo pueden tener efectos significativos o, en otros casos, marginales al momento de determinar la probabilidad de impago de los clientes en la cartera comercial del sector financiero.

Por su parte, Salazar (2013) evidencia la utilidad del uso de modelos econométricos para analizar el riesgo de crédito, permitiendo demostrar cómo la cartera vencida es elástica al ciclo económico; lo que quiere decir que el riesgo de incumplimiento no solo está asociado a la estructura financiera del cliente, sino también al entorno macroeconómico, propiamente al ciclo económico. Adicionalmente, muestra cómo variables tales como género, sector económico, endeudamiento paralelo y fallas de información son determinantes al momento de evaluar el riesgo de incumplimiento del cliente, por lo que el autor sugiere plantear acciones gerenciales para controlar más el riesgo ligado a estas variables.

Para desarrollar su trabajo, Salazar utiliza dos tipos de modelos estadísticos, el logit binomial para la definición de los determinantes microeconómicos del riesgo de crédito y el cálculo de la probabilidad de incumplimiento, y un modelo log-log que permite establecer los factores explicativos del default crediticio a una escala macroeconómica. Así da una aproximación a dos modelos usados para estimar la probabilidad de incumplimiento en los créditos y aclara cuáles son los principales determinantes del incumplimiento en el crédito en Colombia, al ser un estudio desarrollado en la ciudad de Popayán.

Para lograr el objetivo planteado del presente proyecto, se hará uso de la información histórica con que cuenta la empresa GFO, como las tasas asignadas a cada uno de los clientes, el monto de cada una de las operaciones, el plazo pactado para el pago, así como el cumplimiento de este; también se recurrirá a las variables explicativas como la información financiera a partir de la cual se determinan los

indicadores financieros, sector al que pertenece el cliente, calidad en el reporte de la información financiera y condiciones actuales del cliente. El modelo desarrollado debe permitir sintetizar toda esta información, asignándole un peso a cada una de las variables previamente mencionadas para poder establecer la probabilidad de incumplimiento del cliente.

Los resultados obtenidos a partir del modelo econométrico se contrastarán con los datos que actualmente maneja la empresa de sus clientes, esto permitirá confirmar o debatir qué tan acertado es el manejo del riesgo de crédito de la compañía GFO.

Para el desarrollo de los objetivos planteados se seguirá la siguiente estructura: después de la introducción se revisará la literatura existente para modelos de riesgo de crédito mediante los modelos de árboles de decisión y de regresión logística binaria; desarrollado este paso, a partir de la información suministrada por la empresa, se determinará cuáles son las variables cuantitativas y cualitativas que serán definitivas para el desarrollo de los modelos; la tercera parte estará dedicada a la construcción de los modelos y sus respectivas pruebas de valoración y validación por medio de test estadísticos y pruebas de backtesting, para posteriormente emitir las conclusiones y recomendaciones propias del caso.

2. OBJETIVOS

2.1 Objetivo general

Desarrollar un modelo econométrico de seguimiento que pueda predecir la probabilidad de incumplimiento de los clientes de GFO a partir de su información histórica.

2.2 Objetivos específicos

- Definir cuáles son las variables cuantitativas y cualitativas determinantes en los modelos que se van a plantear.
- Desarrollar los modelos logit y árboles de decisión de acuerdo con la información definida para el objeto de estudio.
- Medir la efectividad de los modelos por medio de procedimientos estadísticos y pruebas de backtesting.

3. MARCO TEÓRICO

Para realizar el presente trabajo se considerarán diferentes modelos econométricos que permitan desarrollar un método apropiado de medición y mitigación del riesgo de crédito; a continuación, se detallan algunos conceptos clave y métodos que se deben indagar:

3.1 El riesgo de crédito

“La palabra riesgo proviene del latín *riscaré*, que significa atreverse a transitar por un sendero peligroso. En realidad, tiene una connotación negativa, sin embargo, el riesgo es parte inevitable de los procesos de toma de decisiones en general y de los procesos de inversión en particular” (Haro, 2008, s. p.); por lo tanto, el origen del riesgo es tan antiguo como el riesgo de crédito y se puede definir como la probabilidad de incumplimiento de pago de una operación financiera de acuerdo con los términos establecidos, como resultado de los problemas o las situaciones que pueda presentar el deudor a lo largo de la vida del activo financiero o al vencimiento de este.

El objetivo de los modelos de riesgo de crédito es determinar la función de probabilidad de las pérdidas del crédito; es importante tener en cuenta que todas las empresas están expuestas a esto. Los tipos de riesgo son:

- Riesgo de impago o default: es el riesgo a las pérdidas por impago de la obligación.
- Riesgo de rebaja crediticia: es el riesgo de pérdida de valoración del crédito por parte de las entidades calificadoras.
- Riesgo de exposición: se refiere al riesgo sobre los pagos futuros de la obligación.
- Riesgo de spread de crédito: es el riesgo de que aumente la rentabilidad de un instrumento financiero respecto a otro, con la misma fecha de vencimiento.

El riesgo de crédito es medido por la pérdida esperada (PE) mediante el uso de la siguiente ecuación:

Ecuación 1. $PE = PD * EAD * LGD$

Donde:
PD: probabilidad de default. Es la probabilidad de incumplimiento de la obligación**.**
EAD: exposición a default. Es el monto del capital e intereses adeudado al momento del incumplimiento
LGD: es la pérdida por incumplimiento para el prestamista cuando la contraparte incumple la obligación.

3.2 Modelos de medición

3.2.1 Árboles de decisión

Surgen a partir de la teoría de juegos de John von Neumann y Oskar Morgenstern en 1944, quienes a partir de un tipo de gráficos representaron la estructura temporal de un juego en forma extensiva, siendo así un modelo que permite predecir el resultado de una variable dependiente por medio de combinaciones o particiones de ciertas variables independientes, llegando a un resultado a partir de una serie de decisiones o datos relacionados y teniendo como uno de sus principales objetivos el aprendizaje inductivo a partir de la visualización del árbol y de las construcciones lógicas. Gráficamente se representa como un conjunto de nodos de decisión, nodos de probabilidad y ramas (Breiman, Friedman, Stone y Olshen, 1984) que se detallan a continuación:

- Nodos de decisión: se representan con un cuadro e indican la necesidad de tomar una decisión.
- Nodos de probabilidad: se representan por medio de círculos que indican que en ese punto del proceso ocurre un evento aleatorio.

- Ramas: se representan mediante flechas que indican los diferentes caminos que surgen al hacer la elección de los eventos o las probabilidades.

Este modelo resulta útil cuando no se conoce o no se ha revisado el comportamiento de la población objeto de estudio, ya que permite, entre múltiples variables, definir las más significativas sin necesidad de contar con información previa sobre lo que se está modelando. Suele ser aplicado en campos como las ciencias médicas, la biología, la política y en diversas ramas de la economía.

En finanzas, este modelo suele ser usado en riesgo de crédito con el fin de determinar probabilidades de ocurrencia de default o no default introduciendo una variable categórica como nodo raíz. Normalmente los modelos econométricos que son usados para predecir posibilidades de default incluyen el ajuste de los datos hacia determinada distribución de probabilidad; dichos ajustes son estimados e incluidos en la ecuación de predicción, sin embargo, en árboles de decisión, el conjunto de datos suele ser dividido sucesivamente de acuerdo con el grado de relación existente entre las variables independientes y la variable dependiente y se indica el rango en donde la variable independiente tiene un mayor nivel de asociación con la variable dependiente y la fortaleza de sus correlaciones.

Una de las ventajas de estos modelos es que permiten categorizar los clientes evaluados en clientes cumplidos o incumplidos, o segmentarlos por categorías de riesgo como alto, medio o bajo. La categorización o segmentación de clientes de acuerdo con el resultado de la variable dependiente permite adicionalmente estimar probabilidades de pérdida futura dependiendo del perfil de riesgo asignado al cliente. Otras ventajas de este modelo son las de aprobar el uso de un gran número de variables independientes, capturar relaciones que no se encuentran fácilmente de los modelos lineales estándar y no requerir de supuestos distribucionales (Rayo, Lara y Camino, 2010).

Para trabajar con árboles de decisión existen múltiples metodologías, entre las que se destacan CART, CHAID, CHAID exhaustivo, QUEST y C4.5, las cuales permiten cambiar las reglas de asignación, las reglas de partición y los criterios con los que se corta el nodo final (Cardona, 2004). Una de las principales desventajas al momento de trabajar con este tipo de modelos es la imposibilidad de determinar la magnitud en la que cada variable aporta a la predicción de incumplimiento.

En el trabajo realizado por López (2007) acerca de los determinantes de riesgo de crédito para carteras con bajo nivel de incumplimiento en los bancos españoles, se adoptó árboles de decisión como una de las metodologías de estudio. En la primera fase se evaluaron todas las variables disponibles (indicadores de liquidez, endeudamiento, rotación, rentabilidad y productividad): 65 indicadores en total, y mediante diferentes procedimientos estadísticos se eligieron las variables más relevantes para el modelo. En dicho estudio se consideraron tres modelos diferentes para aplicar, los cuales son: análisis discriminante, análisis logit y árboles de decisión. El modelo árboles de decisión determinó que las variables más significativas para la predicción de incumplimiento eran los indicadores cash flow/resultado neto, activos líquidos/pasivo exigible y resultado neto/capital + reservas. Los resultados le permitieron a López concluir que las variables que mejor explican el incumplimiento en este tipo de carteras están dadas para los tres modelos en mayor medida por indicadores de rentabilidad y liquidez, y en menor medida por indicadores de endeudamiento. Así mismo, el modelo más relevante para predecir el incumplimiento durante el año posterior a la información evaluada es el modelo logit con un acierto del 90%, y el mejor modelo para predecir dos años antes de que ocurran los eventos es el modelo árboles de decisión con una efectividad del 85%.

