DESARROLLO DE UN MODELO ESTADÍSTICO QUE PERMITA ESTIMAR LA PROBABILIDAD DE INCUMPLIMIENTO DE LOS CLIENTES QUE TENIENDO EXPERIENCIA CREDITICIA, SOLICITAN CUPO DE TARJETA DE CRÉDITO

BRÍTER IVÁN GONZÁLEZ MORALES

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
DEPARTAMENTO DE CIENCIAS BÁSICAS
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ D.C.
2015

DESARROLLO DE UN MODELO ESTADÍSTICO QUE PERMITA ESTIMAR LA PROBABILIDAD DE INCUMPLIMIENTO DE LOS CLIENTES QUE TENIENDO EXPERIENCIA CREDITICIA, SOLICITAN CUPO DE TARJETA DE CRÉDITO

BRÍTER IVÁN GONZÁLEZ MORALES

Trabajo de grado para optar el título de Especialista en Estadística Aplicada

Director

JUAN PABLO MOJICA MACÍAS

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
DEPARTAMENTO DE CIENCIAS BÁSICAS
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ D.C.
2015

Nota de Aceptación

_____

_____

_____

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

Bogotá, D.C., _____

Con todo mi amor, a mi esposa Elsa Patricia, por la confianza que ha depositado en mí, por su apoyo y entrega incondicional; a mis hijos Daniela y Bríter Andrés, sin duda alguna fueron el motor de este proyecto.

A mi madre y mi padre que me formaron en el camino del bien, del trabajo duro, y de la perseverancia para alcanzar las metas propuestas.

Agradecimientos especiales al Instituto Técnico Central La Salle por la estupenda formación que recibí en las áreas de matemáticas, física y química.

Bríter Iván González

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

LISTA DE ECUACIONES

pág.

GLOSARIO

***DEFAULT*:** se califica con este estado las obligaciones crediticias que han superado determinado nivel de morosidad, donde se considera que la recuperación del crédito es altamente improbable.

MINERÍA DE DATOS: es la actividad que permite la formulación de un modelo estadístico a partir del análisis exploratorio de la información, buscando patrones ocultos de comportamiento en cada atributo del individuo en estudio, respecto de la variable objetivo (Molina, 2014).

MODELO ESTADÍSTICO: es el resultado de un proceso analítico de información, que estima el valor que tomará la variable respuesta a partir de los valores dados por otras variables o atributos denominadas explicativas (Fernández & Pérez, 2005).

ODDS RATIO: es el término utilizado para señalar la proporción de casos buenos (donde no se presentó el evento de incumplimiento) frente a los casos malos (donde se presentó *default*). Su análisis permite establecer la diferenciación real, a la luz del *default*, entre categorías de una misma variable, o entre rangos de probabilidad de incumplimiento.

PARSIMONIA: es el principio que deben cumplir los modelos para garantizar que guardando la mayor simplicidad desde el punto de vista de número de variables, tienen la mayor efectividad posible, medida a través de la verosimilitud de sus estimaciones. (Guillermo de Ockham 1349).

RANGOS DE UNA VARIABLE: son las subdivisiones del contenido de cada variable de modelo a partir del estudio de probabilidades de la ocurrencia del evento analizado; en el caso de las variables discretas corresponde a cada valor o grupo de valores que guarda similitud en su nivel de incumplimiento observado; para las variables continuas se conforman intervalos con la mayor homogeneidad posible respecto del nivel de default observado (Quezada, 2014).

REGRESIÓN MATEMÁTICA: es el resultado final del proceso de modelado, se representa mediante una ecuación que incluye tantos términos como variables independientes resultaron estadísticamente significativas para estimar el comportamiento de la variable objetivo o dependiente.

VARIABLE OBJETIVO: también conocida como variable explicada o dependiente, es el atributo a predecir del individuo en estudio, aquel cuyo comportamiento se modela, estimando su valor a partir de atributos denominados variables explicativas o independientes.

VARIABLE INDEPENDIENTE: se considera aquella variable o atributo de los individuos en estudio, que es incluida en la regresión matemática dada la evidencia estadística de su correlación con el comportamiento de la variable objetivo.

VARIABLE CONTINUA: variable de contenido numérico, de la que se conoce su dominio, más no su valor preciso, ejemplo el valor del ingreso, la edad, años trabajando en la empresa, entre otras (Ezequiel, 2013).

VARIABLE DISCRETA: variable que toma valores con dominio establecido previamente, dentro del que solo toma valores exactos, tales como género, nivel de estudios, estado civil. También se conocen como variables categóricas y su mayor aplicabilidad está en las variables dicotómicas que toman valor uno o cero (Ezequiel, 2013).

RESUMEN

El presente trabajo persigue establecer el perfil de riesgo crediticio de cada cliente del producto tarjeta de crédito, estimando su probable nivel de incumplimiento a partir de información sociodemográfica y de comportamiento disponible.

En el desarrollo se evidencia la aplicación de metodologías estadísticas, en particular las que tienen que ver con el establecimiento de probabilidades, descripción y análisis del contenido de variables, prueba de hipótesis, árboles de decisión y método de regresión logística.

La aplicación de pruebas de desempeño sobre la solución encontrada, muestra buen ajuste de la estimación de incumplimiento frente a los eventos observados, así como una gran capacidad de discriminación sobre los niveles de morosidad.

Palabras clave: riesgo crediticio, probabilidad de incumplimiento, árboles de clasificación, regresión logística.

ABSTRACT

This document pretends to assign the credit risk profile of each costumer of credit card product, through estimation of his probability of default, based on demographic and behavioral information.

The development reveals the aplication of statistical methods, especially those related with establishing probabilities, description and analysis of the data contents, hypothesis pruf, decition trees and logistic regression.

The aplication of performance test for disegned solution, shows good fit of the breach estimation versus to observed events, as well as a big discrimination capacity about the non performing loans.

Key words: credit risk, probability of default, decition trees, logistic regression.

1. INTRODUCCIÓN

El incumplimiento de los clientes o materialización del riesgo de crédito es una realidad incluida en los presupuestos y proyecciones financieras de los bancos. Sin embargo, cuando este incumplimiento se presenta muy desviado respecto a los niveles estimados, genera problemas de liquidez y pérdidas no esperadas sobre los estados financieros de la institución, obligando que en la siguiente formulación anual de presupuesto se eleven los costos de los productos e incluso se tomen medidas para reducir la colocación de créditos.

Por esta razón, las entidades bancarias requieren aprobar créditos a clientes que atiendan los pagos dentro de los términos convenidos, generando los ingresos suficientes para atender las obligaciones adquiridas con los actores del sistema financiero que suministran los fondos para la colocación de crédito, tales como titulares de cuentas de ahorro o corriente, depositantes de certificados de depósito a término, otros establecimientos de crédito, y el propio accionista de la compañía.

De otro lado, se tiene una sociedad en crecimiento con gran necesidad de bancarización, donde sus habitantes pretenden obtener crédito mediante la solicitud de cupos rotativos que les permita financiar a corto plazo la adquisición de bienes o servicios que mejoren su calidad de vida.

El método IRB (Internal Ratings Based) permite valorar la probabilidad de incumplimiento a partir de estimaciones propias de las entidades financieras, con el fin de incluir sus particularidades en la medición del riesgo de crédito. Se cuenta así con estimaciones adecuadas del incumplimiento probable para cada cliente, según algunas variables sociodemográficas y de comportamiento disponibles en la entidad o en alguna central de riesgo, garantizando la generación de pérdidas dentro de los límites tolerados por política interna (Álvarez & Osorio, 2014).

En este orden de ideas, es viable y necesario que los establecimientos de crédito cuenten con metodologías que permitan responder concreta, eficiente y oportunamente a cada solicitud de tarjeta de crédito procurando dar viabilidad a la mayor cantidad de clientes que sea posible, pero teniendo certeza de que aún el cliente admitido con el peor perfil de riesgo crediticio, generará determinada rentabilidad mínima.

Las metodologías citadas parten de estudios de rentabilidad que señalan el nivel máximo de incumplimiento tolerado; es ahí donde se requiere de un modelo estadístico que permita estimar el incumplimiento para cada cliente y así evaluar si al estar dentro del límite tolerado, se le permite acceder al crédito.

La utilización de modelos estadísticos para identificar perfiles según probabilidad de incumplimiento es una práctica que satisface plenamente la necesidad planteada y es contemplada ampliamente en literatura disponible sobre estimación de expectativas de impago a partir de métodos de base actuarial (Norman, 2011).

La regresión logística es la herramienta más utilizada en las actividades de modelado para predecir eventos de incumplimiento en riesgo de crédito, dado el mejor ajuste que tiene su distribución frente a la de los eventos de incumplimiento observados (Fernández & Pérez, 2005).

Con esta ambientación se formula como problemática a resolver, si es posible desarrollar un modelo estadístico que estime la probabilidad de incumplimiento de un cliente potencial de tarjeta de crédito a partir de su información sociodemográfica y su historial de comportamiento.

1.1. OBJETIVOS

1.1.1 OBJETIVO GENERAL

Desarrollar un modelo estadístico que permita estimar la probabilidad de incumplimiento de los clientes con algún tipo de experiencia crediticia que solicitan cupos de tarjeta de crédito, a partir de información residente en la central de riesgo y al interior de la entidad financiera.

1.1.2 OBJETIVOS ESPECÍFICOS

- Establecer una muestra de los clientes con experiencia crediticia que se acercaron a una entidad financiera en determinada ventana de tiempo para solicitar el producto de tarjeta de crédito.

- Precisar la variable objetivo de forma tal que revele el posible incumplimiento de los clientes, es decir el hecho de que no atiendan los compromisos de pago adquiridos deteriorando su perfil de riesgo y afectando los estados financieros de la entidad.

- Desarrollar análisis de correlación para cada una de las variables independientes frente a la variable dependiente u objetivo, estableciendo así la mejor segmentación en grupos homogéneos según la incidencia que presenten sobre el incumplimiento observado.

- Generar un modelo de regresión logística que presente la mayor capacidad de discriminación entre clientes con alto y bajo nivel de incumplimiento, así como la mayor bondad de ajuste de las estimaciones frente a las observaciones históricas.

1.2. ALCANCE Y DELIMITACIONES

El entregable del proyecto es un modelo estadístico que estime probabilidades de incumplimiento, incluyendo la documentación pertinente sobre población de desarrollo, definición de variable objetivo, análisis descriptivo de las variables independientes, ecuación resultante de la regresión y pruebas de desempeño.

Para la formulación del modelo se utilizarán árboles de clasificación y regresión logística sobre los atributos socio demográficos y de comportamiento crediticio que se tengan disponibles en la base de datos de la institución financiera que los proporcionó.

Este proyecto no incluye captura de información ni aplicación de pruebas para identificación de observaciones influyentes; se parte de la directriz de construir sobre data existente capturada con procesos de calidad al interior de un banco comercial, que garantizan integridad, veracidad y limpieza de los datos.

2. MARCO TEÓRICO Y CONCEPTUAL

RIESGO DE CRÉDITO

La Superintendencia Financiera de Colombia en su circular básica contable define el riesgo de crédito como la posibilidad de que una entidad financiera incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor incumpla sus obligaciones.

TÉCNICA ESTADÍSTICA

La teoría y la aplicación de las estadísticas descriptiva e inferencial facilitan la comprensión de los eventos de incumplimiento observados en la actividad de colocación masiva de créditos de consumo.

En el desarrollo de este proyecto se utilizarán ampliamente los dos frentes, estudiando el contenido de la información disponible, los tipos de dato y sus distribuciones, así como la correlación entre variables independientes y variable objetivo, para finalmente definir un modelo que se ajuste a la realidad observada.

MUESTREO ALEATORIO SIMPLE

Como la inferencia estadística se formula con base en una muestra de objetos de la población de interés, el proceso mediante el cual se obtiene debe asegurar la selección de una buena muestra garantizando que todos los registros tienen la misma probabilidad de formar parte de ella (Canavos, 1988).

Con este fin se aplicará técnica de muestreo aleatorio simple coordinado negativo asignando números aleatorios a toda la población, ordenándola por este número aleatorio y seleccionando el volumen de registros requerido.

Mediante la ecuación 1 se validará la suficiencia del tamaño de la muestra; ésta debe ser mayor al tamaño n, dónde N es el tamaño poblacional, Z es el valor crítico sobre una distribución normal para nivel de confianza determinado por $\propto$, p es la probabilidad de ocurrencia del evento en estudio, $1-p$ es la probabilidad de no ocurrencia y err es el margen de error tolerado.

$$n = \frac{N * Z_{\frac{\propto}{2}}^{2} * p * (1-p)}{err^2 * (N-1) + Z_{\frac{\propto}{2}}^{2} * p * (1-p)}$$

Ecuación 1. Tamaño de la muestra

VARIABLE OBJETIVO

Es la variable cuyo contenido se quiere estudiar o explicar mediante su relación con las demás variables del cliente, por esta razón también se le denomina variable dependiente o explicada.

En modelos de comportamiento crediticio se le denomina *default* y representa una condición dicotómica de fallido o no fallido, catalogando al cliente en un estado de fallido si su nivel de morosidad observado superó determinado umbral dónde la mayoría de clientes no vuelve a ponerse al día (Fernández & Pérez, 2005).