Por otra parte, Cardona (2004) usa modelos basados en árboles de decisión para estimar las probabilidades de incumplimiento de créditos de libre inversión en una entidad bancaria colombiana; este modelo ofrece una ventaja que para el autor es fundamental, y es que el método es de fácil entendimiento para personas que no

cuentan con amplios conocimientos estadísticos, debido a que permite diferentes usos como la clasificación de clientes por rangos y modelos de cobranza de acuerdo con los perfiles de cada cliente. Cardona pretende evaluar el uso del modelo árboles de decisión desde tres aspectos distintos: simplicidad, potencia y estabilidad,[1] logrando resultados satisfactorios en dichos aspectos y permitiendo calcular de manera confiable las provisiones de cartera que impactan los resultados de la entidad.

3.2.2 Modelo logit o modelo de regresión logística

El término “curva logística” fue usado por primera vez por Edward Wright, pero es hasta el siglo XIX cuando es desarrollada la ecuación logística tal como hoy se conoce; fue aplicada en dos áreas de manera independiente: química y demografía, a partir de esto fue utilizada por diferentes investigadores, en diversas áreas, y es en 1944 con Joseph Berkson que se acuña el término “logit model” (Martínez, 2008).

El modelo logit es una técnica estadística que permite estimar la relación existente entre una variable dependiente no métrica o nominal y otras variables independientes que pueden ser métricas o no. Este modelo es usado para la toma de decisiones en una situación en donde hay solo dos posibles respuestas, por ejemplo, calificar a un cliente como “bueno” o “malo” a partir de las características cuantitativas y cualitativas que este pueda tener.

La ecuación usada por dicho modelo es la siguiente:

Ecuación 2. $P = \frac{1}{1+e^{-(Z)}}$

[1] Simplicidad: que el modelo sea entendido por cualquier persona de la entidad. Potencia: que discrimine correctamente a los clientes buenos y malos. Estabilidad: que el modelo sea consistente a lo largo del tiempo.

En donde:

P: es la probabilidad de incumplimiento.
z: es el scoring logístico.
e: es el número de Euler.

Esta metodología es empleada casi siempre en las ciencias económicas para la construcción de modelos de credit scoring, por sus propiedades que son más adecuadas a estadísticas respecto a otros modelos, su capacidad de admitir variables categóricas y la posibilidad de determinar la influencia de las variables independientes en el resultado de la variable dependiente.

En un estudio realizado por Rayo, Lara y Camino (2010), en instituciones de microfinanzas de Perú, los autores recomiendan que cada entidad financiera tenga su propio modelo de credit scoring de acuerdo con su historial de cartera, con el fin de medir la probabilidad de impago de los créditos que se van a otorgar; también proponen que dicho modelo permita clasificar a los clientes como solventes o insolventes usando la metodología de balanceo, con el fin de determinar el punto de corte óptimo para definir la insolvencia. En el estudio se distribuyen las variables cualitativas y cuantitativas del modelo a lo largo de seis fases,[2] buscando relacionarlas con el incumplimiento de los créditos. El resultado del modelo arroja la participación de ocho variables determinantes, que se encuentran en las seis fases mencionadas anteriormente, donde dos de estas influyen de manera positiva en la probabilidad de pago y seis de manera negativa. Una vez los autores logran determinar la ecuación del modelo, se procede con la validación de este, logrando porcentajes de acierto del 89% para determinar la probabilidad de pago y del 67% para determinar la probabilidad de impago.

[2] Investigación de mercado, informes de crédito para clientes nuevos o recurrentes, evaluación del expediente de crédito, evaluación de las garantías, aprobación de la solicitud y variables macroeconómicas.

En Salazar (2013) se aborda la medición del riesgo de crédito a la cartera de libre inversión en una de las instituciones financieras de la ciudad de Popayán. El autor usó dos metodologías:

- El modelo logit: que centra la medición econométrica en dos determinantes de la probabilidad o riesgo de impago: una, la información de los clientes, teniendo en cuenta variables como el género, el sector económico, el endeudamiento paralelo, el periodo de liquidez enfrentado al adquirir la deuda, el monto del préstamo, el número de periodos, la tasa de interés, los costos de la transacción y la valoración monetaria de los fallos en la información. Dada una de las variables se van incluyendo conforme se desarrolla el modelo y la relación que exista con otras; y la otra, de tipo macroeconómica, usando series de tiempo del municipio.

- El modelo log-log: en el cual se plantea una regresión donde su coeficiente periódico de cartera vencida en créditos de libre inversión depende del nivel de actividad económica local, de la tasa de interés y de una variable dummy que refleja el cambio estructural experimentado por la cartera vencida a razón de un problema en la regulación financiera, como fue la incursión en la economía local de captadores ilegales de dinero.

Salazar llega a la conclusión de que ambos modelos son útiles a la hora de determinar y evaluar el riesgo de crédito, y se demuestra con el modelo logit que la línea de crédito objeto de estudio es de bajo riesgo, para el caso 2,47% de probabilidad de incumplimiento severo, y con el modelo log-log se logra deducir que el índice de cartera vencida en la línea de crédito puede explicarse mediante factores macroeconómicos.

3.3 Pruebas para la selección del mejor modelo

Después de analizar la información en los modelos estadísticos previamente explicados, se selecciona el modelo más apropiado, haciendo uso de las siguientes pruebas:

3.3.1 Curva ROC

Acrónimo de Receiver Operating Characteristic (o Característica Operativa del Receptor). La curva ROC permite determinar la exactitud diagnóstica del modelo; es utilizada para establecer el punto de corte en el que se alcanza la sensibilidad y especificidad más alta, evaluar la capacidad discriminativa de una prueba diagnóstico (diferenciar por ejemplo entre incumplimiento y no cumplimiento) y comparar la capacidad discriminativa de dos o más pruebas diagnósticas que expresan sus resultados como escalas continuas.

Se representa gráficamente mediante una figura llamada “figura de la curva ROC”, donde cada punto de la curva corresponde a un punto de corte de la prueba diagnóstica, informando respecto a la sensibilidad (eje Y) y 1-especificidad (eje X) del modelo. Ambos ejes de la figura incluyen valores entre 0 y 1 (0% a 100%). La línea trazada desde el punto 0,0 al punto 1,1 recibe el nombre de diagonal de referencia, o línea de no-discriminación. Cuanto más se aproxime la curva ROC a la diagonal de referencia menor poder discriminativo tendrá la prueba diagnóstica; esto significa menor capacidad de determinar entre los casos de cumplimiento o incumplimiento en los pagos. Por el contrario, cuanto más se acerque a 1 mayor será este poder discriminativo (“Análisis ROC: visualización”, s. f.).

3.3.2 Prueba de Kolmogorov-Smirnov o K-S

Esta es una prueba no paramétrica de bondad de ajuste que permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica específica. Su objetivo es señalar si los datos provienen de una población que tiene la distribución teórica especificada.
Las ventajas del uso de esta prueba son:

- Es más eficaz que la prueba chi-cuadrado (χ^2).
- Es fácil de calcular y usar, y no requiere agrupación de los datos.
- El estadístico es independiente de la distribución de frecuencias esperada, solo depende del tamaño de la muestra (García, González y Jornet, 2010).

Para determinar el resultado de esta prueba se debe contrastar la hipótesis nula:

$$\begin{Bmatrix} H_0: La\ variable\ si\ cumple\ con\ la\ distribución\ teórica\ (p > 0.05) \\ H_1: La\ variable\ no\ cumple\ con\ la\ distribución\ teórica\ (p < 0.05) \end{Bmatrix}$$

3.3.3 Test de Hosmer-Lemeshow

Se trata de una prueba para evaluar la bondad de ajuste, es decir, si el modelo propuesto puede explicar lo observado al medir la distancia entre lo observado y lo esperado. Se realiza ordenando de menor a mayor las N probabilidades estimadas y agrupándolas en diez grupos o intervalos. Se cuenta para cada intervalo el valor esperado (el valor calculado a partir del modelo) y el observado (los valores que se tienen) para cada uno de los dos resultados posibles de la variable dependiente dicotómica. El estadístico de esta prueba se obtiene calculando el ji-cuadrado de Pearson a partir de las frecuencias observadas y estimadas para cada uno de los intervalos (Sánchez, 2012).

3.3.4 Tabla de contingencia

Es una de las formas más comunes para resumir datos categóricos. Es usada para medir la influencia de una variable independiente sobre una independiente y calcular la intensidad de dicha asociación.

4. METODOLOGÍA PARA EL DESARROLLO DEL MODELO

Para obtener el resultado esperado, en el capítulo 3 de este documento se realizó una investigación bibliográfica acerca de los modelos de credit scoring más usados y que mejor se adaptan a la situación de estudio, entendiendo sus ventajas y desventajas al momento de aplicación.

La población de la investigación está constituida por la base de datos de las operaciones de factoring de los últimos tres años de la compañía (2017, 2018, 2019), la cual se encuentra registrada en el software usado por GFO, esta se conforma por 47.863 operaciones realizadas por 2.653 clientes emisores, con un total de 396.211 facturas negociadas.

La base de datos fue organizada y depurada como corresponde para proceder con el análisis de esta y así entender el comportamiento de los clientes a lo largo de los años; así mismo, se realizó un análisis de cartera por edades para identificar la magnitud de la compañía y su riesgo de crédito.

Debido a que la empresa no cuenta con una base de datos organizada y generalizada de la información financiera de los clientes emisores, el paso siguiente fue iniciar una búsqueda en fuentes externas con el fin de conseguir la mayor cantidad de información de los clientes correspondiente a los años donde se realizaron las operaciones de factoring con GFO, para lograr un análisis minucioso de los resultados financieros mediante una serie de indicadores, los cuales se cruzaron con la base de datos de la compañía.

Así pues, se delimitó de manera considerable la muestra sobre la cual se realizó el trabajo, dados los siguientes criterios de inclusión y exclusión:

Criterios de inclusión:

- Clientes emisores que hayan realizado operaciones de factoring con la compañía entre los años 2017, 2018 y 2019.

- Clientes emisores que hayan realizado al menos 10 operaciones de factoring con la compañía.
- Clientes que hayan estado activos con la compañía durante los años 2017, 2018 y 2019.

Criterios de exclusión:

- Clientes emisores que hayan adoptado NIIF para pymes o plenas, que no se encuentran sometidas a inspección o vigilancia por la Superintendencia de sociedades pertenecientes al sector real de la economía, y por tanto que no reporten sus estados financieros a través del SIREM. (única fuente de información financiera disponible al alcance).
- Clientes emisores con operaciones que no presentasen fecha coherente de desembolso, estimada de pago y real de pago, en la base de datos suministrada por la empresa.
- Clientes emisores en cuyas operaciones presentasen datos faltantes o duplicados.
- Clientes emisores cuya información cualitativa presente datos faltantes o duplicados.

De acuerdo con los criterios anteriores, la muestra para el desarrollo de este trabajo quedó conformada por 419 clientes emisores, que hayan realizado más de 10 operaciones de factoring y que hayan estado activos durante los 3 años objeto de estudio.

Una vez obtenida y depurada toda la información requerida, se analizaron las diferentes metodologías que de acuerdo con el tamaño de la muestra y la calidad de la información podían ser usadas. Luego de seleccionar los mejores modelos se organizó la información de acuerdo con el software estadístico SPSS, versión 25.0; también se realizó el análisis de los resultados y finalmente se emitieron las conclusiones y recomendaciones.

5. RESULTADOS

5.1 Desarrollo del modelo

Para el desarrollo del modelo se tomó una muestra obtenida a partir de la base de datos suministrada por GFO, la cual cumplía a cabalidad con los criterios de inclusión y exclusión determinados en la metodología. La muestra incluye clientes de todos los sectores económicos ubicados en más de 20 ciudades diferentes de Colombia e incluye operaciones realizadas con condiciones de negociación que oscilan entre el 80% y el 100% del valor nominal del instrumento financiero, con plazos de pago pactados entre 10 y 180 días y tasas de interés entre el 14% y el 30% NA.

Como punto de partida se calculó la cartera por edades de vencimiento a corte 31 de diciembre de cada año, tomando como periodo de gracia los primeros 30 días después del vencimiento de la operación, calificando esta como corriente y clasificando el incumplimiento por rangos de edades entre 31 a 60 días, 61 a 90 días, 91 a 120 días, 121 a 150 días y más de 150 días de vencimiento; de acuerdo con éstos la cartera en cifras de la compañía presentó los resultados que se muestran en los Anexos 1 y 2.

Una vez verificada la edad de la cartera se definió como variable dependiente del modelo o variable explicativa el incumplimiento en la fecha estimada del pago de la obligación más 30 días de gracia, entendiendo que en muchas ocasiones los clientes se toman un tiempo prudencial adicional para el pago de sus obligaciones. Si bien el factoring, no se considera microcrédito, este se tomó como referencia para definir los días de gracia adicional que marcan y/o separan una operación como cumplida o incumplida[3].

[3] De acuerdo con la circular externa No. 100 de 1995, capitulo II, sección 1 de la Superintendencia Financiera, la cual determina los componentes de la perdida esperada en los microcréditos que se encuentren con una mora mayor o igual a 30 días, tiempo a partir del cual se considera incumplimiento y amerita seguimiento.

Una vez definido el punto a partir del cual se considera incumplida una operación, la variable dependiente se categorizó como una variable binaria así 0 = Sí presentó incumplimiento en la fecha estimada de pago, con una mora superior a 30 días, opción que se denominará de ahora en adelante como sí o incumplimiento y 1 = No presentó incumplimiento en la fecha estimada de pago, con una mora superior a 30 días, opción que se denominará de ahora en adelante como no o cumplimiento.

Después de definir la variable dependiente, se evalúa la muestra para determinar dentro del universo de clientes de esta, que porcentaje de ellos llegaron a presentar incumplimiento en alguna de sus operaciones, arrojando como resultado que el 33% de los clientes, han llegado a presentar incumplimiento en al menos 1 factura de alguna de las operaciones realizadas y el 67% restante de los clientes nunca han presentado incumplimiento, tal como se muestra en la figura 1.

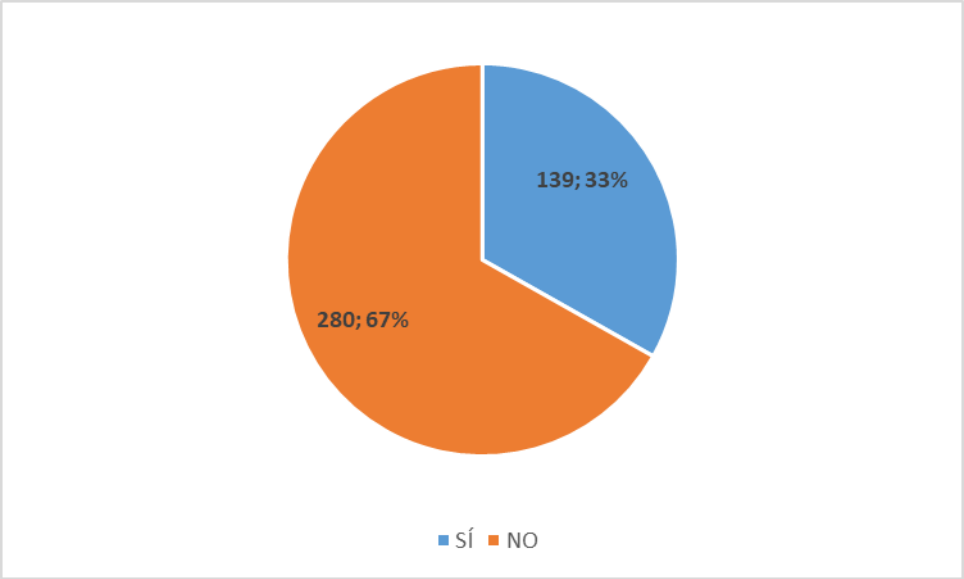


Figura 1. Clientes con incumplimiento en alguna de sus operaciones

Fuente: elaboración propia con datos internos de GFO.

Es necesario aclarar, que si bien en la figura 1, se evidencia que el 33% de los clientes han presentado incumplimiento en alguna de sus operaciones, esto no es igual a decir que el 33% de la cartera de la empresa se encuentre en mora, tal como

se evidencia en el anexo 2, donde en promedio la cartera en mora para los años objeto de estudio es del 6% aproximadamente.

Posteriormente, para establecer las variables independientes, las cuales fueron clasificadas como cualitativas y cuantitativas se procedió a obtener la mayor información posible de las operaciones como de los clientes, teniendo como fuentes las bases de datos internas de GFO y las bases de datos externas como los estados financieros reportados por las compañías a la Superintendencia de Sociedades.

5.1.1. Variables cualitativas

Se concretó que las variables cualitativas que se iban a analizar en el modelo serían principalmente las reportadas anualmente por las compañías a la Superintendencia de Sociedades, clasificando estas en 3 grandes grupos: confiabilidad, experiencia y ubicación geográfica y sector.

- Confiabilidad: Es el grupo de las variables que dan una idea de la confiabilidad que puede tener la empresa cliente, en esta se clasifican las variables: Estado actual de la organización (activa, acuerdo de reorganización o acuerdo de reestructuración), la información financiera presenta información reexpresada (Si / No), ¿Se reunió el máximo órgano social para considerar la información financiera? (Si/no), ¿La compañía está obligada a tener revisor fiscal? (Si/No), ¿El revisor fiscal pertenece a una firma? (Si/No), ¿Los estados financieros están acompañados del dictamen del revisor fiscal? (Si/No) y ¿Cuál es el concepto del revisor fiscal sobre los estados financieros? (Limpio, con salvedad, negativo).

- Experiencia: en este grupo, se encuentra únicamente la variable “Años de constitución de la empresa cliente”, determinada a partir de la fecha de inscripción ante Cámara de Comercio. De la muestra resultante, los clientes presentan un mínimo de 2 años de constituidas y un máximo de 76 años, siendo la media de estas, 26 años.

- Ubicación geográfica y sector: por medio de este grupo de variables se pretende determinar si la ubicación geográfica y sector económico al cual pertenecen las empresas cliente, tienen algún tipo de influencia en su comportamiento de pago.

5.1.2 Variables cuantitativas

En cuanto a las variables cuantitativas, se logró obtener información tanto de los estados financieros de las compañías, así como de las condiciones de cada operación realizada como las tasas de interés negociadas con cada cliente, la fecha de desembolso, fecha estimada y fecha real de pago y el monto de la operación.

Con los estados financieros, que comprendían estado de situación financiera, estado de resultados integral y estado de flujos de efectivo de los años en que se realizaron las operaciones, se procedió a elaborar una matriz de indicadores financieros la cual se puede observar en el Anexo 3, con la finalidad de evaluar la influencia de estos en la variable dependiente (incumplimiento); adicionalmente se cuenta con las variables ingresos, cuentas por cobrar y capital.

5.2 Análisis de datos en software

Para el desarrollo de los modelos se usó el programa estadístico SPSS, versión 25.0, el cual ofrece un análisis estadístico avanzado, una amplia biblioteca de algoritmos de aprendizaje automático, el análisis de texto, la extensibilidad de código abierto, la integración con Big Data y una implementación en aplicaciones, siendo de fácil comprensión y uso.

Para iniciar el análisis de las variables fue necesario depurar la información: a partir de la base de datos, conformada finalmente por 47 variables, se categorizaron algunas de estas, teniendo en cuenta que las opciones fueran excluyentes entre sí

y exhaustivas. A unas de las variables se les asignó valores de 1 = Sí, 0 = No, y otras fueron categorizadas de acuerdo con las opciones que ofrecían.