VARIABLES INDEPENDIENTES

Las variables independientes conforman el conjunto de atributos del sujeto en estudio, sobre los cuales se basa la hipótesis nula de que no son estadísticamente significativos respecto al nivel de incumplimiento observado, mientras que la hipótesis alterna por su parte señala que sí hay evidencia estadística de correlación entre el contenido de estas variables y el comportamiento de la variable en estudio o variable objetivo (Hosmer & Lemeshow, 2013).

Las variables independientes consideradas categóricas, tales como el Género, la Región, el Estado Civil o el Nivel de Estudios de las personas, deberán transformarse en k-1 variables dicotómicas, donde k es el número de categorías de la variable original y para cada una se señalará con uno o cero la presencia o ausencia de la categoría respectiva en cada elemento de la data a modelar (Hosmer & Lemeshow, 2013).

MODELO DE REGRESIÓN LOGÍSTICA

La regresión logística es el resultado del modelado; como se puede ver en la ecuación número 2, a partir del contenido de m variables independientes, estima la PI (probabilidad de caer en *default* o presentar incumplimiento).

$$PI = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+ \ldots\ldots +\beta_m X_m)}}$$

Ecuación 2. Probabilidad de Incumplimiento

La bondad de este tipo de regresión es su capacidad de ajustarse a la distribución de los eventos observados; permite asignar a cada elemento de la población estudiada una probabilidad de que el evento *default* se presente, y facilita el establecer los *odds ratio* entre diferentes intervalos según la probabilidad de incumplimiento.

ESTADÍSTICO DE PRUEBA DE SIGNIFICANCIA

El modelo basado en regresión logística plantea para cada una de las variables independientes la hipótesis nula de que no son estadísticamente influyentes en la variable *default*, y por el contrario se plantea como hipótesis alterna, que sí hay relación significativa entre cada variable independiente y la dependiente. Esto es,

Ho: $\beta i = 0$ El coeficiente de cada variable en la ecuación 2 es cero, dado que no influye significativamente en la probabilidad de *default.*

Ha: $\beta i \neq 0$ El coeficiente de cada variable en la ecuación 2 es diferente de cero, dado que sí influye significativamente en la probabilidad de *default.*

Para probar la significancia de cada variable en el modelo, la hipótesis se contrasta mediante el estadístico de Wald (Hosmer & Lemeshow, 2013) el cual se estima mediante la ecuación 3.

$$Wald = \left(\frac{\beta_i}{Err\ Estándar\ de\ \beta_i}\right)^2$$

Ecuación 3. Estadístico de Wald sobre significancia de las variables

La evaluación del estadístico, más exactamente de su raíz cuadrada, se hace frente a la distribución normal fijando como punto crítico el $Z_{(1-\propto/2)}$ con $\propto$ correspondiente a una zona de rechazo de 5%. En otras palabras, si la raíz cuadrada del estadístico de Wald sobre una distribución normal arroja una probabilidad inferior al 97.5% se acepta la hipótesis nula de que la respectiva variable no ejerce influencia sobre la variable objetivo.

CAPACIDAD DE DISCRIMINACIÓN

La prueba conocida como KS (Kolmogorov Smirnov), diseñada por el matemático ruso Andrei Kolmogorov, se basa en la máxima distancia obtenida entre las curvas de distribución acumulada de clientes malos y clientes buenos según deciles (grupos de 10% de la población) ordenados por las probabilidades de incumplimiento estimadas por el modelo.

Un modelo con alto KS acumula mayor cantidad de malos en los deciles con probabilidad de incumplimiento alta, mientras que en rangos de baja probabilidad de incumplimiento agrega un mínimo de registros en condición de *default*.

El coeficiente GINI, llamado así por el italiano Corrado Gini, fue creado para analizar la desigualdad en los ingresos de un país; su utilidad se ha extendido a estos modelos logísticos permitiendo evaluar su capacidad para clasificar individuos entre los que caen en *default* y los que no lo hacen.

Este coeficiente señala la proporción que cubre el área bajo la curva de la distribución acumulada de malos, obteniendo un cubrimiento nulo en modelos donde la acumulación de malos sea igual a la acumulación de buenos, y cubrimiento de 100% en modelos donde en el decil de mayor PI, se acumula la totalidad de individuos malos.

El uso de estas medidas estadísticas para evaluar la capacidad de discriminación en los modelos de riesgo de crédito es una práctica común entre los analistas de crédito. No obstante que los parámetros de evaluación pueden variar entre diferentes analistas, las prácticas de líderes internacionales en modelado para riesgo de crédito, señalan como aceptables los valores obtenidos entre 20% y 30%, buenos los resultados entre 30% y 40%, muy buenos entre 40% y 50%, excelentes entre 50% y 60%, y califican como poco usuales los resultados superiores a 60% (Laredo, Pedroso & Okaze, 2012).

BONDAD DE AJUSTE

La bondad de ajuste del modelo se verifica mediante el test de Hosmer y Lemeshow al comparar el número de casos observados con incumplimiento frente al número de casos pronosticados en cada uno de diez niveles clasificados por la probabilidad de incumplimiento (Hosmer & Lemeshow, 2013).

El estadístico de validación se calcula con la ecuación número 4 donde k es el número de rangos de la tabla de desempeño, MO_i es el número de malos observados en cada uno de los k rangos, n es el número de casos totales de cada rango, y PI es la probabilidad de incumplimiento estimada por el modelo para el rango respectivo.

$$\hat{C} = \sum_{i=1}^{k} \frac{(MO_i - n_i * PI_i)^2}{n_i * PI_i * (1 - PI_i)}$$

Ecuación 4. Estadístico para prueba de bondad de ajuste

El resultado del estadístico de validación se evalúa sobre una distribución χ^2_{k-2} donde k es el número de rangos de la tabla de desempeño, es decir una distribución Chi cuadrado con 10 - 2 grados de libertad.

Dado que el numerador del estadístico refleja la magnitud de los errores, éste será más pequeño entre menos error se observe, por tal razón a mayor p-valor (probabilidad a la derecha del valor crítico según la distribución chi cuadrado) mayor es el ajuste de los valores observados respecto de lo esperados.

3. MARCO METODOLÓGICO

La entidad financiera que suministró la información para la construcción de este modelo es un establecimiento de crédito que cuenta con gran experiencia en el otorgamiento de crédito de consumo y tiene prioridad en sus objetivos estratégicos respecto a incrementar de manera sólida y rentable su participación en el volumen de tarjetas de crédito colocadas en el mercado Colombiano.

Al interior de la entidad financiera, se han establecido los términos y mecanismos de acceso a la información requerida para el desarrollo del modelo, los cuales acepta el autor de este trabajo y con los cuales se compromete a: utilizar la data sin ningún tipo de delegación a terceros, presentar el trabajo de grado de manera individual, ejecutar los análisis en el computador asignado por el Banco y con el software que para tal fin provee el Banco, y proporcionar información a terceros tales como director de trabajo y jurados solo en estado agregado, nunca con detalle de productos o números de identificación de clientes.

Este trabajo está enmarcado dentro del tipo de estudio estadístico correlacional, bajo el entendido de que pretende encontrar variables explicativas cuyo contenido esté correlacionado con el comportamiento de la variable *default.*

Respecto al software a utilizar en el desarrollo del modelo, SPSS 22 es un programa estadístico informático usado ampliamente en este tipo de actividades dada la variedad de fuentes de ayuda y capacitación disponibles así como su capacidad para trabajar grandes bases de datos (Quezada, 2014).

La base de datos está constituida por las solicitudes de cupo rotativo o tarjeta de crédito que recibió el Banco por parte de clientes con experiencia crediticia entre julio del año 2013 y junio del año 2014; esto permite analizar el comportamiento más reciente y eliminar sesgos por la estacionalidad de algunos meses del año.

Para adelantar este estudio, se cuenta con autorización de acceso de lectura al repositorio de datos donde se archiva la información socio demográfica, financiera y de comportamiento que se capturó al momento del estudio de crédito; cada registro de la base de datos es un elemento de la población a estudiar.

Adicionalmente se cuenta con acceso de lectura a la información del comportamiento crediticio posterior al momento del estudio, permitiendo establecer para cada elemento de la población, si presentó o no el evento que se defina como variable objetivo.

La muestra de información a extractar de la base de datos que contiene la información sociodemográfica y de comportamiento crediticio, se establece mediante técnica de muestreo aleatorio simple, garantizando que su composición respecto a las principales variables de estudio sea similar a la poblacional.

En la tabla 1 se listan las variables que serán sujetos de análisis descriptivo y generación de regresiones, con el fin de establecer la variable explicada (dependiente) y las variables explicativas (independientes).

Variable	Observaciones
Origen de la solicitud de crédito	Preaprobado, mercado natural (solicitud)
Antigüedad con el Banco en años	>=0
Edad	>18 y <=90
Estado civil	Soltero, unión libre, casado, separado, viudo
Nivel de estudios	Ninguno, primaria, bachiller, técnico, universitario, especialista
Actividad económica	Empleado, independiente, rentista, transportador, pensionado, otros
Tipo de contrato	Fijo, indefinido, provisional, libre remoción, otros

Tabla 1. Lista de variables

Variable	Observaciones
Tipo de vivienda	Familiar, en arriendo, propia, hipotecada
Personas a cargo	>=0 y <=11
Región del país	Bogotá, Cali, Medellín, Barranquilla, Bucaramanga
Indicador de cuenta de nómina	S/N
Tipo de cupo	Tarjeta crédito, rotativo
Canal de trámite	Oficinas, fuerza externa, internet
Ingreso en SMLMV	>=0.5 y <=200
Valor solicitado en SMLMV	>=0.5 y <=200
Número de consultas	>=0 y <=31
Uso rotativos	>=0 Se identifica con -1 a quienes no tienen crédito rotativo
Antigüedad en meses del crédito más reciente	>0 y <=476
Antigüedad en meses del crédito más antiguo	>0 y <=608
Antigüedad en meses de la TC más antigua	>=0 Se identifica con -1 a quienes no tienen crédito rotativo
Mora Máxima en los 12 meses previos	>=0
Puntaje del score genérico en la solicitud	>0 y <=950
Mora al momento de la solicitud	>=0
Puntaje del score genérico 12 meses después	>0 y <=950
Mora Máxima en los siguientes 12 meses	>=0

Tabla 1. Lista de variables

La muestra seleccionada se segmenta aleatoriamente en dos particiones, 50% para entrenamiento del modelo y 50% para aplicación de pruebas de desempeño.

A partir del puntaje score genérico entregado por la central de riesgo y el nivel de mora máximo observado posteriormente, se define la variable a usar como *default.*

Mediante metodología de árboles de decisión se evalúa la existencia de correlación entre el comportamiento de la variable *default* y el contenido de cada una de las demás variables; este análisis se evidencia mediante gráficos de composición y nivel de incumplimiento por categoría.

El desarrollo del modelo incluye la evaluación de diferentes escenarios de regresión a partir del set de variables, buscando el modelo con mayor capacidad de discriminación, mejor ajuste y menor número de variables, conservando el principio de parsimonia.

Las tablas de desempeño generadas tanto en la partición de entrenamiento como en la partición de prueba, incluyen pruebas estadísticas del modelo sobre su capacidad de discriminación, tales como indicador KS y Coeficiente GINI, así como sobre su bondad de ajuste mediante test de Hosmer Lemeshow.

4. RESULTADOS

4.1 SELECCIÓN DE LA MUESTRA

La población de clientes con experiencia crediticia que solicitaron tarjeta de crédito o cupo rotativo durante el período mencionado, julio 2013 a junio 2014, está conformada por 114.302 registros.

Las solicitudes de crédito llegaron por dos canales diferentes, por un lado el que llamaremos mercado natural hace referencia a las solicitudes que el cliente hace directamente en la oficina, mientras que los preaprobados se refieren a solicitudes generadas con posterioridad a una oferta comercial efectuada por el Banco.

Teniendo en cuenta que la entidad financiera autorizó consultar el comportamiento crediticio de un número cercano a 80.000 registros, se decide incluir en la muestra el censo de las solicitudes de origen preaprobado y una muestra de las solicitudes originadas por mercado natural, reduciendo el tamaño de la población a una muestra de 82.550 registros.

El tamaño muestral obtenido se considera suficiente ya que supera ampliamente el mínimo requerido de 14.485 elementos asumiendo nivel de confianza del 99% y margen de error del 1% (ver ecuación número 1).

Adicionalmente, para cada uno de los orígenes mencionados, se aplica muestreo aleatorio simple para segmentar la muestra en dos partes de similar tamaño, de tal forma que una partición se usará para desarrollar o entrenar el modelo y con la otra se practicarán pruebas de desempeño.

	Casos
Preaprobados - Partición para desarrollo	24,644
Preaprobados - Partición para prueba	24,573
Mercado natural - Partición para desarrollo	16,660
Mercado natural - Partición para prueba	16,673
Total general	82,550

Tabla 2. Distribución de la muestra

Como se aprecia en la tabla 2, las particiones para desarrollo y prueba quedan conformadas por similar número de registros, 41.304 para entrenamiento y 41.246 para prueba.