Una vez organizada la información, teniendo en cuenta los requerimientos del software, se procedió a analizarla mediante los siguientes modelos:

5.2.1 Modelo de regresión logística binaria

El primer paso para desarrollar este modelo fue determinar las variables, siendo la variable dependiente el incumplimiento en la obligación de pago. Como variables independientes se establecieron todas las que conformaban la base de datos, excepto el valor de las utilidades decretadas, actividad económica según el código CIIU, nombre de la firma a la cual pertenece el revisor fiscal, ingresos, capital y cuentas por cobrar, ya que sus resultados eran dispersos, variables entre sí e incompletos en algunos casos, lo que hacía muy complejo categorizarlas; adicionalmente, las variables facturas incumplidas, total facturas negociadas y tasa de la operación son propias del comportamiento del cliente, por lo que no aplicarían para clientes nuevos.

El segundo paso fue dividir la muestra de 419 datos en dos submuestras de manera aleatoria, tal y como se muestra en la tabla 1, con el fin de realizar *a posteriori* la validación del modelo de regresión logística estimado. Para este propósito se destinó el 75% de la muestra, es decir, 314 datos para la estimación del modelo estadístico y el 25% para la muestra de entrenamiento o backtesting, lo que significa los restantes 105 datos. La totalidad de los datos que hacen parte de la muestra corresponden a operaciones de factoring finalizadas, cada una de estas realizadas con personas jurídicas, las cuales surtieron los procesos de aprobación habituales de acuerdo con los procedimientos y las políticas internas de crédito de la compañía GFO.

Tabla 1. Observaciones muestra de comprobación y entrenamiento-modelo logit

Grupo Factoring de Occidente S. A. S. Periodo estudio 2017/2019		
Observaciones (N)		
Pagos	Impagos	Total
280	139	419
Muestra de entrenamiento (75%)		
205	109	314
Muestra de validación (25%)		
75	30	105

Fuente: datos arrojados por el software SPSS.

Usando los valores de sensibilidad y especificidad, que pueden verse en el Anexo 4, se determinó que el punto de corte óptimo para el modelo está situado en 0,3350, siendo este donde se encuentra la mayor probabilidad de acierto, la cual es de 0,766.

Teniendo en cuenta lo anterior, se aplica en el modelo el punto de corte del 33%, arrojando así la siguiente ecuación:

Ecuación 3.

$$Z = 8{,}205 + 0{,}00\ Capital\ de\ Trabajo - 0{,}295\ Margen\ Operacional + 0{,}141\ Margen\ Ebitda + 0{,}207\ Margen\ Neto - 0{,}085\ Endeudamiento + 20{,}074\ Impacto\ Carga\ Financiera - 11{,}459\ Solvencia$$

Donde:

Ecuación 4.

$$P(Probabilidad\ de\ incumplimiento) = \frac{1}{1+e^{-(Z)}}$$

Las variables que intervienen en el modelo final se muestran a continuación:

Tabla 2. Variables en la ecuación-modelo logit

Variables en la ecuación

	B	gl	Sig.
Capital de trabajo	0,000	1	0,074
Margen operacional	-0,295	1	0,000
Margen ebitda	0,141	1	0,021
Margen neto	0,207	1	0,000
Endeudamiento	-0,085	1	0,003
Impacto carga financiera	20,074	1	0,000
Solvencia	-11,459	1	0,000
Constante	8,205	1	0,004

Fuente: datos arrojados por el software SPSS.

Una vez arrojada la ecuación, se procede a analizar los coeficientes de las variables, donde se evidencia que los indicadores financieros más relevantes al momento de definir la probabilidad de incumplimiento son el impacto en la carga financiera y la solvencia, siendo estos los coeficientes más altos, así mismo, se observa que tanto el margen ebitda, el margen neto y el nivel de endeudamiento, cuyos coeficientes son bajos, muestran inconsistencias en sus signos como por ejemplo: a mayor endeudamiento menor probabilidad de incumplimiento, y a mayor margen neto y margen ebitda mayor probabilidad de incumplimiento, indicadores que dan evidencia de limitaciones en la ecuación arrojada por el modelo.

Esta situación puede tener explicación tanto en el tamaño de la muestra ya que los datos son limitados, o en el tipo de clientes, ya que en general este servicio es prestado a compañías con poco acceso a la banca tradicional por sus regulares condiciones financieras. Por ejemplo, el coeficiente de endeudamiento puede determinar que el cliente, mientras mayor sea su nivel de endeudamiento, más cumplido es en sus obligaciones de pago con GFO, puesto que esta puede ser su única fuente de liquidez inmediata.

Valoración del modelo: para determinar si el modelo mediante esta metodología es adecuado, se procede a realizar las pruebas establecidas previamente en el marco teórico.

De acuerdo con la tabla 2, de los coeficientes de las ocho variables que integran el modelo, tres de ellas influyen de manera negativa y el resto de manera positiva en la probabilidad de que un cliente presente o no incumplimiento en el pago. Dado que se fija p-value en 0,05 y al contrastarse el nivel de significancia de las variables independientes se concluye que estas influyen en el comportamiento de pago. Pese a que la variable capital de trabajo que tienen $P\,Value > 0{,}05$ no es significativa, no fue retirada del modelo, ya que al eliminarla resultaría afectado el scoring de incumplimiento.

En el Anexo 6 se evalúa la importancia global del modelo por medio de la prueba ómnibus de coeficientes, donde se observa que el estadístico chi-cuadrado, arrojado para el modelo es 87,24, el cual al contrastar su nivel de significancia $P = 0{,}00 < 0{,}005$ hace que la hipótesis nula sea aceptada, por tanto, se determina que el modelo puede predecir la variable dependiente.

Para medir la bondad de ajuste del modelo, el software calcula los coeficientes R^2 de Cox y Snell, con un valor de 0.268 y una versión corregida de este, y el R^2 de Nagelkerke, con un valor de 0.371 como se muestra en el Anexo 7, los cuales son indicativos de un ajuste que se considera aceptable para el modelo.

Otra prueba para medir la bondad de ajuste del modelo es la prueba de Hosmer y Lemeshow, dado su nivel de significancia $P = 0{,}802 > 0{,}05$ como se muestra en el Anexo 8; no se rechaza la hipótesis nula, lo que indica que el modelo se ajusta lo suficiente a los datos.

Lo anterior es confirmado en el Anexo 9, en donde se detalla la pequeña diferencia entre los valores observados y esperados para cada una de las frecuencias calculadas.

Para evaluar la capacidad predictiva del modelo se analizan los datos de la tabla 3, en ella se encuentra el resumen de los datos observados y los datos pronosticados por el modelo, distinguiendo entre los clientes que han presentado incumplimiento en el pago y los que no.

Tabla 3. Clasificación-modelo logit

Tabla de clasificación

Observado		Pronosticado Casos seleccionados Incumplimiento		Porcentaje correcto
		0	1	
Incumplimiento	0	152	33	82,2
	1	28	67	70,5
Porcentaje global				78,2

a. El valor de corte es 0,330

Fuente: datos arrojados por el software SPSS.

El nivel de sensibilidad está dado por la capacidad del modelo para estimar los casos más relevantes, a saber, los casos de incumplimiento; en este modelo, como se observa en la tabla 3, se estimaron de manera correcta 67 casos de incumplimiento versus un observado de 95, lo que significa un acierto o nivel de sensibilidad del 70,5%.

El nivel de especificidad está dado por el porcentaje de estimación de los datos de cumplimiento; en este caso, como se observa en la tabla 3, el modelo estimó de manera correcta 152 casos de cumplimiento versus un observado de 185; pronosticó de manera correcta el 82,2% de los casos.

Finalmente, el nivel de exactitud está dado por la capacidad global del modelo de predecir tanto datos de cumplimiento como de incumplimiento, en este caso, el modelo presenta un porcentaje global de exactitud del 78,2%; así mismo, se evidencia que el modelo es más específico que sensible pues tiene mayor capacidad para predecir con certeza las empresas que pueden llegar a presentar cumplimiento en los pagos.

En el Anexo 10 se muestra el área bajo la curva; para el modelo su resultado es de 0.826, siendo este un resultado positivo, ya que entre más se acerque a 1 mayor es su poder de discriminación entre los casos de incumplimiento versus los de cumplimiento.

Al observarse la figura 2 se evidencia que la curva del modelo está cercana a la esquina superior izquierda, confirmando que este tiene una buena capacidad de discriminación entre ambas opciones.

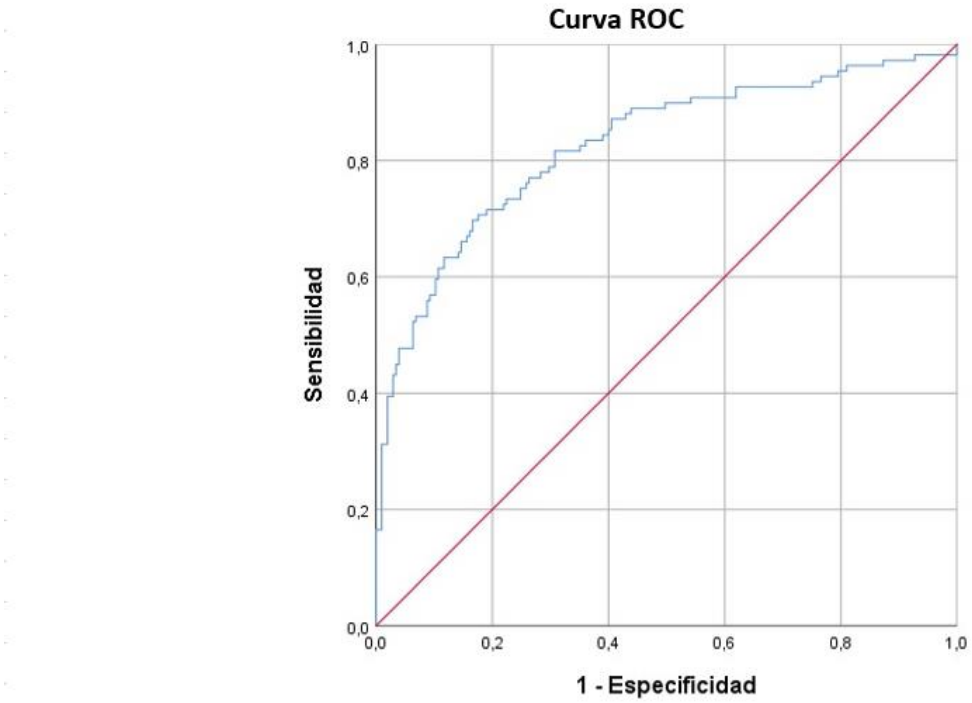


Figura 2. Curva ROC-modelo logit

Fuente: datos arrojados por el software SPSS.

Posteriormente se evalúa si la distribución poblacional del modelo se ajusta a la distribución de los datos por medio del estadístico p y mediante la prueba de Kolmogorov-Smirnov, el cual al ser igual a 0,00, tal como se muestra en el Anexo 11, indica que la distribución de la frecuencia del modelo no es consistente con la distribución teórica, presentando un grado de dispersión de 0,163.

Validación del modelo:

Tabla 4. Clasificación backtesting-modelo logit

Tabla de clasificación

Observado		Pronosticado					
		Casos seleccionados			Casos no seleccionados		
		Incumplimiento		%	Incumplimiento		%
		0	1	correcto	0	1	correcto
Incumplimiento	0	152	33	82,2	51	10	83,6
	1	28	67	70,5	7	17	70,8
Porcentaje global				78,2			80,0

a. El valor de corte es ,330

Fuente: datos arrojados por el software SPSS.

Para realizar el proceso de validación del modelo logit se realizó el proceso de backtesting por medio del software SPSS; el software corrió el modelo inicial sobre 314 datos, de los cuales descartó 34 como se muestra en el Anexo 13. Para el proceso de backtesting quedaron disponibles 105 datos, de los cuales el software descartó 20 por presentar valores perdidos, dejando 85 datos para la validación final como se muestra en la tabla 4, donde se logró obtener un nivel de aciertos totales del 80%, un nivel de sensibilidad del 70,8% y especificidad del 83,6%, confirmando la viabilidad del modelo.

5.2.2 Modelo árboles de decisión

Para este modelo también se usó el software SPSS. A partir de la base de datos ya depurada se corrió el modelo con la metodología CHAID (chi-square automatic interaction detection), la cual elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente; para este caso, al tener una muestra pequeña de datos se eligió una configuración de 50/20 para los nodos padres e hijos respectivamente, lo que señala que cada nodo padre deberá tener como mínimo 50 casos y cada nodo hijo 20 (el software viene predeterminado con la configuración 100/50).

Otro punto que es importante aclarar en la configuración del modelo es la definición de los niveles de profundidad que se requieren y que la muestra esté en capacidad de dar, para este caso, adicional a la configuración 50/20 se estableció un máximo de cinco niveles de profundidad para el árbol, sin embargo, por el tamaño de la muestra y la configuración 50/20 elegida, el máximo de niveles obtenidos fue de tres. De haber tenido una muestra con un tamaño poblacional mayor, el árbol se cortaría o en el quinto nivel de profundidad o en el punto donde los nodos padres e hijos no cumplieran la condición 50/20, como en este caso. Esta combinación permite obtener resultados más útiles para el trabajo desarrollado, ya que, de correrse el modelo con la configuración predeterminada, el árbol arrojado tendría únicamente dos niveles de profundidad.

Posterior a definir la configuración de los nodos padre e hijo y de la profundidad del árbol, el modelo separa el conjunto de datos en dos partes, una para construir la muestra de entrenamiento del modelo y otra para construir la muestra de comprobación.

El modelo arroja un árbol con las siguientes variables independientes: impacto en la carga financiera, sector económico, días de rotación de proveedores y cobertura de intereses, conformado por diez nodos y tres niveles de profundidad.

La figura 3 representa el árbol de entrenamiento.

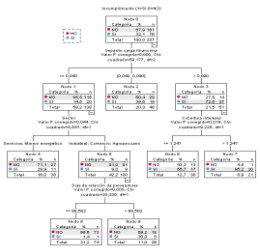


Figura 3. Muestra de entrenamiento-modelo árboles de decisión
Fuente: datos arrojados por el software SPSS.

En el nodo 0 se encuentra la variable dependiente, es decir, incumplimiento en el pago de las obligaciones financieras contraídas por los clientes con la empresa GFO; para este caso, la muestra de entrenamiento fue de 237 clientes, de los cuales el 32,1% presentó incumplimiento y el 67,9% no. La variable dependiente se ramifica en tres nodos de acuerdo con la variable “impacto en la carga financiera”, lo que indica que esta es la variable predictora principal. Las empresas con dicho

indicador menor o igual a 0.04 fueron 138, el 58,2% del total de la muestra de entrenamiento, de las cuales el 14,5% presentaron incumplimiento; las empresas con este indicador entre 0.04 y 0.09 fueron 48, lo que significa el 20,3% de la muestra de entrenamiento, de las cuales el 36,9% presentaron incumplimiento, y por último, en el nodo 3 se clasificaron las empresas con un indicador mayor a 0.09, en el cual se situó el 21,5% de la población, es decir, 51 empresas, de las cuales el 72,5% presentaron incumplimiento.

El nodo 1 se ramifica en los nodos 4 y 5 correspondientes al sector económico al cual pertenece la empresa, presentándose mayor incumplimiento en las empresas de los sectores servicios y minero energético con un 28,9%, versus un 9% de incumplimiento en las empresas de los sectores industrial, comercial y agropecuario.

A su vez, el nodo 5 se ramifica en los nodos 8 y 9 de acuerdo con los días de rotación de proveedores, presentándose mayor incumplimiento en las empresas con este indicador mayor a 88,5 días.

El nodo 3, donde se encuentran las empresas con el indicador impacto en la carga financiera mayor a 0,09, se ramifica en los nodos 6 y 7 de acuerdo con el indicador cobertura de intereses; en este, las empresas con una cobertura de intereses menor o igual a 1,24 presentaron un nivel de incumplimiento del 56,7%, y las que tienen este indicador superior a 1,24 presentaron incumplimiento en el 95,2% de los casos.

A partir de esto se puede concluir que la última rama del modelo presenta inconsistencias, ya que un mayor nivel de cobertura de intereses se asocia con un desempeño positivo de la empresa el cual no debería estar relacionado con un mal comportamiento de pago, este tipo de inconsistencia en el modelo puede estar dado por el tamaño de la muestra, debido a que por ejemplo en esta rama solo se analizan 51 de los 237 clientes.

Valoración del modelo: para determinar si el modelo mediante esta metodología es adecuado, se procede a realizar las pruebas establecidas previamente en el marco teórico.

Se inicia evaluando la capacidad predictiva del modelo tal como se muestra en la tabla 5; el porcentaje global de predicción del modelo es 77,6%, teniendo una alta especificidad y una baja sensibilidad, entonces el modelo tiene una buena capacidad para determinar los casos de cumplimiento (91,3%), pero una baja capacidad para detectar los casos de incumplimiento (48,7%).

Tabla 5. Clasificación-modelo árboles de decisión

Clasificación

Muestra		Pronosticado		
		NO	SÍ	Porcentaje correcto
Entrenamiento	NO	147	14	91,3%
	SÍ	39	37	48,7%
	Porcentaje global	78,5%	21,5%	77,6%

Método de crecimiento: CHAID

Fuente: datos arrojados por el software SPSS.

Una vez determinado el nivel de sensibilidad y especificidad del modelo de entrenamiento, se calcula el punto de corte óptimo para este mediante las coordenadas de la curva ROC, tal como se muestra en el Anexo 15. En ella se determina que el punto de corte óptimo de la curva está en 0.46; este es el punto donde el modelo logra discriminar de una mejor manera los casos de incumplimiento de los casos de cumplimiento.

Paso siguiente, se evalúa el área bajo la curva, tal como lo muestra el Anexo 16, la cual arroja un resultado de 0.841, siendo este un resultado positivo, ya que entre más se acerque a 1 mayor es el poder de discriminación entre los casos de incumplimiento versus los de cumplimiento.

El análisis anterior se puede visualizar en la figura 4, donde se evidencia que la curva del modelo está cercana a la esquina superior izquierda, confirmando que este tiene una buena capacidad de discriminación entre ambas opciones.

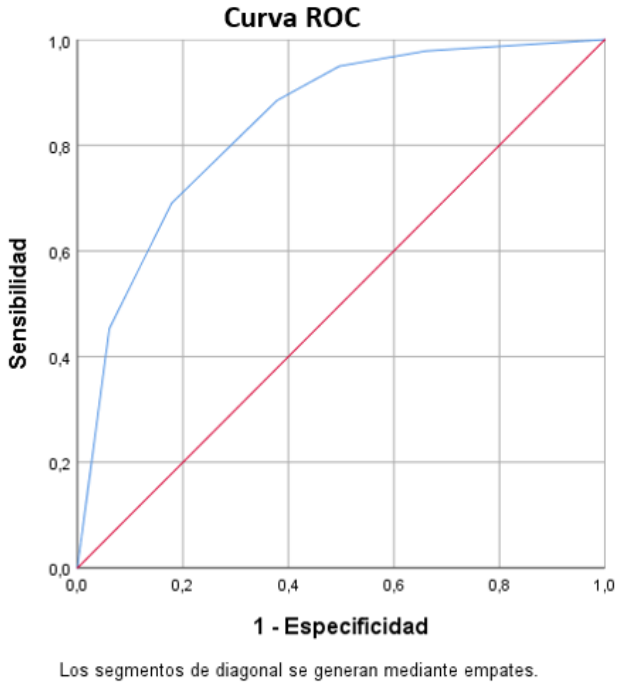


Figura 4. Curva ROC-modelo árboles de decisión

Fuente: datos arrojados por el software SPSS.