La figura 1 permite verificar que tanto la selección de la muestra como la asignación de particiones efectivamente se efectuaron de manera aleatoria. La distribución de cada grupo respecto al ingreso de los clientes muestra similaridad al evaluar percentiles 1 y 99, mediana, y cuartiles 1 y 3.

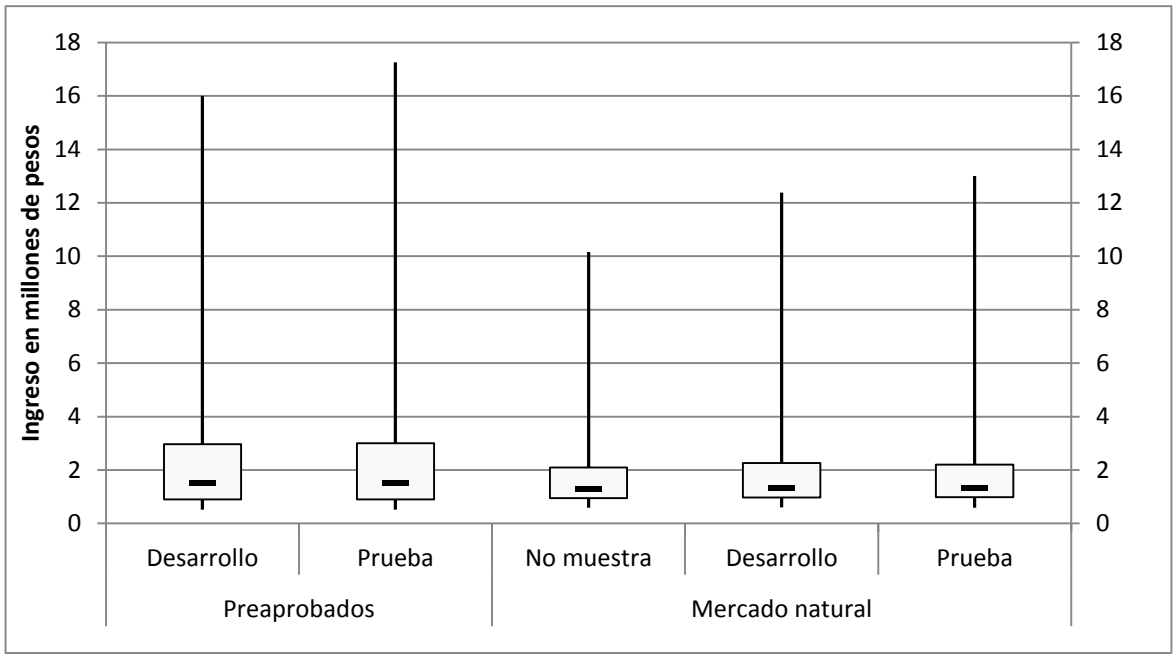


Figura 1. Distribución del ingreso en las particiones

4.2 ESTABLECIMIENTO DEL *DEFAULT*

El score genérico es un puntaje calculado por la central de riesgo para cada cliente; éste puntaje refleja el perfil de riesgo crediticio, asignando puntajes bajos a los clientes con mayor nivel de morosidad y sobreendeudamiento. Por lo anterior, se propone como variable *default* el hecho de que en una medición posterior (12 meses) el cliente presente score genérico muy bajo o nivel de mora superior a 90 días. Como se puede apreciar, se está definiendo una variable dicotómica que señala si se cayó o no en estado de *default.*

Con el fin de establecer el nivel de score genérico límite para el *default* se efectúa un análisis gráfico sobre el contenido de la variable, buscando con criterio experto el punto a partir del cual se aprecia simetría en la distribución, permitiendo descartar los casos con puntaje inferior.

En la figura 2 se presenta la distribución de la muestra en los diferentes niveles del score genérico proporcionado por la central de riesgo. Se aprecia que con los puntajes superiores a 500 se consigue una distribución relativamente simétrica que deja a su izquierda los clientes de peor perfil de riesgo crediticio.

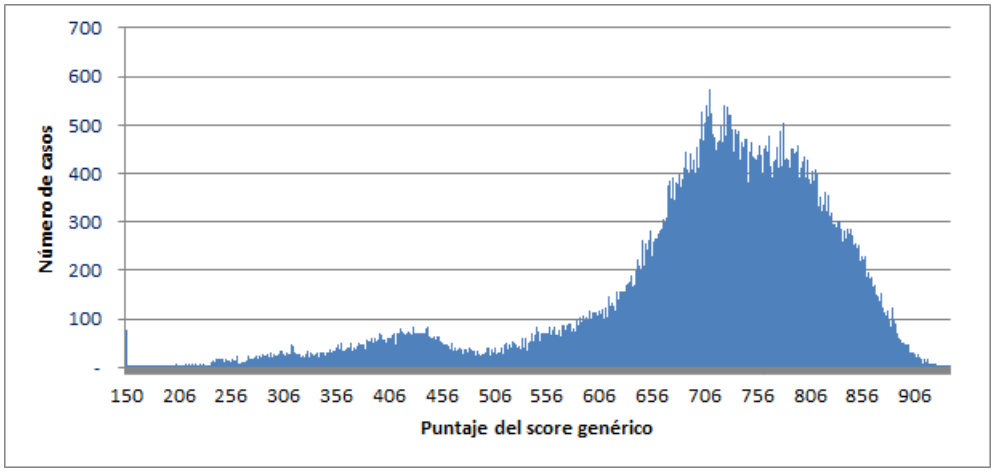


Figura 2. Distribución del score genérico

El concepto de “malo evidente” hace referencia a los casos que no se deben incluir en el proceso de modelado en razón a que al momento de la observación ya están en *default,* es decir que ya no hace sentido estimar una probabilidad de incumplir porque el evento de incumplimiento ya ocurrió.

Se debe hacer precisión que el malo evidente es aquel que al momento de solicitar el crédito ya está en condición de *default*, mientras que los casos que más adelante se citan como malos son los que cayeron a *default* durante los siguientes 12 meses al momento de la solicitud.

Con la definición de *default,* mora superior a 90 días o score genérico inferior a 500 se filtra la base de datos obteniendo la muestra definitiva para desarrollo o entrenamiento y prueba del modelo, ver cifras en la tabla 3.

	Malos evidentes	Muestra definitiva	Total general
Partición para desarrollo	5,947	35,357	41,304
Partición para prueba	6,000	35,246	41,246
Total general	11,947	70,603	82,550

Tabla 3. Separación de malos evidentes y muestra definitiva

La evaluación de caída a *default* se efectúa para cada registro en la ventana de tiempo de los 12 meses siguientes al momento en que se estudió el crédito.

	No default	Sí default	Total	Tasa de malos	odds ratio
Partición para desarrollo	29,858	5,499	35,357	15.6%	5.4
Partición para prueba	29,706	5,540	35,246	15.7%	5.4
Total	59,564	11,039	70,603	15.6%	5.4

Tabla 4. Tasa de malos y odds ratio en las dos particiones

La tasa de malos es el porcentaje de casos que cayeron en *default* sobre el número de casos totales; en la tabla 4 se aprecia que la tasa de malos para las dos particiones es prácticamente la misma 15.6% y 15.7%.

Respecto al *odds ratio* es un indicador que hace referencia a la proporción del número de casos buenos, es decir registros que no cayeron en *default,* frente al número de casos que sí cayeron en incumplimiento; para una y otra partición se concluye que por cada incumplido se aprecian 5.4 clientes que no lo son.

La tasa de malos es la que el modelo debe estimar; debe tener la capacidad de reconocer grupos de clientes con diferentes probabilidades de incumplir y acertar en esa estimación de incumplimiento con un razonable margen de error.

4.3 ANÁLISIS DESCRIPTIVO

Las variables independientes son el conjunto de atributos del cliente, sobre los cuales se basa la hipótesis nula de que no son estadísticamente relevantes, es decir que no inciden en el comportamiento de la variable objetivo o *default*; mientras que identificaremos como hipótesis alterna el hecho de que sí haya relación significativa entre cada variable explicativa y la explicada.

A continuación se presenta la segmentación de cada una de las variables estudiadas según el nivel de incumplimiento o *default* observado en las diferentes categorías o intervalos definidos por el software de clasificación mediante árboles de decisión. La categoría identificada como “Sí” hace referencia a los casos que cayeron en *default*.

Origen de la solicitud de crédito

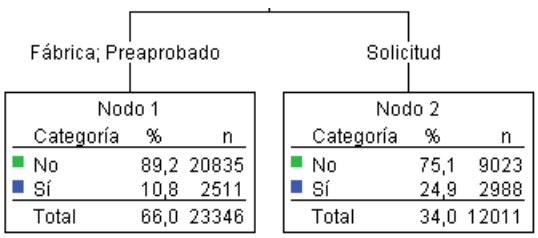


Figura 3. Segmentación de la variable Origen

El origen de la solicitud explica si el cliente hizo la solicitud de tarjeta de crédito directamente o por un proceso de preaprobación a partir de otro producto o por una campaña masiva de oferta comercial. Se sugiere la hipótesis nula de que el incumplimiento observado no guarda relación con el origen de la solicitud.

El análisis de esta variable muestra discriminación sobre el incumplimiento; en la figura 3 se aprecia que de los 12.011 casos cuyo origen fue el mercado natural (solicitud), es decir el 34% de los casos, 2.988 llegaron a *default* es decir un 24.9%; mientras que en los procesos de preaprobación (66% de los casos) la tasa de *default* tan solo fue del 10.8%.

Antigüedad en años del cliente con el Banco

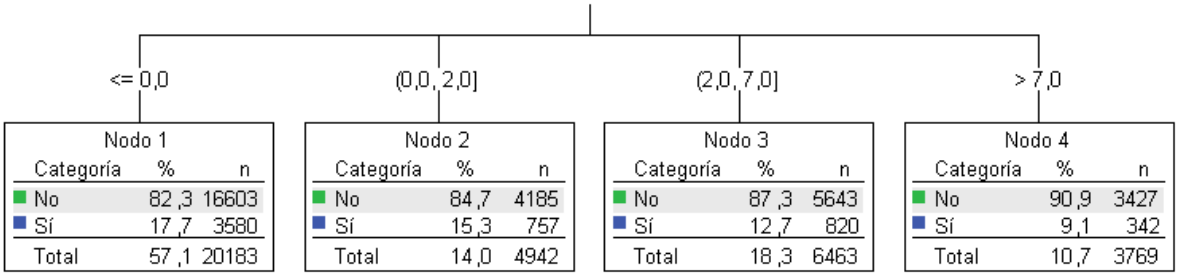


Figura 4. Segmentación de la variable Antigüedad en el Banco

Sobre la variable antigüedad en el Banco se sugiere la hipótesis nula de que el incumplimiento observado no guarda relación con la antigüedad que el cliente ha acumulado en el establecimiento de crédito al momento del estudio. La hipótesis alterna por su parte sugiere que los clientes recién vinculados al Banco tendrán mayor probabilidad de incumplir y los que tienen mayor antigüedad con el establecimiento de crédito atenderán mejor sus obligaciones crediticias.

La figura 4 muestra una segmentación que establece cuatro nodos discriminando la tasa de malos desde el 9.1% para los clientes con más de siete años de antigüedad en la institución hasta un 17.7% observado en clientes nuevos.

Edad del cliente en años

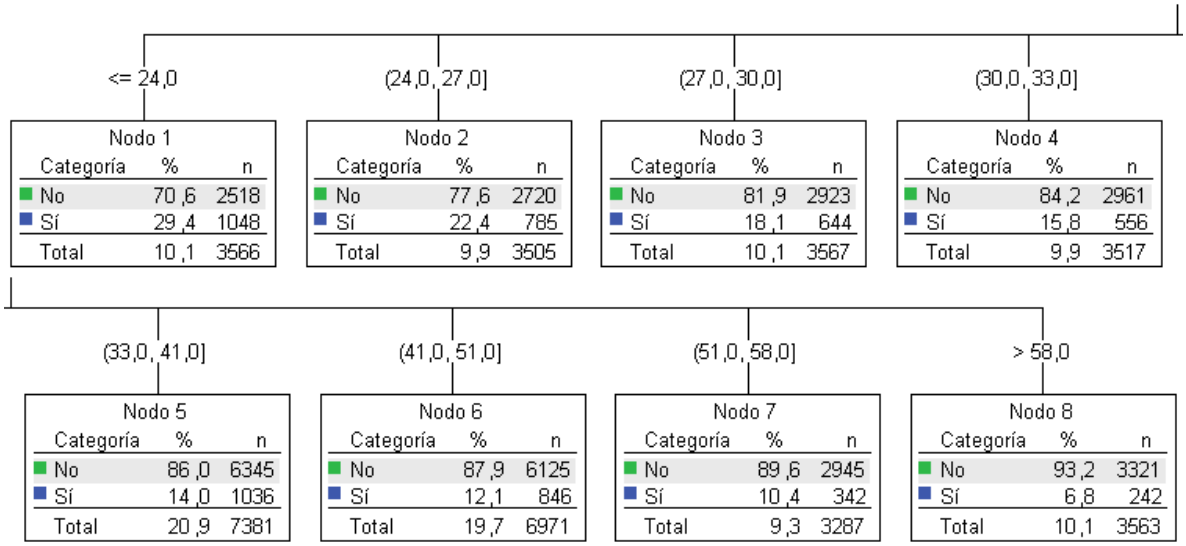


Figura 5. Segmentación de la variable Edad del Cliente

La edad del cliente es indiferente frente al nivel de incumplimiento observado, este es el planteamiento de la hipótesis nula; por su parte la hipótesis alterna señala que a menor edad de las personas, más se tiende a incumplir en el pago de los créditos.