Posteriormente, se evalúa el ajuste de la distribución poblacional del modelo y la distribución de los datos, haciendo uso de la prueba Kolmogorov-Smirnov. De acuerdo con el Anexo 17 esta arroja un estadístico $p = 0.00$, indicando que la distribución de frecuencia del modelo no es consistente con la distribución teórica, presentando un grado de dispersión de 0.209.

Por último, con los datos relacionados en el Anexo 18 se analiza el comportamiento de pago de los clientes de acuerdo con el sector al cual pertenecen. Del total de la muestra, la mayor concentración de los clientes está en el sector industrial con un 33,2% del total, seguido del comercio con un 32,2% y servicios con un 28,6%; por

otra parte, la menor concentración la presentaron el sector agropecuario y minero energético con un 1,9% y 4,1% respectivamente.

Del total de la muestra, el porcentaje de clientes que presentaron incumplimiento también está concentrado en clientes del sector servicios con un 37,41% del total, seguido por el sector industrial con un 35,97%, comercio 20,86% y minero energético 5,04%.

Finalmente, el sector que presenta mejor comportamiento de pago es el agropecuario, sin embargo, como la frecuencia esperada del incumplimiento en este sector es inferior a 5, este dato no es muy confiable.

Validación del modelo: para realizar la validación del modelo el software dejó por fuera de la muestra de entrenamiento 182 datos, equivalentes al 43,44% de la muestra total; con estos se construyó el árbol para la muestra de comprobación tal y como se observa en la figura 5.

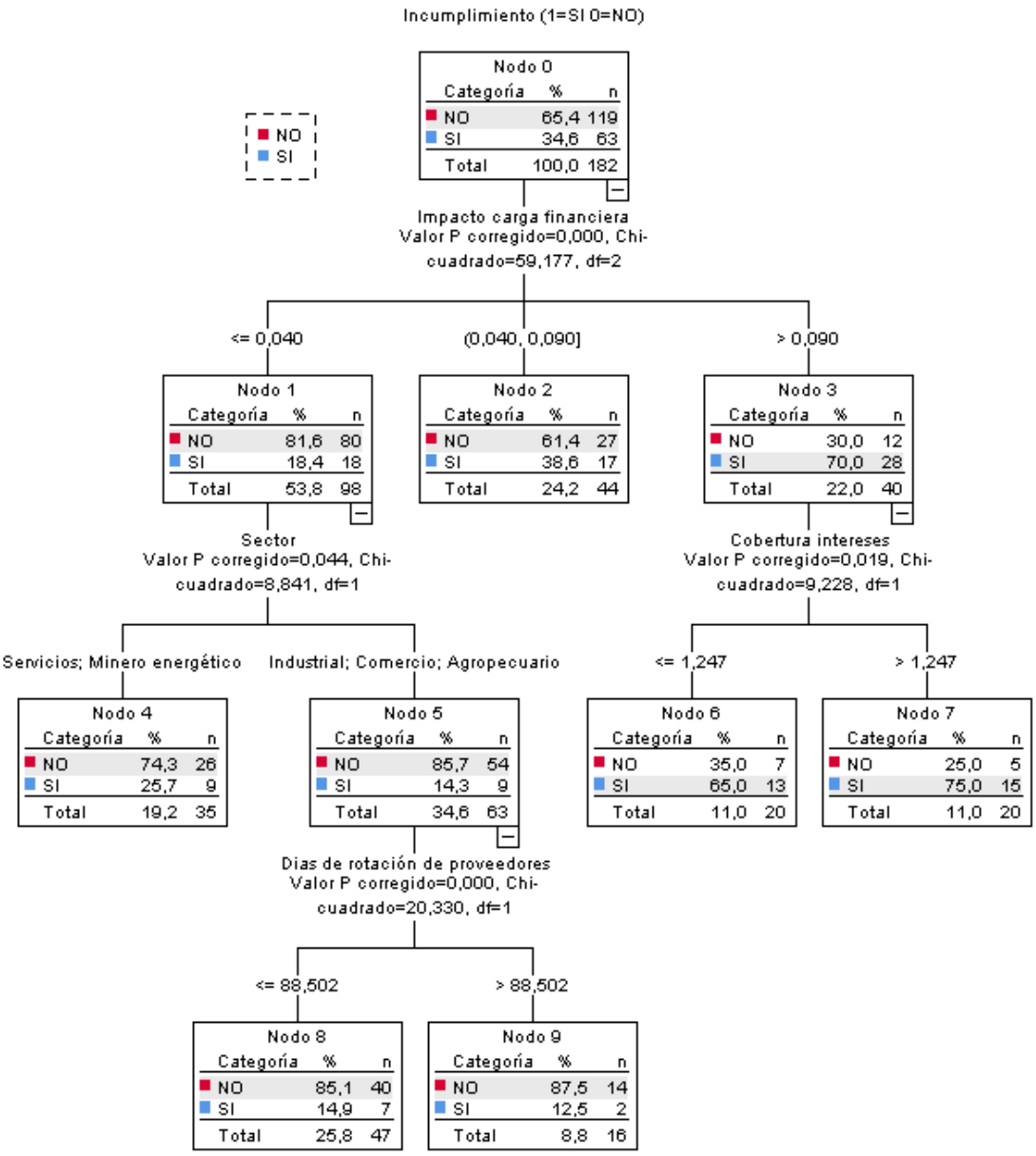


Figura 5. Muestra de comprobación-modelo árboles de decisión

Fuente: datos arrojados por el software SPSS.

Por medio de la tabla de clasificación del modelo de comprobación (tabla 6) se evalúa la capacidad predictiva del modelo, observando un porcentaje global de aciertos del 74,2%, siendo similar al del modelo de entrenamiento (77,6%).

Tabla 6. Clasificación muestra de comprobación-modelo árboles de decisión

Clasificación

Muestra		Pronosticado		
		NO	SÍ	Porcentaje correcto
Prueba	NO	107	12	89,9%
	SÍ	35	28	44,4%
	Porcentaje global	78,0%	22,0%	74,2%

Método de crecimiento: CHAID

Fuente: datos arrojados por el software SPSS.

5.3 Comparación entre los modelos

Para finalizar se sintetiza la validación de los modelos desarrollados por medio de las dos metodologías, tal como se muestra en la tabla 7, comparando entre ellos el nivel de especificidad, sensibilidad y porcentaje global, tanto para los modelos de entrenamiento como para los backtesting realizados. Adicionalmente se comparan los resultados de las pruebas de bondad de ajuste, como lo son la curva ROC y la prueba Kolmogorov-Smirnov.

Tabla 7. Comparación entre el modelo logit y el modelo árboles de decisión

	Modelo logit		Modelo árboles de decisión	
	Entrenamiento	Prueba	Entrenamiento	Prueba
Especificidad	82,2	83,6	91,3	89,9
Sensibilidad	70,5	70,8	48,7	44,4
Porcentaje global	78,2	80,0	77,6	74,2
Curva ROC	0,826		0,841	
Kolmogorov-Smirnov	0,163		0,209	

Fuente: datos arrojados por el software SPSS.

6. CONCLUSIONES Y RECOMENDACIONES

En el desarrollo de este trabajo se diseñó un modelo de credit scoring para la cartera de la empresa Grupo Factoring de Occidente, compañía dedicada al factoring en Colombia y vigilada por la Superintendencia de Sociedades. Se identificó la poca literatura existente con relación a la medición del riesgo de crédito en este tipo de empresas y la falta de regulación al respecto para compañías no bancarias que se dedican a esta actividad, dejando un margen bastante amplio para el desarrollo de futuros trabajos en esta línea de investigación.

La base de datos para el objeto de estudio comprendía 396.212 operaciones realizadas entre los años 2017 a 2019, correspondientes a 2.530 clientes, sin embargo, esta no incluía información relevante del cliente en el momento de realizar la venta de sus facturas.

Dada la escasez de datos financieros en los historiales de las operaciones, fue necesario recurrir a fuentes externas para acceder a la mayor cantidad de información posible, siendo esta la principal limitación encontrada a nivel de la muestra, ya que solo se contaba con la información reportada en la Superintendencia Financiera, adicionalmente, el servicio de factoring con entidades no financieras, como es el caso de la compañía objeto de estudio, son tomados por personas naturales o jurídicas que no cuentan con acceso a la banca, debido a un mal historial crediticio, ya sea por el comportamiento de pago, reportes en centrales de riesgo, por ser compañías con estructuras financieras muy pequeñas o por manejar informalidad en sus sistemas contables lo que hace aún más complejo poder contar con información financiera veraz y actual, pese a esto finalmente se logró la consecución de información financiera de 419 clientes que realizaron la venta de 123.622 facturas al Grupo Factoring de Occidente, reduciendo de manera considerable el tamaño muestral para la elaboración y evaluación del modelo.

Para el análisis de la información se llevó a cabo el desarrollo de dos modelos, uno mediante árboles de decisión y otro por medio de la regresión logística binaria; ambos presentaban una especificidad superior al 74% e incluían entre las variables explicativas el indicador impacto en la carga financiera, días de rotación de proveedores y el sector económico al cual pertenecen los clientes; la capacidad de predicción del incumplimiento para ambos era superior al 74%.

Sin embargo, la técnica de regresión logística binaria es la que diseña el modelo con mayor capacidad de predicción global, con un 80,8%; además integra más variables en el modelo, cada una con una capacidad de influencia sobre la variable independiente, las cuales al ser valoradas por separado son significativas.

A pesar de que ambos modelos tienen una capacidad de predicción global superior al 74%, se recomienda adoptar el modelo de regresión logística binaria y no el de árboles de decisión, puesto que el primero tiene una mayor capacidad de predicción para los casos de incumplimiento, es decir, un mejor nivel de sensibilidad (82,4% versus 48,7%).

Por último, este modelo es más útil para discriminar entre los clientes que pueden presentar incumplimiento respecto de los que no, comparando sus áreas bajo la curva ROC, siendo de 87,4% en la regresión logística binaria y del 84,1% en el modelo árboles de decisión.

Otra de las ventajas del modelo de regresión logística binaria es que permite calcular la probabilidad de incumplimiento del cliente al momento de realizar la solicitud, dato necesario para el desarrollo de modelos basados en calificaciones internas, que, si bien no son requerimiento para empresas de factoring, hacen parte de las recomendaciones del Acuerdo de Basilea II para instituciones financieras.

Entre tanto, en el proceso de backtesting el modelo árboles de decisión mostró resultados satisfactorios, al comprobarse que el porcentaje global de acierto es

superior al 74%, sin embargo, su resultado fue inferior al arrojado en la muestra de entrenamiento. Por el contrario, en el proceso de backtesting del modelo de regresión logística binaria se obtuvieron incluso mejores resultados que con la muestra de entrenamiento.

Dadas las limitaciones en la cantidad y sesgo de la información hacia clientes más grandes, se considera que el modelo desarrollado debe ser usado como modelo de seguimiento y no de otorgamiento, puesto que no abarca todo el universo de clientes de la empresa. Así mismo se considera que este modelo solo tiene validez al interior de la compañía; además, a pesar de que tanto para el modelo de regresión logística binaria y el modelo de árboles de decisión las pruebas estadísticas muestran que tienen buenos ajustes, los coeficientes de algunas variables independientes no son consistentes, sin embargo, estos modelos son solo una aproximación de la realidad por lo que se sugiere sean actualizados mínimo con periodicidad anual, con el fin de ampliar la muestra poblacional, eliminar estos sesgos y verificar su idoneidad.
Pese a lo anterior, el sistema desarrollado sin ser óptimo, puede ser útil como herramienta para crear alertas tempranas para los clientes recurrentes de la entidad, ya que un cliente puede ingresar teniendo unos indicadores financieros que lo califiquen como de bajo riesgo, indicadores que con el tiempo pueden deteriorarse hasta llevarlo a un posible incumplimiento, para esto es necesario una continua actualización de los datos financieros de los clientes que permitan una toma de decisiones oportuna.

En la actualidad, la compañía Grupo Factoring de Occidente no tiene un sistema de calificación de clientes estandarizado y alejado de las perspectivas cualitativas del analista de crédito, por lo tanto se recomienda a la empresa iniciar con dos pasos fundamentales: primero, la estructuración de una hoja de vida de cada operación y sus participantes (emisor y pagador), con el fin de tener bases de datos sólidas y completas para futuras actualizaciones del modelo y un horizonte más amplio en cuanto a la cantidad de operaciones con información suficiente para introducir en ellos. Segundo, iniciar con la adopción del modelo de regresión logística binaria

desarrollado en este trabajo, con el fin de hacer seguimiento a los clientes para tomar decisiones de manera oportuna.

Si bien el modelo de credit scoring desarrollado en el presente trabajo no reemplazará la labor del analista de crédito de la compañía, se puede convertir en una herramienta primordial para condensar y analizar rápidamente altos volúmenes de información que, ante la escalabilidad y el crecimiento de este tipo de negocios, se vuelve cada día más importante y compleja de examinar.

Al 31 de diciembre de 2019 la compañía tuvo una cartera con vencimiento superior a 30 días de $14.837 millones de pesos, y una cartera superior a 90 días de $2.395 millones de pesos, que ocasionó la realización de provisiones contables de deterioro equivalentes a 934 millones de pesos. Se estima que con la adopción del modelo dichas provisiones puedan llegar a disminuir hasta en un 82,4%.

REFERENCIAS

Bank for International Settlements (2017). *Basel III: Finalising Post-crisis Reforms.* BIS.

Breiman, L., Friedman, J., Stone, Ch. y Olshen, R. A. (1984). *Classification and Regression Trees.* Chapman and Hall.

Cardona, P. A. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de Estadística*, *27*(2), 139-151.

García, R., González, J. y Jornet, J. (2010). *SPSS: Pruebas No Paramétricas.* https://www.uv.es/innomide/spss/SPSS/SPSS_0802A.pdf.

González, A. G. (2010). Determinantes del riesgo del crédito comercial en Colombia. *Reporte de Estabilidad Financiera*. https://1library.co/document/q7rk70ry-determinantes-del-riesgo-de-credito-comercial-en-colombia.html.

Haro, A. (2008). *Medición y control de riesgos financieros*. Limusa.

López, R. F. (2007). Análisis de los determinantes del riesgo de crédito en presencia de carteras de bajo incumplimiento. Una propuesta de aplicación. *Revista Europea de Dirección y Economía de la Empresa*, *16*(2), 71-92.

Martínez, E. M. (2008). Logit Model como modelo de elección discreta: origen y evolución. *Anuario Jurídico y Económico Escurialense*, (41), 469-484.

Rayo, S., Lara, J. y Camino, D. (2010). A Credit Scoring Model for Institutions of Microfinance under the Basel II Normative. *Journal of Economics, Finance and Administrative Science*, *15*(28), 90-124.

ROC: visualización (s. f.). *IBM Knowledge Center*. https://www.ibm.com/support/knowledgecenter/es/SSLVMB_subs/statistics_mainhelp_ddita/spss/base/idh_roc_display.html

Salazar, F. E. (2013). Cuantificación del riesgo de incumplimiento en créditos de libre inversión: un ejercicio econométrico para una entidad bancaria del municipio de Popayán, Colombia. *Estudios Gerenciales*, *29*(129), 416-427.

Sánchez, G. (2012). Regresión logística. En T. Luque, *Técnicas de análisis de datos en investigación de mercados* (págs. 431-468). Pirámide.

ANEXOS

Anexo 1. Cartera por edades valores absolutos (expresada en millones de pesos)

Corte	Corriente	Vencido					Total
		31 a 60	61 a 90	91 a 120	121 a 150	más de 150	
31/12/2017	122.996	4.331	2.580	281	182	1.234	131.605
31/12/2018	250.274	9.085	1.377	474	110	837	262.158
31/12/2019	232.369	10.321	2.121	1.373	289	733	247.206

* Cifras en millones de pesos COP, la cartera corriente también incluye la cartera vencida de 0 a 30 días, ya que los autores definen el punto de incumplimiento a partir del día 31.

Fuente: elaboración propia con datos internos de GFO.

Anexo 1. Cartera por edades valores relativos

Corte	Corriente	Vencido					Total
		31 a 60	61 a 90	91 a 120	121 a 150	más de 150	
31/12/2017	93,46%	3,29%	1,96%	0,21%	0,14%	0,94%	100,00%
31/12/2018	95,47%	3,47%	0,53%	0,18%	0,04%	0,32%	100,00%
31/12/2019	94,00%	4,18%	0,86%	0,56%	0,12%	0,30%	100,00%

Fuente: elaboración propia con datos internos de GFO.

Anexo 2. Indicadores financieros

N.º	Nombre del indicador	Fórmula
1	Razón corriente o índice de liquidez	Activo corriente/Pasivo corriente
2	Importancia de la cartera	Cartera/Total de activo corriente
3	Capital de trabajo	Activo corriente-Pasivo corriente
4	Prueba súper ácida	Activo corriente-Inventario-Cartera/ Pasivo corriente
5	Solidez	Activo total/Pasivo total
6	Ebitda	Utilidad operativa-Depreciaciones-Amortizaciones
7	Días de rotación de cartera	365/(Ingresos/Deudores)
8	Días de rotación de proveedores	365/(Costos/Proveedores)
9	Días de rotación de inventarios	365/(Costos/Inventarios)
10	Días de ciclo de efectivo	Días de cartera + Días de Inventario - Días de Proveedores
11	Días de rotación del activo total	365/(Ingresos/Activos)
12	Margen bruto	Utilidad bruta/Ingresos
13	Margen operacional	Utilidad operacional/Ingresos
14	Margen ebitda	Ebitda/Ingresos
15	Margen neto	Utilidad neta/Ingresos
16	ROI	Utilidad operativa/Activo total
17	ROE	Utilidad neta/Patrimonio
18	Endeudamiento	Pasivo total/Activo total
19	Endeudamiento financiero	Obligaciones financieras/Total de activos
20	Impacto carga financiera	Gastos financieros/Ventas netas
21	Cobertura intereses	Ebitda/Gastos financieros
22	Índice de solvencia	Activo/Pasivo

Fuente: elaboración propia.

Anexo 3. Coordenadas de la curva COR-modelo logit

Coordenadas de la curva

Variables de resultado de prueba

Punto de corte	Sensibilidad	Especificidad	Total
0,000	1,000	0,000	0,500
0,054	0,972	0,073	0,523
0,100	0,963	0,185	0,574
0,150	0,927	0,302	0,615
0,202	0,890	0,522	0,706
0,250	0,826	0,649	0,737
0,300	0,734	0,766	0,750
0,335	0,697	0,834	0,766
0,352	0,661	0,849	0,755
0,400	0,624	0,883	0,753
0,451	0,569	0,907	0,738
0,500	0,532	0,932	0,732
0,553	0,477	0,961	0,719
0,603	0,431	0,966	0,699
0,649	0,394	0,980	0,687
0,701	0,330	0,980	0,655
0,742	0,312	0,980	0,646
0,820	0,266	0,990	0,628
0,850	0,248	0,990	0,619
0,900	0,165	0,990	0,578
0,957	0,101	1,000	0,550
1,000	0,009	1,000	0,505

Fuente: datos arrojados por el software SPSS.