Los nodos de la figura 5 ordenan la tasa de malos para diferentes rangos de edad; los peores riesgos crediticios con tasa de incumplimiento del 29.4% corresponden a menores de 24 años, mientras que la tasa de malos más baja 6.8% se aprecia en los mayores de 58 años.

Estado civil de los clientes

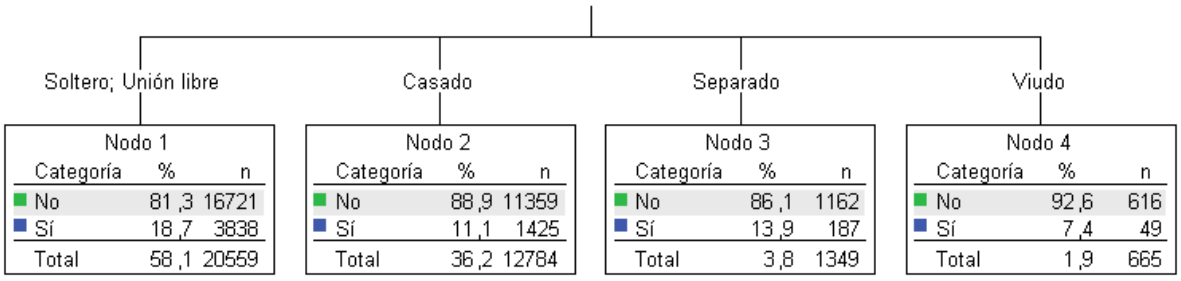


Figura 6. Segmentación de la variable Estado Civil

En la figura 6 se aprecia el mayor incumplimiento en los clientes solteros o que viven en unión libre; en segundo lugar se encuentran los separados; en el tercer puesto los casados, y finalmente, en los viudos se apreció la menor tasa de *default*.

Nivel de estudios

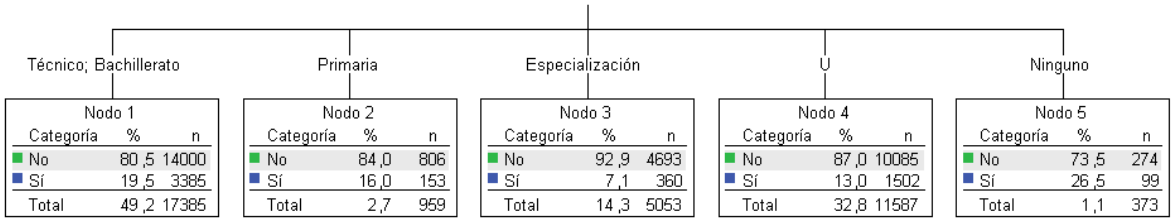


Figura 7. Segmentación de la variable Nivel de Estudios

Esta variable, al igual que las anteriores, es muy probable que sea aceptada en la regresión rechazando la hipótesis nula, dado que hay una gran diferencia en la tasa de malos según el nivel educativo; la figura 7 muestra en los nodos extremos tasa del 26.5% para los clientes sin ningún estudio y tasa del 7.1% para los clientes con especialización.

Actividad económica del solicitante

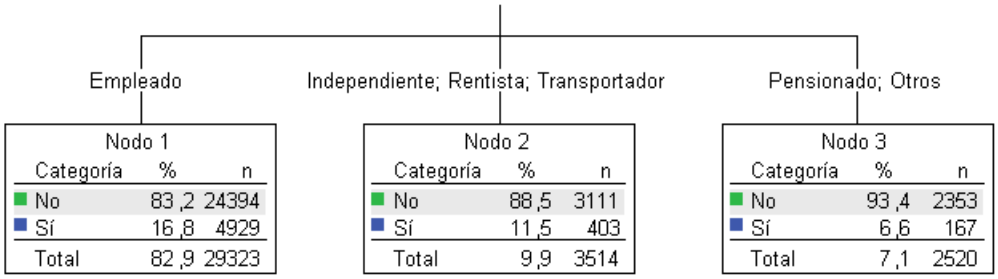


Figura 8. Segmentación de la variable Actividad Económica

En la figura 8 se señala que el 16.8% de los clientes asalariados cayó en *default,* mientras que en los independientes incluyendo rentistas de capital y

transportadores el incumplimiento baja al 11.5%, y en pensionados se observa la tasa de malos más baja, tan solo 6.6%; lo que sugiere que se aceptará la hipótesis alterna ya que la actividad económica sí es reveladora del probable incumplimiento.

Tipo de contrato

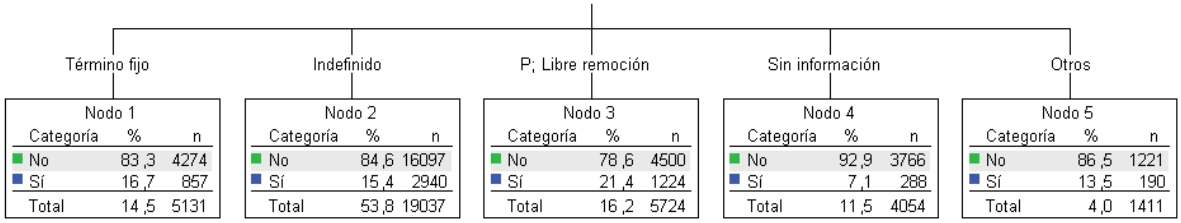


Figura 9. Segmentación de la variable Tipo de Contrato

Las categorías de la variable tipo de contrato se clasifican en 5 nodos que diferencian la tasa de incumplimiento; el nodo de peor perfil en la figura 9 tiene tasa de malos de 21.4% y hace referencia a los contratos de provisionalidad y libre remoción, mientras que los contratos a término indefinido incumplen en 15.4%.

Tipo de vivienda

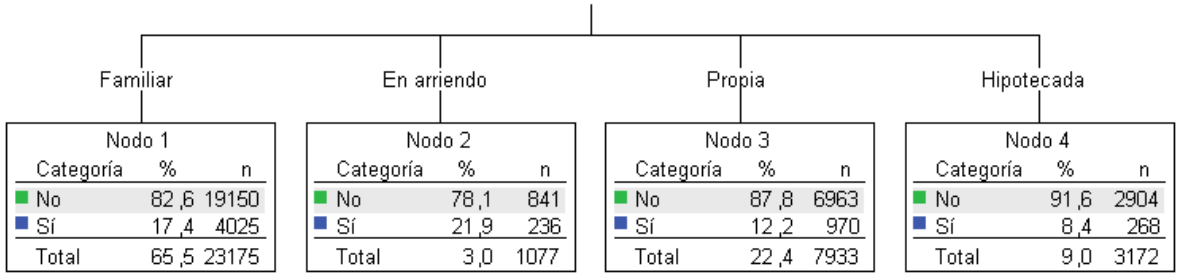


Figura 10. Segmentación de la variable Tipo de Vivienda

Los clientes que afirman vivir en arriendo son los de mayor tendencia a incumplir, 21.9% según la figura 10, los que viven con sus padres (vivienda familiar) están en segundo lugar con un 17.4%. Con mejor perfil y una tasa del 12.2% están los que tienen casa propia y finalmente los que tienen su vivienda hipotecada en el mejor puesto con el 8.4% de *default*.

Número de personas a cargo

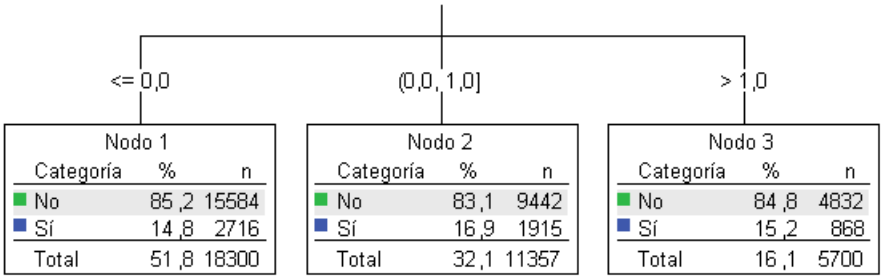


Figura 11. Segmentación de la variable Número de Personas a Cargo

Esta variable presenta poca discriminación, la figura 11 muestra que las tasas de malos son relativamente cercanas, entre el 14.8% y el 16.9%; sin embargo su ordenamiento es correcto al sugerir menores incumplimientos para el grupo cuyos solicitantes no tienen personas económicamente dependientes.

Región del país

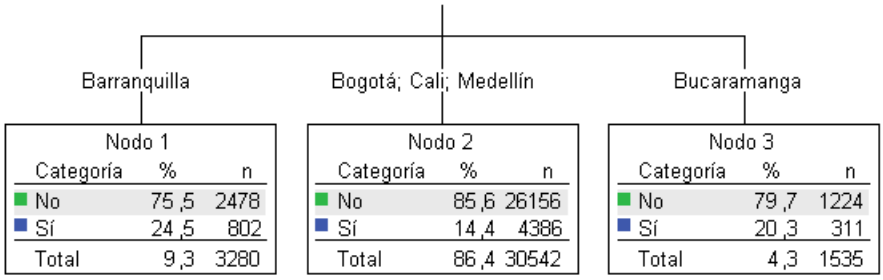


Figura 12. Segmentación de la variable Región

La figura 12 presenta la segmentación por las cinco regionales del país; en ella se aprecia que el 86.4% de la muestra se concentra en las regionales correspondientes a Bogotá, Medellín y Cali, y su nivel de incumplimiento (14.4%) es levemente inferior al porcentaje global (15.6%). Por su parte, el 4.3% que conforma la región Bucaramanga tiene tasa de malos del 20.3%; y finalmente, el restante 9.3% de Barranquilla tiene 24.5% como la peor tasa de *default*.

Indicador de cuenta de nómina con el Banco

	Casos	Tasa de malos
No	24,772	15.7%
Sí	10,585	15.1%
Total general	35,357	15.6%

Tabla 5. Segmentación de la variable Indicador de Nómina

El software no genera árbol de decisión como consecuencia de la diferenciación casi nula que la variable aporta a la tasa de malos, en la tabla 5 se aprecia 15.7% para los que no tienen nómina con el Banco y 15.1% para los que sí la tienen.

Tipo de cupo solicitado

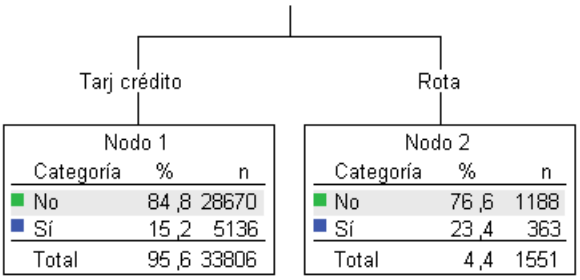


Figura 13. Segmentación de la variable Tipo de Cupo

En la figura 13 se aprecia que los cupos de crédito para utilizar mediante tarjeta de crédito, presentan menor nivel de incumplimiento que aquellos cupos rotativos que se usan mediante tarjeta débito o directamente en oficinas del Banco.

Canal por el cual se tramita la solicitud

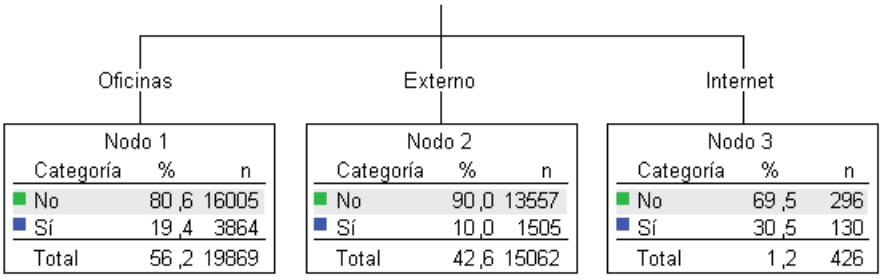


Figura 14. Segmentación de la variable Canal

En la figura 14 se muestra que las solicitudes presentadas por la fuerza externa de vendedores corresponden al 42.6% de la muestra y presentan 10.0% como tasa de incumplimiento; por su parte los clientes que solicitan el crédito en oficinas son un 56.2% del total y tienen incumplimiento del 19.4%. Finalmente, el 1.2% de la muestra corresponde a clientes que hacen la solicitud a través de internet y el 30.5% de ellos llegan a *default.*

Ingreso en SMLMV (salarios mínimos legales mensuales vigentes)

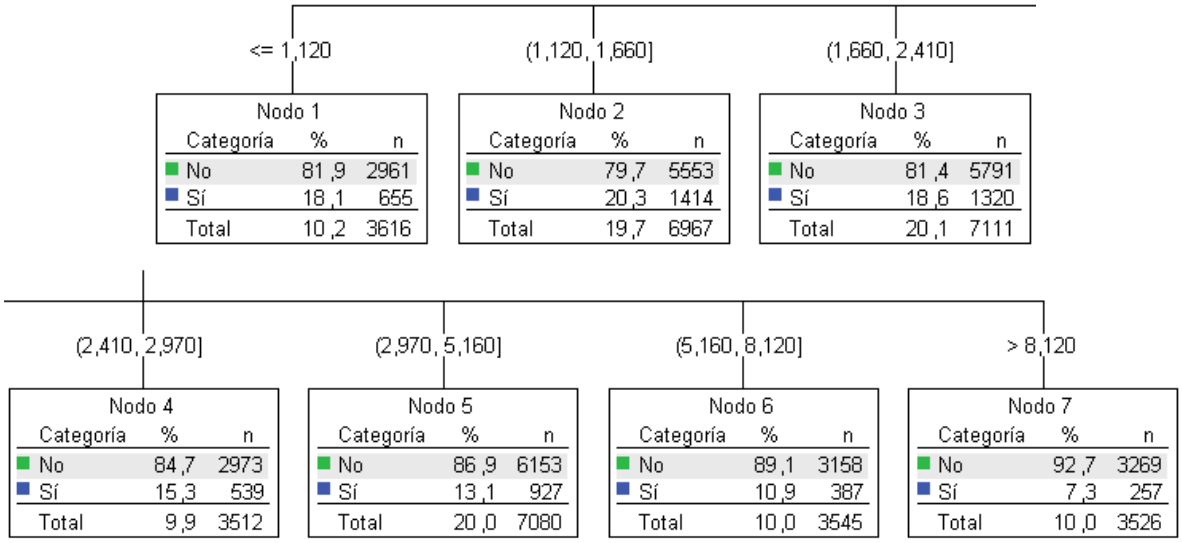


Figura 15. Segmentación de la variable Ingreso en SMLMV

En la figura 15 se aprecia que las tres ramas ubicadas en la parte superior del gráfico, incluyen los clientes con ingresos hasta de 2.4 SMLMV y presentan los incumplimientos más altos entre 18.1% y 20.3%; de otro lado, las ramas ubicadas en la parte inferior del gráfico muestran que la tasa de malos va disminuyendo de manera importante hasta llegar a 7.3% en los clientes de ingreso superior a 8.1 SMLMV.

Monto solicitado en SMLMV

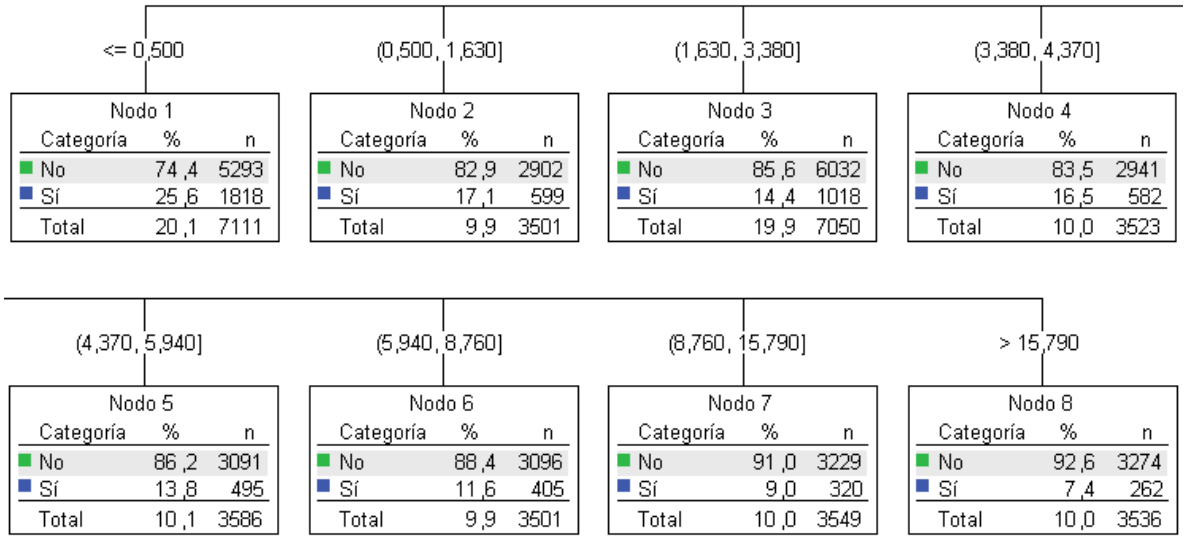


Figura 16. Segmentación de la variable Monto Solicitado en SMLMV

Esta variable, en general presenta ordenamiento lógico en la tasa de malos, a mayor cupo solicitado se aprecia menor nivel de incumplimiento; se afirma que es un comportamiento lógico ya que el mayor cupo lo solicitan quienes tienen mayor ingreso y por ende menor probabilidad de incumplir.

Número de consultas recientes en la central de riesgo

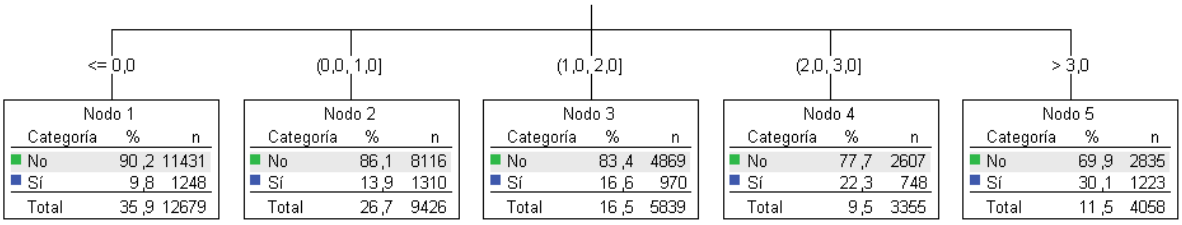


Figura 17. Segmentación de la variable Número de Consultas en la Central de Riesgo

La hipótesis alterna sobre esta variable señala que si un cliente está siendo muy consultado recientemente (últimos seis meses) en la central de riesgo, tiene mayor probabilidad de incumplir que uno que no está siendo consultado. El árbol sugiere que esta hipótesis alterna se aceptará dado que los clientes sin consultas

recientes presentan tasa de *default* del 9.8%, los que tienen una consulta incumplen en un 13.9%, los de dos consultas lo hacen en un 16.6%, los de 3 en un 22.3% y los que tienen más de tres consultas incumplen en el 30.1% de los casos (ver figura 17).

Uso de los rotativos

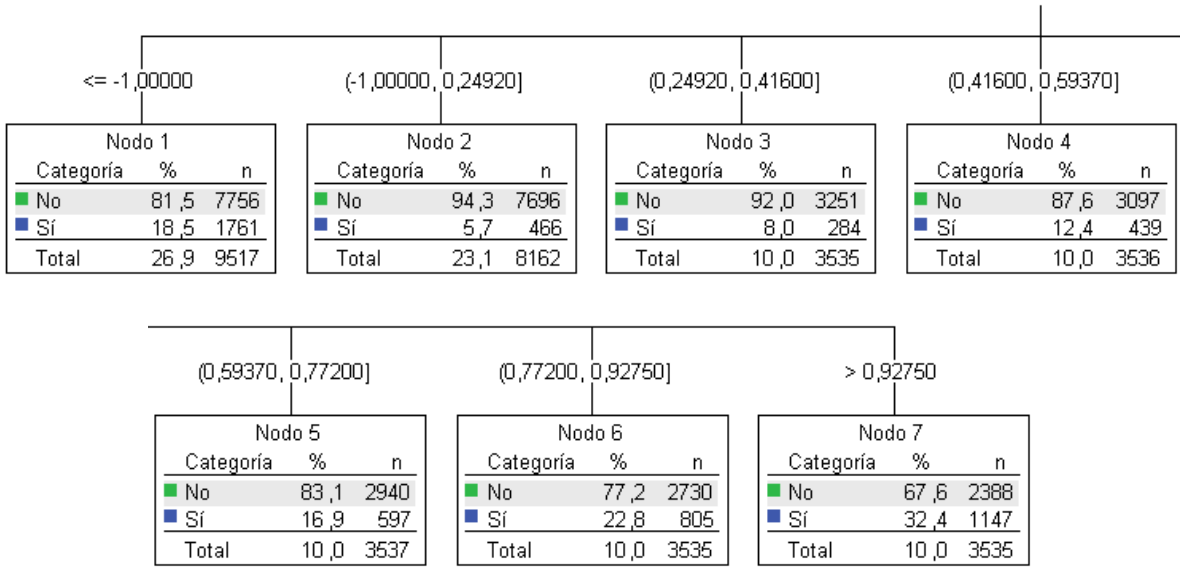


Figura 18. Segmentación de la variable Uso de los Rotativos

Esta variable se calcula dividiendo el saldo utilizado en créditos rotativos sobre el respectivo cupo de crédito; es la variable más utilizada en los modelos para riesgo de crédito dada su gran capacidad de discriminación. En la figura 18 se corrobora lo afirmado ya que el árbol de decisión muestra la mayor distancia en tasa de *default* de todas las variables, va desde el 5.7% en aquellos clientes que usan sus cupos en menos del 25%, hasta incumplimientos de 32.4% en los que su deuda representa más del 93% de sus cupos.

El nodo 1 hace referencia al 26.9% de clientes que aunque sí tienen experiencia crediticia, no tienen cupos rotativos aprobados; su tasa de malos es del 18.5%, un poco más alta que el porcentaje global.

Antigüedad en meses del crédito más reciente

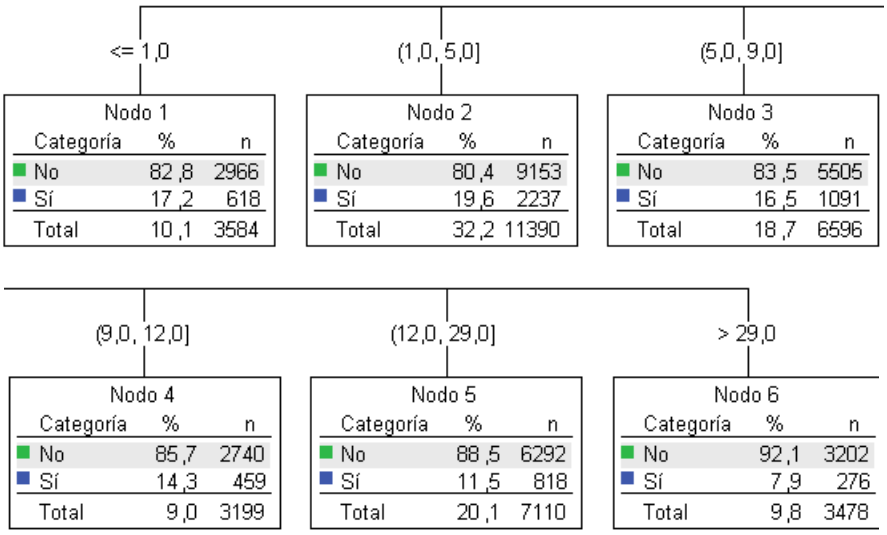
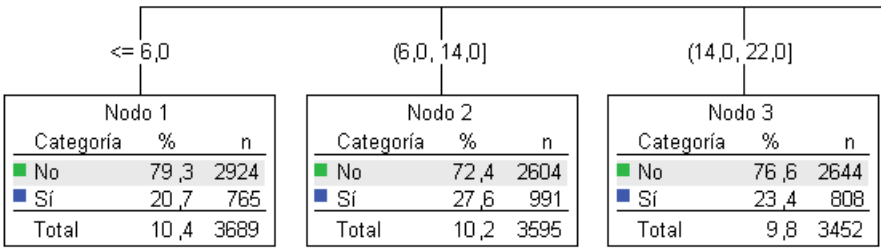


Figura 19. Segmentación de la variable Antigüedad del Crédito más Reciente

La segmentación de esta variable permite afirmar que los clientes con créditos otorgados recientemente tienden a incumplir más sus obligaciones dado que incrementan su endeudamiento, mientras que aquellos que llevan mucho tiempo sin tramitar nuevos créditos presentan menor tasa de *default*.