Anexo 4. Variables en la ecuación-modelo logit

Variables en la ecuación

	B	Error estándar	Wald	gl	Sig.	Exp(B)
Capital de trabajo	0,000	0,000	3,182	1,000	0,074	1,000
Margen operacional	-0,295	0,073	16,502	1,000	0,000	0,745
Margen ebitda	0,141	0,061	5,317	1,000	0,021	1,151
Margen neto	0,207	0,040	26,575	1,000	0,000	1,230
Endeudamiento	-0,085	0,029	8,668	1,000	0,003	0,919
Impacto carga financiera	20,074	4,120	23,744	1,000	0,000	522432318,575
Solvencia	-11,459	2,994	14,644	1,000	0,000	0,000
Constante	8,205	2,878	8,127	1,000	0,004	3658,213

Fuente: datos arrojados por el software SPSS.

Anexo 5. Prueba ómnibus de coeficientes-modelo logit

Pruebas ómnibus de coeficientes de modelo

	Chi-cuadrado	Gl	Sig.
Paso	5,343	1	0,021
Bloque	87,240	7	0,000
Modelo	87,240	7	0,000

a. Un valor negativo de chi-cuadrado indica que el valor de chi-cuadrado ha disminuido del paso anterior

Fuente: datos arrojados por el software SPSS.

Anexo 6. Resumen del modelo-modelo logit

Resumen del modelo

Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
271,474[d]	0,268	0,371

Fuente: datos arrojados por el software SPSS.

Anexo 7. Prueba de Hosmer y Lemeshow-modelo logit

Prueba de Hosmer y Lemeshow

Chi-cuadrado	gl	Sig.
4,572	8	0,802

Fuente: datos arrojados por el software SPSS.

Anexo 8. Tabla de contingencia para la prueba de Hosmer y Lemeshow-modelo logit

Tabla de contingencia para la prueba de Hosmer y Lemeshow

	Incumplimiento = 0		Incumplimiento = 1		
	Observado	Esperado	Observado	Esperado	Total
1	26	26,783	2	1,217	28
2	24	24,621	4	3,379	28
3	26	23,426	2	4,574	28
4	25	22,522	3	5,478	28
5	22	21,408	6	6,592	28
6	19	20,141	9	7,859	28
7	18	18,533	10	9,467	28
8	14	15,247	14	12,753	28
9	9	9,777	19	18,223	28
10	2	2,541	26	25,459	28

Fuente: datos arrojados por el software SPSS.

Anexo 9. Área bajo la curva ROC-modelo logit

Área bajo la curva

Variables de resultado de prueba

Área	Desv. error	Significación asintótica	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
0,826	0,026	0,000	0,775	0,878

Fuente: datos arrojados por el software SPSS.

Anexo 10. Prueba de Kolmogorov-Smirnov para una muestra-modelo logit

Prueba de Kolmogorov-Smirnov para una muestra

		Probabilidad pronosticada
N		314
Parámetros normales	Media	0,340
	Desv. desviación	0,263
Máximas diferencias extremas	Absoluto	0,163
	Positivo	0,163
	Negativo	-0,098
Estadístico de prueba		0,163
Sig. asintótica(bilateral)		,000[c]

a. La distribución de prueba es normal
b. Se calcula a partir de datos
c. Corrección de significación de Lilliefors

Fuente: datos arrojados por el software SPSS.

Anexo 11. Estadísticos descriptivos Kolmogorov-Smirnov-modelo logit

Estadísticos descriptivos

	N	Media	Desviación	Mínimo	Máximo	Percentiles 25	50 (Mediana)	75
Probabilidad pronosticada	314	0,340	0,263	0,000	1,000	0,157	0,255	0,479

Fuente: datos arrojados por el software SPSS.

Anexo 12. Resumen de procesamiento de casos-modelo logit

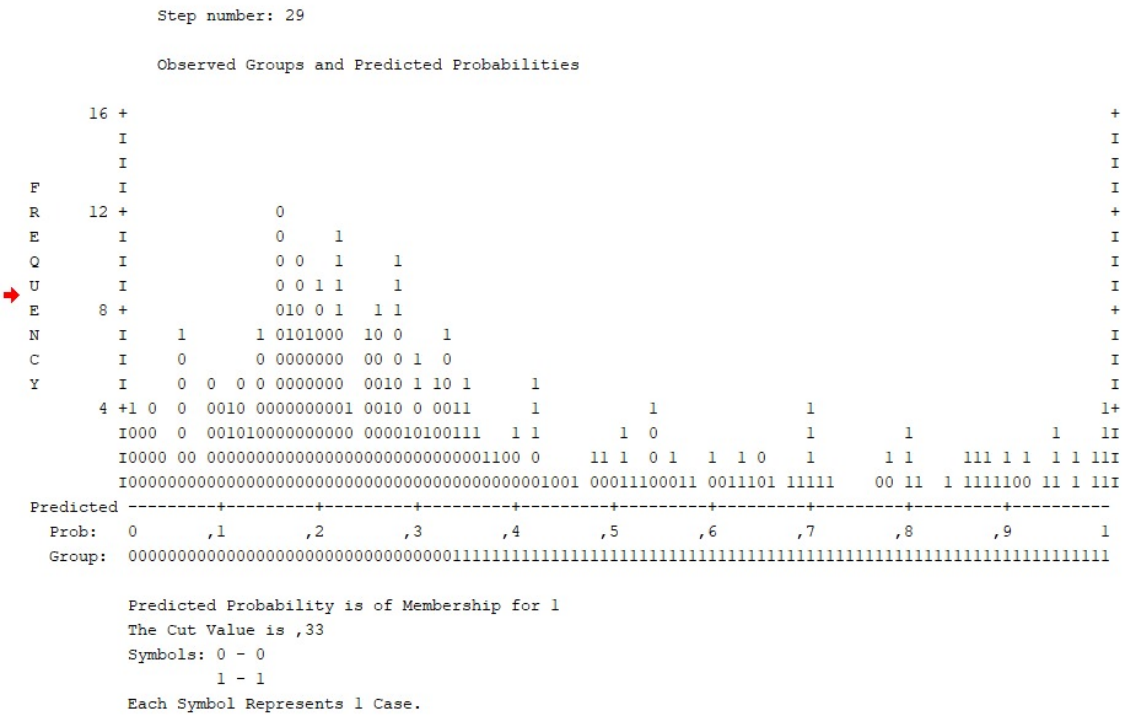
Resumen de procesamiento de casos

Casos sin ponderar		N	Porcentaje
Casos seleccionados	Incluido en el análisis	280	66,8
	Casos perdidos	34	8,1
	Total	314	74,9
Casos no seleccionados		105	25,1
Total		419	100,0

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos

Fuente: datos arrojados por el software SPSS.

Anexo 13. Histograma de probabilidades pronosticadas-modelo logit



Fuente: datos arrojados por el software SPSS.

Anexo 14. Coordenadas de la curva ROC-modelo árboles de decisión

Coordenadas de la curva

Variables de resultado de prueba

Positivo si es mayor o igual que	Sensibilidad	Especificidad	Total
0,0000	1,000	0,000	0,500
0,0304	0,978	0,339	0,659
0,0856	0,950	0,504	0,727
0,2690	0,885	0,621	0,753
0,4639	0,691	0,821	0,756
0,6476	0,453	0,939	0,696
1,0000	0,000	1,000	0,500

Fuente: datos arrojados por el software SPSS.

Anexo 15. Área bajo la curva ROC-modelo árboles de decisión

Área bajo la curva

Variables de resultado de prueba

Área	Desv. error[a]	95% de intervalo de confianza asintótico	
		Límite inferior	Límite superior
0,841	0,020	0,802	0,880

Fuente: datos arrojados por el software SPSS.

Anexo 16. Prueba de Kolmogorov-Smirnov-modelo árboles de decisión

Prueba de Kolmogorov-Smirnov para una muestra

		Prob. pronosticada
	N	419
Parámetros normales	Media	0,33311
	Desv. Desviación	0,28190
Máximas diferencias extremas	Absoluto	0,209
	Positivo	0,209
	Negativo	-0,140
Estadístico de prueba		0,209
Sig. asintótica(bilateral)		,000[c]

a. La distribución de prueba es normal. b. Se calcula a partir de datos. c. Corrección de significación de Lilliefors

Fuente: datos arrojados por el software SPSS.

Anexo 17. Tabla de contingencia-modelo árboles de decisión

Tabla cruzada sector* incumplimiento (1 = SÍ 0 = NO)

			Incumplimiento		
			NO	SÍ	Total
Sector	Agropecuario	Recuento	7	1	8
		% dentro de sector	87,5%	12,5%	100,0%
		% dentro de incumplimiento	2,5%	0,72%	1,9%
		% del total	1,7%	0,2%	1,9%
	Comercio	Recuento	106	29	135
		% dentro de sector	78,5%	21,5%	100,0%
		% dentro de incumplimiento	37,9%	20,86%	32,2%
		% del total	25,3%	6,9%	32,2%
	Industrial	Recuento	89	50	139
		% dentro de sector	64,0%	36,0%	100,0%
		% dentro de incumplimiento	31,8%	35,97%	33,2%
		% del total	21,2%	11,9%	33,2%
	Minero energético	Recuento	10	7	17
		% dentro de sector	58,8%	41,2%	100,0%
		% dentro de incumplimiento	3,6%	5,04%	4,1%
		% del total	2,4%	1,7%	4,1%
	Servicios	Recuento	68	52	120
		% dentro de sector	56,7%	43,3%	100,0%
		% dentro de incumplimiento	24,3%	37,41%	28,6%
		% del total	16,2%	12,4%	28,6%
	Total	Recuento	280	139	419
		% dentro de sector	66,8%	33,2%	100,0%
		% dentro de incumplimiento	100,0%	100,0%	100,0%
		% del total	66,8%	33,2%	100,0%

Fuente: datos arrojados por el software SPSS.