Antigüedad en meses del crédito más antiguo



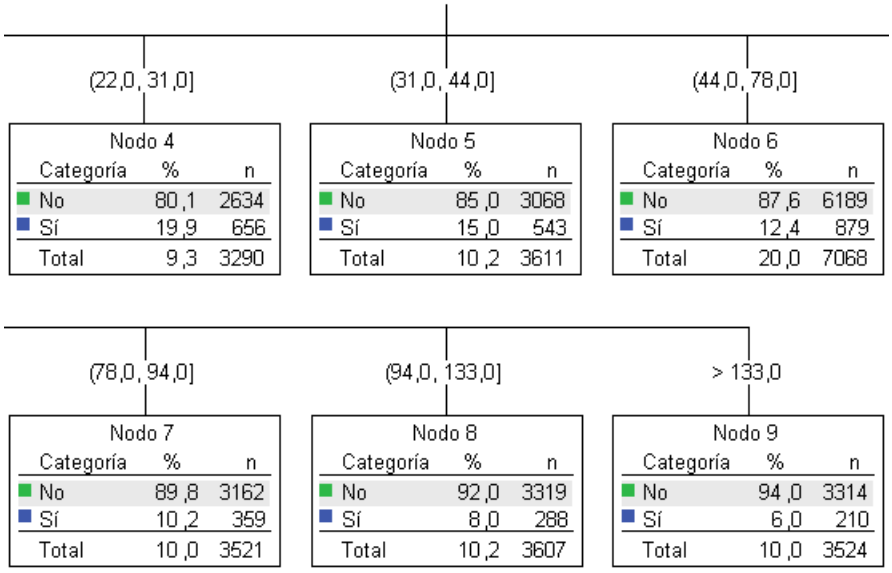


Figura 20. Segmentación de la variable Antigüedad del Crédito más Antiguo

Bajo similar concepto al de la variable anterior, entre más antigüedad tenga un cliente en el sector financiero mejor será su comportamiento crediticio, así en la figura 20 se observa que los clientes cuyo crédito vigente más antiguo es de más de 133 meses solo incumplen en un 6.0%, mientras que aquellos cuyo crédito más antiguo es de menos de 22 meses incumplen entre un 20.7% y un 27.6%.

Antigüedad en meses de la tarjeta de crédito más reciente

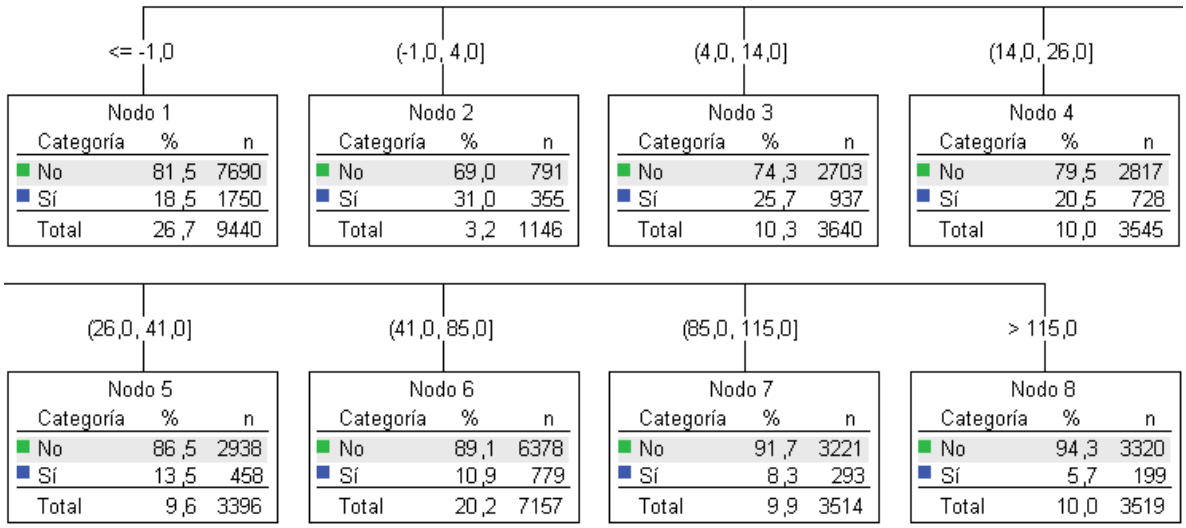


Figura 21. Segmentación variable Antigüedad de la Tarjeta de Crédito más Reciente

Como ya se había comentado, los clientes que no tienen créditos rotativos aprobados (nodo 1 de la figura 21) presentan tasa de incumplimiento del 18.5%, un poco superior al 15.6% de porcentaje global.

Adicionalmente, la figura 21 muestra ordenamiento en las tasas de *default* según la antigüedad de su tarjeta de crédito más reciente; revisando los casos extremos, los clientes que tienen tarjetas de crédito hace aproximadamente 10 años o más (más de 115 meses) son los que menos incumplen con el 5.7%, mientras que aquellos con tarjetas recientemente otorgadas, menos de 5 meses para citar el caso del nodo 2, incumplen en un 31.0%.

Mora máxima en los 12 meses previos

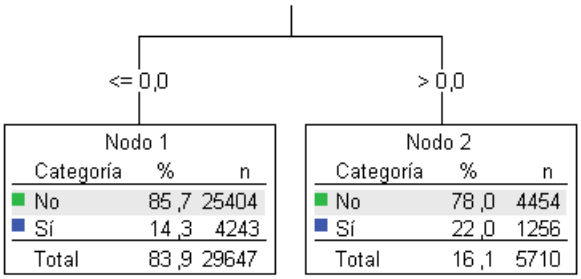


Figura 22. Segmentación variable Mora Máxima en los 12 meses previos

Para esta variable la hipótesis nula afirma que el incumplimiento no se ve afectado por la morosidad que haya presentado el cliente en los anteriores doce meses al estudio; la hipótesis alterna por su parte afirma que sí es relevante, que un cliente con morosidades previas es más propenso a entrar en incumplimiento.

El árbol de decisión de la figura 22 segmentó dos particiones, en la primera, con un 83.9% de clientes que no tuvieron mora durante los 12 meses previos a la solicitud, se observó un incumplimiento del 14.3%; mientras que en el 16.1% de clientes que sí tuvieron moras, la tasa de *default* es del 22.0%; muy probablemente el modelo tendrá evidencia suficiente para rechazar la hipótesis nula aceptando la significancia de la variable.

4.4. FORMULACIÓN DEL MODELO

4.4.1 REGRESIÓN

Teniendo en cuenta los objetivos del presente trabajo, se presentan a continuación los resultados relacionados con la formulación de un modelo de regresión logística que permita estimar la probabilidad de incumplimiento.

Con el fin de asegurar la elección del modelo adecuado se ejecutan los diferentes métodos de regresión logística ofrecidos por el software. Los resultados obtenidos se evalúan con la ecuación 5, criterio de Información de Akaike, que permite medir la calidad de un modelo haciendo un balance entre su máxima verosimilitud y la cantidad de variables que requiere para lograrlo (Akaike, 1974).

$$AIC = 2 * (número\ de\ variables + 1) - 2 * Ln(Verosimilutud)$$

Ecuación 5. Criterio de información de Akaike

Como se aprecia en la tabla 6, el único método de regresión que genera un resultado levemente diferente es el denominado Intro, que se basa en la inclusión simultánea de todas las variables; su resultado se califica de diferente ya que es el único con valor AIC desviado, 41.6 frente a 37.6 obtenido en los demás modelos.

Los demás métodos, denominados por pasos, generan resultados similares entre sí; en particular los métodos hacia adelante se identifican como los de mejor desempeño si se evalúa el indicador AIC a nivel del cuarto decimal, 30.6066 frente a 30.6068 del método Wald hacia atrás y y 30.6071 de los métodos LR y condicional hacia atrás.

Método de selección de variables	Número de variables aceptadas	Máxima verosimilud	LN máxima verosimilud	AIC (criterio de información de Akaike)
Intro. Todas las variables se introducen en un solo paso.	30	26797.211	10.19605309	41.60789381
Hacia adelante (Condicional). Selección de avance por pasos basada en estimaciones condicionales de los parámetros.	28	26815.071	10.19671936	37.60656128
Hacia atrás (LR). Se retrocede por pasos basando la decisión en la razón de verosimilitud.	28	26808.118	10.19646003	37.60707994
Hacia adelante (Wald). Selección de avance por pasos basada en el estadístico de Wald.	28	26815.071	10.19671936	37.60656128
Hacia atrás (Condicional). Se retrocede por pasos basando la decisión en estimaciones condicionales de los parámetros.	28	26808.118	10.19646003	37.60707994
Hacia atrás (Wald). Se retrocede por pasos basando la decisión en el estadístico de Wald.	28	26811.776	10.19659647	37.60680705

Tabla 6. Comparativo de criterio AIC entre modelos

A partir del criterio AIC se selecciona como modelo a implementar el del método por pasos hacia adelante basado en el estadístico de Wald.

La tabla 7 presenta las variables aceptadas para ser incluidas en el modelo, sus coeficientes, errores típicos, estadístico de Wald y los p-valor obtenidos al contrastar la prueba de hipótesis de significancia.

	Beta	Err estándar	Wald (Beta /ErrEst)^2	Prob (Wald^(1/2))	p-valor
Antigüedad con el Banco en años	(0.025319)	0.004759	28.310243	100.00%	0.00%
Edad	(0.013413)	0.001732	59.963989	100.00%	0.00%
Personas a cargo	0.111063	0.020022	30.769992	100.00%	0.00%
Ingreso en SMLMV	(0.018965)	0.005300	12.803904	99.98%	0.03%
Valor solicitado en SMLMV	(0.005876)	0.002300	6.525336	99.47%	1.06%
Número de consultas	0.166663	0.008191	413.960256	100.00%	0.00%
Uso rotativos	0.309570	0.022559	188.303487	100.00%	0.00%
Antigüedad en meses del crédito más reciente	(0.006734)	0.001287	27.369988	100.00%	0.00%
Antigüedad en meses de la TC más antigua	(0.004761)	0.000472	101.926438	100.00%	0.00%
Mora Máxima en los 12 meses previos	0.509445	0.024488	432.813021	100.00%	0.00%
OrigenDum_3 = Solicitud	0.322465	0.038208	71.228929	100.00%	0.00%
EstCivilDum_2 = Separado	0.267942	0.089435	8.975726	99.86%	0.27%
EstCivilDum_3 = Soltero	0.108133	0.044167	5.993973	99.28%	1.44%
EstCivilDum_4 = Unión libre	0.151770	0.045350	11.200206	99.96%	0.08%
NivelEstudioDum_2 = Especialización	(0.873374)	0.068137	164.300085	100.00%	0.00%
NivelEstudioDum_5 = Técnico	(0.242598)	0.040705	35.521042	100.00%	0.00%
NivelEstudioDum_6 = Universitario	(0.494238)	0.042335	136.293084	100.00%	0.00%
TipContratoDum_4 = Provisional	0.364786	0.046123	62.551854	100.00%	0.00%
TipViviendaDum_3 = Hipotecada	(0.499034)	0.070716	49.799936	100.00%	0.00%
TipViviendaDum_4 = Propia	(0.159756)	0.043777	13.317770	99.99%	0.03%
RegionDum_2 = Bogotá	(0.282806)	0.051402	30.270160	100.00%	0.00%
RegionDum_3 = Bucaramanga	(0.223898)	0.080751	7.687810	99.72%	0.56%
RegionDum_4 = Cali	(0.383798)	0.059575	41.503063	100.00%	0.00%
RegionDum_5 = Medellín	(0.445411)	0.066003	45.540014	100.00%	0.00%
TipCreditoDum_2 = Tarjeta Crédito	(0.353810)	0.069048	26.256897	100.00%	0.00%
CanalDum_2 = Internet	0.308533	0.123200	6.271645	99.39%	1.23%
CanalDum_3 = Oficinas	0.213662	0.041864	26.048312	100.00%	0.00%
Constante	(0.720125)	0.123616	33.936254	100.00%	0.00%

Tabla 7. Coeficientes del modelo

Los signos de los betas (valores negativos entre paréntesis) señalan la dirección de dependencia entre la variable y el default, los betas positivos señalan relación directa, por ejemplo a mayor uso de los créditos rotativos, mayor es la probabilidad de caer en *default*, mientras que coeficientes negativos señalan relación inversa, por ejemplo quienes tienen mayor edad tienden a presentar menor incumplimiento.

En las variables dicotómicas se presenta similar interpretación, un coeficiente con signo positivo, como el caso de las solicitudes tramitadas por internet señala que por ser de ese canal, la probabilidad de incumplir se incrementará; mientras que un coeficiente negativo, como el de clientes de Medellín, por estar en esa región, la probabilidad de incumplir será menor.

4.4.2 ANÁLISIS DE CORRELACIONES

		1	2	4	5	7	8	9	10
Antigüedad con el Banco en años	1	100%	-15%	1%	1%	0%	0%	-4%	-1%
Edad	2	-15%	100%	-8%	-4%	5%	-2%	-24%	-3%
Ingreso en SMLMV	4	1%	-8%	100%	-35%	1%	4%	-10%	-7%
Valor solicitado en SMLMV	5	1%	-4%	-35%	100%	5%	1%	-13%	3%
Porcentaje de uso en los rotativos	7	0%	5%	1%	5%	100%	6%	-38%	0%
Antigüedad en meses del Crédito más Reciente	8	0%	-2%	4%	1%	6%	100%	-14%	-1%
Antigüedad en meses de la Tarjeta de Crédito más Antigua	9	-4%	-24%	-10%	-13%	-38%	-14%	100%	-5%
Mora Máxima en los 12 meses previos	10	-1%	-3%	-7%	3%	0%	-1%	-5%	100%

Tabla 8. Análisis de Correlaciones

En la tabla 8 se aprecia en general, bajas correlaciones entre las variables explicativas; esto descarta problemas por colinealidad entre las variables, es decir la posibilidad de que una variable independiente sea explicada por una o varias variables independientes.

Adicionalmente, los valores obtenidos en los errores estándar de los coeficientes estimados lucen bajos, en promedio 4% sobre el valor de los coeficientes estimados. No obstante que no existe un umbral definido para esta evaluación, sí es señalado que los problemas de colinealidad en las regresiones logísticas se aprecian fácilmente al observar valores muy grandes en los errores estándar de los coeficientes estimados (Hosmer & Lemeshow, 2013).

4.5. PRUEBAS DE DESEMPEÑO

4.5.1 CAPACIDAD DE DISCRIMNACIÓN

La tabla de desempeño sobre la partición de entrenamiento (ver tabla 9) segmenta la muestra en deciles (diez grupos de tamaño similar) agrupados según la probabilidad de incumplimiento estimada por el modelo para cada registro.

En el peor decil, en el que se agrupan casos con probabilidades de incumplimiento mayores al 31.7% el modelo estima en promedio un 44.0%, se aprecia efectivamente el peor nivel de incumplimiento observado, 45.2%. De otro lado, en el decil donde el modelo predice menor probabilidad de incumplir, la realidad muestra 3.2% como menor nivel de *default*.

Grupo PI	Número de casos	Buenos	Malos (default)	PI promedio	IO (incumplimiento observado)	Acumulado de Buenos	Acumulado de Malos	KS (máxima distancia)
>0.317	3,536	1,936	1,600	44.0%	45.2%	6%	29%	22.6%
>0.2322 y <=0.317	3,533	2,579	954	26.9%	27.0%	15%	46%	31.3%
>0.1829 y <=0.2322	3,537	2,793	744	20.6%	21.0%	24%	60%	35.5%
>0.1482 y <=0.1829	3,535	2,970	565	16.5%	16.0%	34%	70%	**35.8%**
>0.1216 y <=0.1482	3,537	3,060	477	13.5%	13.5%	45%	79%	34.3%
>0.0992 y <=0.1216	3,533	3,161	372	11.0%	10.5%	55%	86%	30.4%
>0.0792 y <=0.0992	3,536	3,241	295	8.9%	8.3%	66%	91%	24.9%
>0.0595 y <=0.0792	3,537	3,316	221	6.9%	6.2%	77%	95%	17.9%
>0.0384 y <=0.0595	3,536	3,377	159	4.9%	4.5%	89%	98%	9.4%
>0 y <=0.0384	3,537	3,425	112	2.3%	3.2%	100%	100%	0.0%
Totales	35,357	29,858	5,499	15.6%	15.6%			

Tabla 9. Tabla de desempeño partición de entrenamiento

La prueba de KS y el coeficiente GINI ratifican la capacidad de discriminación que tiene el modelo sobre la población analizada. La curva gruesa de la figura 23 muestra como el modelo acumula la mayor proporción de malos en los deciles con mayor probabilidad de incumplir mientras que en los deciles de mejor pronóstico agrega muy pocos; por el contrario en la curva delgada se acumula menos proporción de buenos en el decil de PI mayores al 31.7% y la mayor proporción en el decil de PI menores o iguales a 3.84%.

El indicador KS se establece como la mayor distancia entre las dos curvas de ganancia (gruesa acumula los que caen en *default* y delgada acumula los que no caen en esa condición); en este caso la mayor distancia logra ser un 35.8% apreciado en el decil con PI superiores a 14.82% y hasta de 18.29%. Es decir que el modelo se cataloga con buen desempeño según criterio de los líderes internacionales en desarrollo de este tipo de modelos.

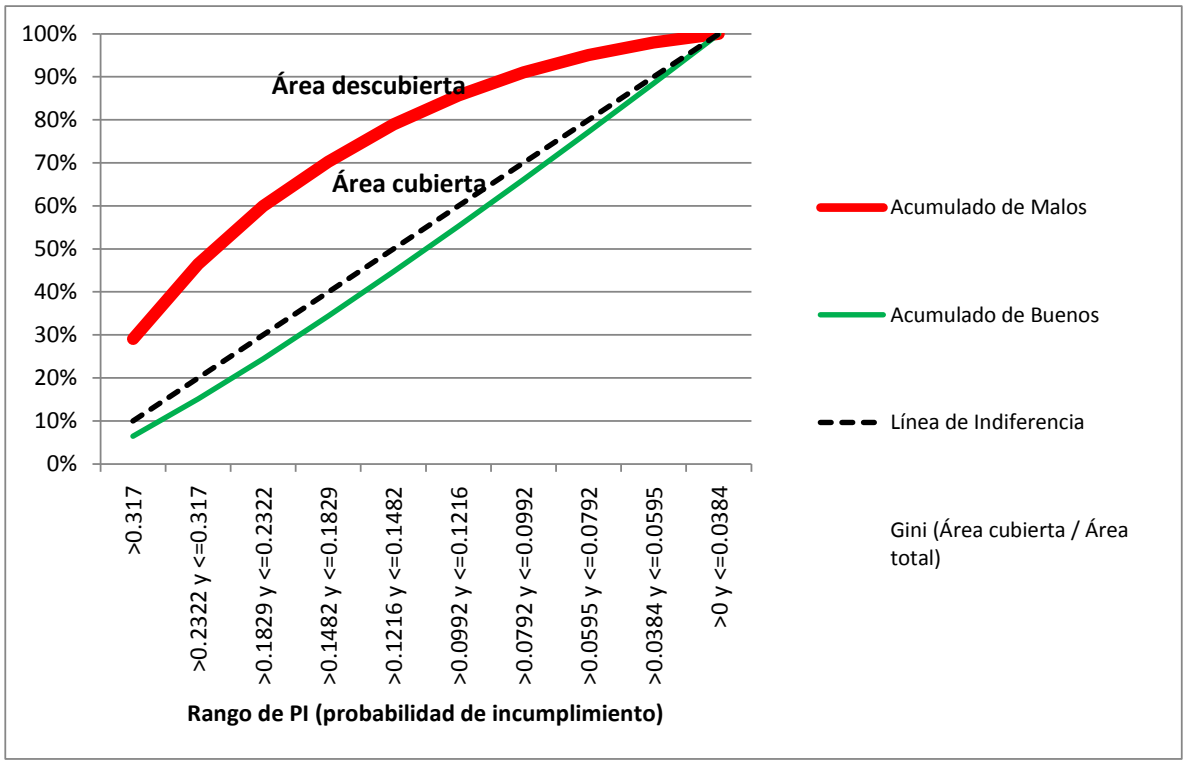


Figura 23. Curvas de ganancia partición de entrenamiento

El coeficiente de GINI por su parte cataloga el modelo como muy bueno según práctica de los líderes internacionales en desarrollo de modelos, ya que el área entre la curva de indiferencia (punteada) y la curva gruesa (acumulado de malos) cubre un 45% del área de incertidumbre total, que se ubica por encima de la línea de indiferencia (punteada) de la figura 23.

4.5.2 BONDAD DE AJUSTE

El siguiente gráfico (figura 24) deja evidencia del adecuado desempeño del modelo en la partición de entrenamiento. Los incumplimientos observados están correctamente ordenados en función de la tasa de malos esperada según el modelo y de hecho los valores observados son muy similares a los esperados.

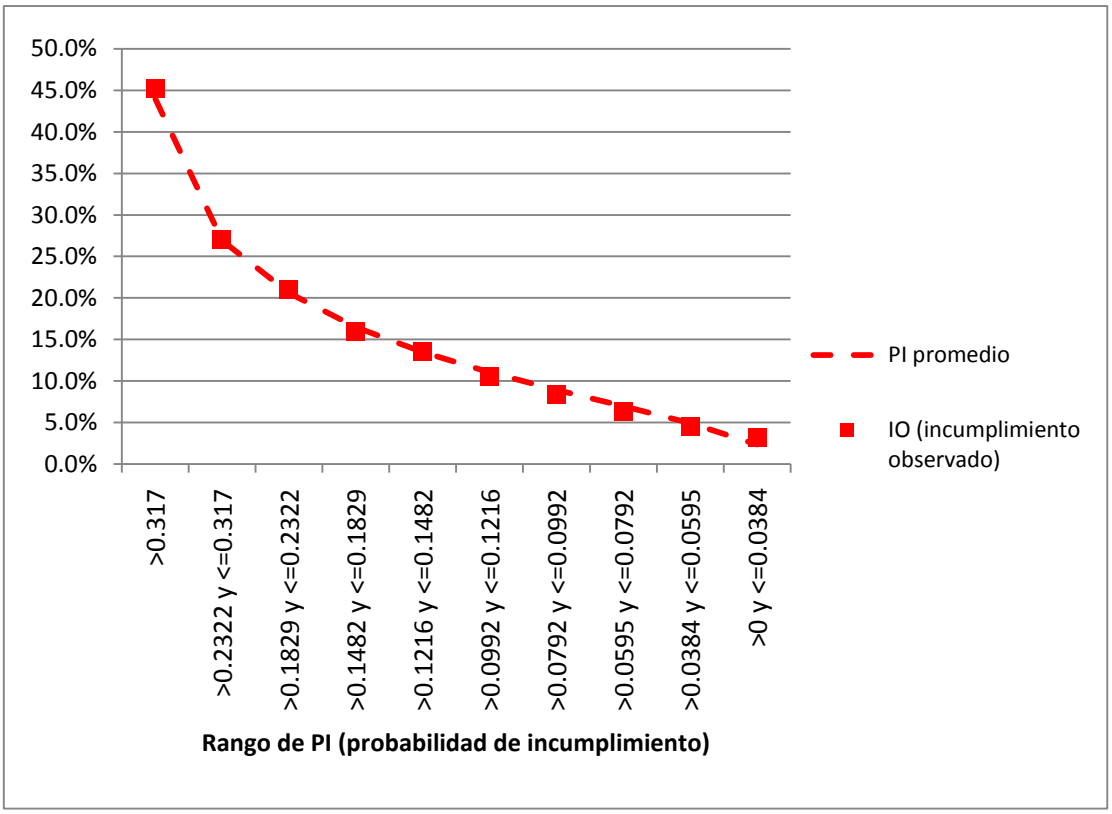


Figura 24. Back testing partición de entrenamiento

Como complemento a este análisis gráfico aplicaremos el test de Hosmer-Lemeshow que establece si las diferencias entre la distribución de valores esperados y la de valores observados se ajustan o no dependiendo de una distribución χ^2_{k-2}, como se vio en la ecuación 4 del marco teórico.

La aplicación de este test para muestras de gran tamaño es cuestionable dada la dificultad para acercar la distribución verdadera a la chi cuadrado (Canavos, 1988); en este ejercicio efectivamente vemos que el estadístico genera un valor crítico de 20.13 dejando un p-valor de tan solo 1% y dado que la prueba exige al menos un 5% se estaría rechazando el ajuste del modelo.

Teniendo en cuenta que este ejercicio realmente se hace sobre un conjunto muy grande de registros y dado que el análisis gráfico practicado sobre la figura 24 muestra un nivel de ajuste muy bueno, se explora la aplicación de la prueba sobre valores transformados a una escala de mil registros, pero conservando exactamente las mismas probabilidades de incumplimiento y las mismas tasas de incumplimiento observado en cada decil de la tabla de desempeño.

Grupo PI	Número de casos	Buenos	Malos (default)	PI promedio	IO (incumplimiento observado)	Test H/L	Test H/L ajustado
>0.317	100	55	45	44.0%	45.2%	2.3423	0.0662
>0.2322 y <=0.317	100	73	27	26.9%	27.0%	0.0130	0.0004
>0.1829 y <=0.2322	100	79	21	20.6%	21.0%	0.4673	0.0132
>0.1482 y <=0.1829	100	84	16	16.5%	16.0%	0.5961	0.0169
>0.1216 y <=0.1482	100	87	13	13.5%	13.5%	0.0027	0.0001
>0.0992 y <=0.1216	100	89	11	11.0%	10.5%	0.9189	0.0260
>0.0792 y <=0.0992	100	92	8	8.9%	8.3%	1.3814	0.0391
>0.0595 y <=0.0792	100	94	6	6.9%	6.2%	2.6231	0.0742
>0.0384 y <=0.0595	100	96	4	4.9%	4.5%	1.4481	0.0410
>0 y <=0.0384	100	97	3	2.3%	3.2%	10.4595	0.2958
Totales	1,000	844	156	15.6%	15.6%	20.2523	0.5728
					p-valor	0.94%	99.98%

Tabla 10. Tabla de desempeño ajustada (1000 casos) partición de entrenamiento

La tabla 10 muestra que las tasas de incumplimiento observado se ajustan muy bien a las probabilidades de incumplimiento, el estadístico ajustado a 1.000 casos arroja un p-valor de 99.98% superando la prueba de bondad de ajuste.

4.5.3 DESEMPEÑO SOBRE LA PARTICIÓN DE PRUEBA

Como se señaló en la sección 4.1 y se precisó en la tabla 2, la muestra seleccionada se segmentó en dos partes mediante muestreo aleatorio simple con el fin de tener una partición de prueba con casi idénticas condiciones a la partición de datos usada para el desarrollo o entrenamiento del modelo.

La tabla 11 presenta los registros de la partición de prueba segmentados en deciles según la probabilidad de incumplimiento estimada por el modelo para cada registro.

Los resultados son muy similares a los obtenidos en la partición de entrenamiento; en el decil con mayor probabilidad de incumplimiento, se estima en promedio 43.6% y se aprecia incumplimiento observado de 42.8%; mientras que en el decil de menor probabilidad de incumplir, se aprecian estimaciones y observaciones de incumplimiento cercanas al 3%.

Grupo PI	Número de casos	Buenos	Malos (default)	PI promedio	IO (incumplimiento observado)	Acumulado de Buenos	Acumulado de Malos	KS (máxima distancia)
>0.3179	3,524	2,017	1,507	43.6%	42.8%	7%	27%	20.4%
>0.2317 y <=0.3179	3,525	2,524	1,001	27.0%	28.4%	15%	45%	30.0%
>0.1819 y <=0.2317	3,524	2,795	729	20.5%	20.7%	25%	58%	33.7%
>0.1475 y <=0.1819	3,522	2,868	654	16.4%	18.6%	34%	70%	**35.9%**
>0.1213 y <=0.1475	3,526	3,057	469	13.4%	13.3%	45%	79%	34.1%
>0.0994 y <=0.1213	3,524	3,180	344	11.0%	9.8%	55%	85%	29.6%
>0.0793 y <=0.0994	3,524	3,210	314	8.9%	8.9%	66%	91%	24.4%
>0.0596 y <=0.0793	3,527	3,282	245	6.9%	6.9%	77%	95%	17.8%
>0.0379 y <=0.0596	3,523	3,359	164	4.9%	4.7%	89%	98%	9.5%
>0 y <=0.0379	3,527	3,414	113	2.3%	3.2%	100%	100%	0.0%
Totales	35,246	29,706	5,540	15.5%	15.7%			

Tabla 11. Tabla de desempeño partición de prueba

La prueba de KS arrojó 35.9% casi idéntico al resultado de la partición de entrenamiento, se ratifica la calificación de buen desempeño del modelo. El coeficiente GINI (44%), disminuye un punto porcentual frente al observado en la partición de desarrollo (45%). Se ratifica como muy bueno su resultado.

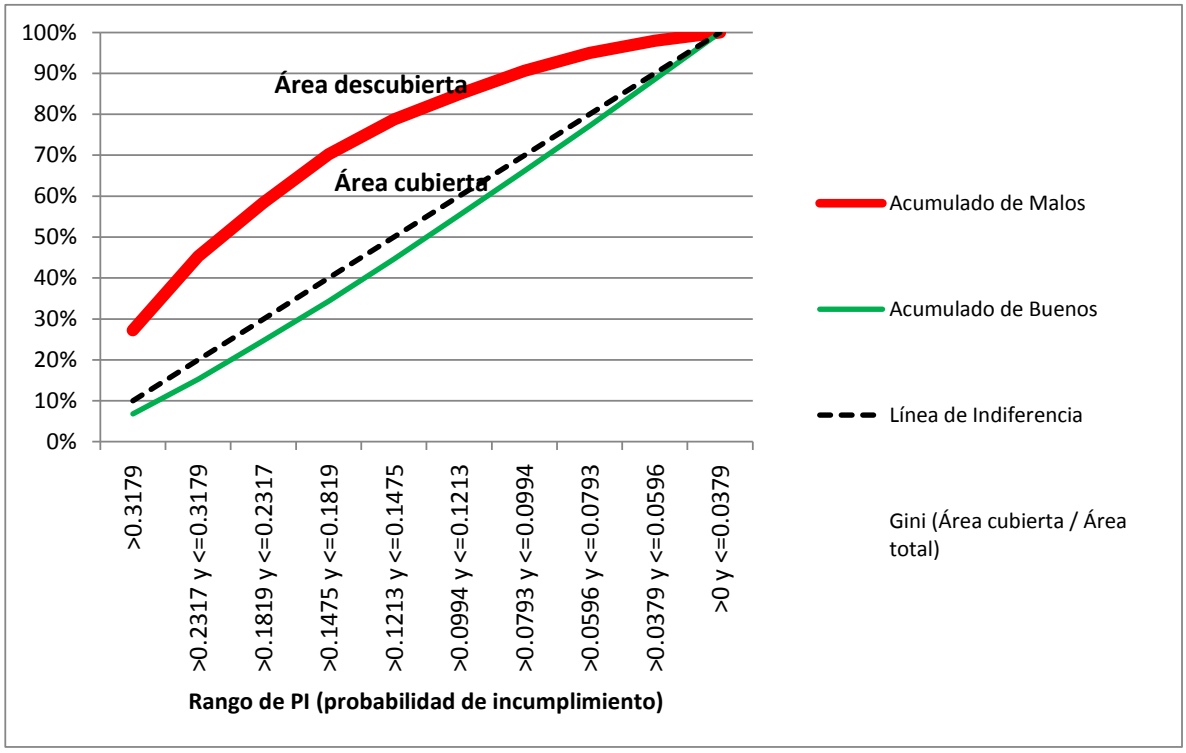


Figura 25. Curvas de ganancia partición de prueba

Las curvas de ganancia de la figura 25 son muy similares a las de la figura 23 (partición de desarrollo), los únicos cambios apreciables son los límites de PI establecidos para cada decil de la segmentación.

La figura 26 confirma el buen ajuste que tiene el modelo. Sobre la partición de prueba los niveles de incumplimiento observado ordenan perfectamente en función de la tasa de malos sugerida por el modelo, y tal como ocurrió en la partición de desarrollo, los incumplimientos observados son muy similares a los esperados.

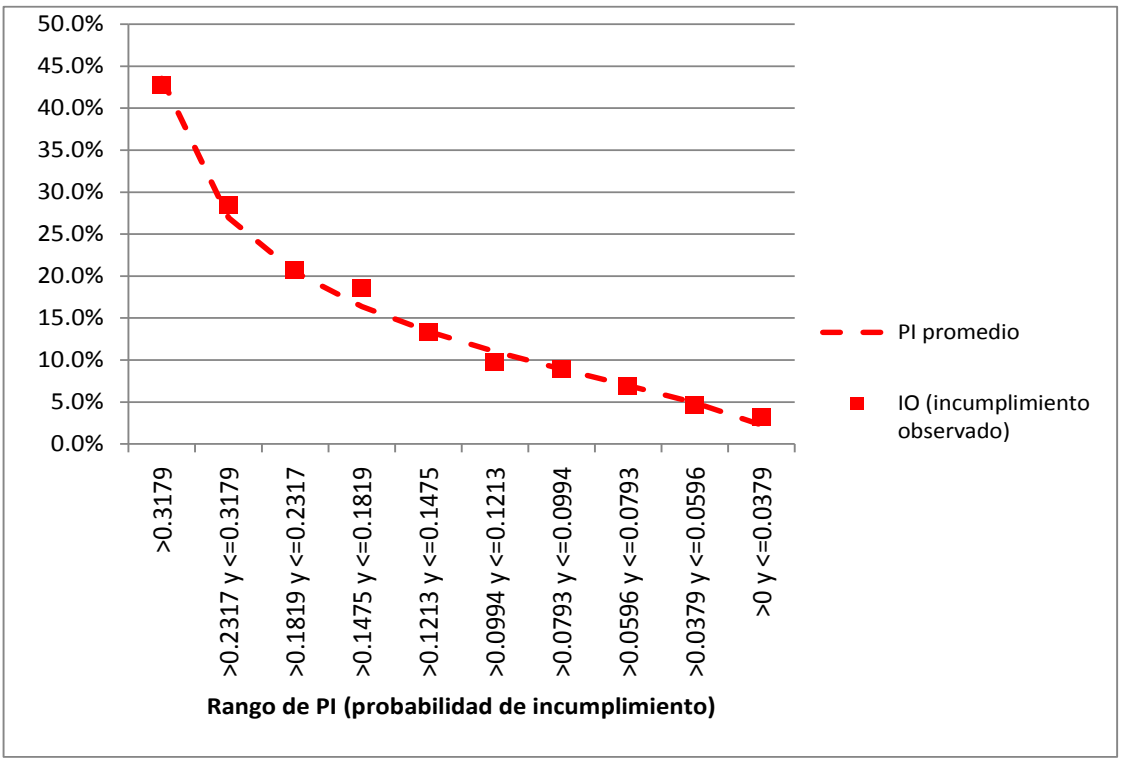


Figura 26. Back testing partición de prueba

El test de Hosmer-Lemeshow aplicado sobre valores transformados a una escala de mil registros, tal como se explicó y aplicó en la sección 4.5.2 sobre la partición de desarrollo, muestra buen ajuste entre las tasas de incumplimiento observado y las probabilidades de incumplimiento, arroja un p-valor de 99.8% (ver tabla 12).

Grupo PI	Número de casos	Buenos	Malos (default)	PI promedio	IO (incumplimiento observado)	Test H/L	Test H/L ajustado
>0.3179	100	57	43	43.6%	42.8%	0.9117	0.0259
>0.2317 y <=0.3179	100	72	28	27.0%	28.4%	3.7435	0.1062
>0.1819 y <=0.2317	100	79	21	20.5%	20.7%	0.0946	0.0027
>0.1475 y <=0.1819	100	81	19	16.4%	18.6%	12.4478	0.3532
>0.1213 y <=0.1475	100	87	13	13.4%	13.3%	0.0268	0.0008
>0.0994 y <=0.1213	100	90	10	11.0%	9.8%	5.6049	0.1590
>0.0793 y <=0.0994	100	91	9	8.9%	8.9%	0.0006	0.0000
>0.0596 y <=0.0793	100	93	7	6.9%	6.9%	0.0000	0.0000
>0.0379 y <=0.0596	100	95	5	4.9%	4.7%	0.4906	0.0139
>0 y <=0.0379	100	97	3	2.3%	3.2%	13.4588	0.3819
Totales	1,000	843	157	15.5%	15.7%	36.7792	1.0435
					p-valor	0.00%	99.80%

Tabla 12. Tabla de desempeño ajustada (1.000 casos) partición de prueba.

5. CONCLUSIONES Y RECOMENDACIONES

Las variables incluidas en este estudio, que describen las características socio demográfica de los clientes y su comportamiento crediticio registrado en la central de riesgo, son reveladoras del perfil de riesgo crediticio de los clientes que solicitan cupo de tarjeta de crédito ante las entidades financieras.

En general, se presentó suficiente evidencia estadística para rechazar la hipótesis nula planteada respecto a que cada una de estas variables no guarda relación con el incumplimiento observado; se aceptó entonces para la mayoría de variables la hipótesis alterna que señaló tener influencia significativa, estadísticamente hablando, en la presencia del *default*.

Del conjunto de modelos ensayados, se estableció como mejor opción el que proporcionó el mejor balance entre menor número de variables y mayor verosimilitud de sus estimadores; esto permite contar con un modelo parsimonioso al tener el menor número de variables posible, con capacidad de discriminación, permitiendo clasificar la muestra estudiada en grupos con diferencias en tasa de *default* desde el 2% hasta el 44%, y finalmente con un muy buen ajuste de las tasas de incumplimiento estimadas frente a las observadas.

El modelo obtenido mostró excelente desempeño tanto en la partición de datos seleccionada para desarrollarlo como en la asignada para prueba; esto permite prever un buen desempeño de la herramienta siempre y cuando se mantenga sobre ella el monitoreo pertinente que asegure estabilidad poblacional mediante pruebas asociadas a homogeneidad de las varianzas, entre otras.

El paso a seguir, no contemplado en el alcance de este trabajo, es fijar el máximo nivel de incumplimiento tolerado (punto de corte) para que aún el cliente de peor perfil admitido contribuya con la rentabilidad mínima requerida.

6. BIBLIOGRAFÍA

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control , 19, 716–723.

Álvarez, Franco, S. I., & Osorio, Betancur, A. (2014). Medición del riesgo crédito Colombia-hacia Basilea III (Doctoral dissertation, Escuela Ingeniería de Antioquia).

Canavos, George (1988). Probabilidad y Estadística, Aplicaciones y Métodos. McGraw Hill. ISBN 968-451-856-0.

Ezequiel, Uriel (2013). Análisis de regresión múltiple con información cualitativa. Universidad de Valencia.

Fernández Castaño, Horacio, Pérez Ramírez, Fredy Ocaris. El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito. Revista Ingenierías Universidad de Medellín 2005. ISSN 1692-3324

Fundación Universitaria Los Libertadores. Guía Metodológica para la Presentación de Trabajos Escritos. Biblioteca Central Hernando Santos Castillo.

Giraldo, Norman (2011). Métodos Estadísticos Aplicados a Finanzas y Gestión de Riesgo. Escuela de Estadística Universidad Nacional de Colombia Medellín.

Hosmer, David W, Lemeshow, Stanley, Sturdivant, Rodney (2013). Applied Logistic Regression ISBN 978-0-470-58247-3.

Instituto Colombiano de Normas Técnicas y Certificación. Presentación de tesis, trabajos de grado y otros trabajos de investigación. NTC 1486.

Laredo, Abraham, Pedroso, Antonio, Okaze, Susana (2012). Análisis Empírico de los Indicadores KS y ROC. Revista Tecnología de Crédito. Serasa & Experian.

Molina Félix, Luis Carlos (2014). Torturando los datos hasta que confiesen.

Quezada Lucio, Nel (2014). Estadística con SPSS 22. Editorial Macro.