



REPORTE DE ESTABILIDAD FINANCIERA

Marzo de 2011

Estimación de la carga financiera en Colombia

Gerencia de Riesgo
Asobancaria - CIFI

Estimación de la Carga Financiera en Colombia

Gerencia de Riesgo
Asobancaria - CIFI

Resumen

Los indicadores de endeudamiento permiten evaluar la posición crediticia de cada individuo. A nivel agregado, contribuyen al seguimiento de la dinámica y sostenibilidad del crédito en el sistema. A pesar de su importancia, en Colombia, la información financiera que se obtiene a nivel micro no se encuentra disponible para la construcción de indicadores a nivel macro. Por esta razón se propone un método de inferencia de la renta de los individuos a partir de su relación con el gasto financiero, que permite medir la carga financiera para la totalidad de titulares de crédito en el país. La metodología parte de la conformación de segmentos para estimar, en cada uno de ellos, regresiones múltiples basadas en Redes Neuronales Artificiales (RNA). La agregación de indicadores micro fundamentados puede contribuir al diseño e implementación de mecanismos de estabilización financiera. Los resultados sugieren exceso probable de endeudamiento para un 10% de la población.

Abstract

Debt indicators contribute to individual credit limits assessment as well as an instrument to monitor the system's credit dynamics and sustainability. As far as the financial information obtained on a micro level basis, is not available at the macro level in Colombia, we develop an income-inference methodology for credit holders in Colombia, relying upon the relationships between individual financial expenditure and income. The methodology considers cluster analysis and Artificial Neural Network (ANN) based multiple regression models for every cluster. The resulting estimates, allow for the construction of micro-founded aggregate indicators, useful for the design and implementation of financial stabilization mechanisms. Results suggest about 10% of population facing excess debt burden.

1 Introducción

El exceso de endeudamiento ha sido señalado por la literatura como factor desencadenante de las crisis financieras - incluyendo la más reciente - (Dey, 2010; Stiglitz, 2010). Previamente, el Comité de Basilea advertía sobre la necesidad de fortalecer los sistemas de valoración de capacidad de pago de clientes y de implementar mecanismos de seguimiento a los límites de crédito. El objetivo era procurar cargas financieras adecuadas que permitieran el pago de las obligaciones (Konovalikhin et al, 2008). En general, excesos de endeudamiento pueden generar burbujas y afectar la capacidad que tienen los individuos para cumplir con sus obligaciones financieras, lo cual puede redundar en problemas sistémicos.

A nivel agregado, los mismos indicadores pueden contribuir al seguimiento y control de la dinámica de endeudamiento de los hogares, ya sea a nivel general o por segmentos de población, regiones, perfiles socio-económicos, entre otros. Sin embargo, la información relativa a los ingresos individuales, no se encuentra disponible para la construcción de indicadores macro. Esta restricción generalizada ha llevado a la construcción de razones globales orientadas a la evaluación de impactos promedio.

Es así como los datos referentes a la renta de los hogares - como encuestas de ingresos y gastos -, o de los individuos - como censos poblacionales -, sólo permiten evaluar aspectos tales como el nivel de ingreso promedio o su relación con factores socio-demográficos a nivel agregado (Sala-i-Martin, 2002; Quah, 2002; Bourguignon y Morrison, 2002; Bhalla, 2002). Sobre estos mismos datos se construyen algunos indicadores generales como el porcentaje de gasto promedio en diversos rubros, entre los que se cuentan, el gasto financiero o los impuestos.

Otras metodologías utilizan la información censal para aproximar la distribución de la renta (Chotikapanich y Griffiths, 2004; Gertel et al, 2001; Kitov, 2009; Victoria-Feser, 1993) o para realizar estimaciones de pobreza con cifras agregadas de ingreso (Fay y Train, 1995, Fisher, 1996). De otro lado, son recurrentes ejercicios de imputación múltiple para completar o actualizar datos sobre la renta de los hogares a partir de encuestas (Fisher, 2006; Cooper, 2010; Mandal y Stasny, 2003) y algunos autores incluyen también, variables censales para estimar la probabilidad de pertenencia de un hogar a un determinado segmento de ingreso (Ferguson, et al 2003). Además de un descriptivo de la estructura y composición de ingresos y gastos de la población, esta información permite observar algunas tendencias intercensales de largo plazo que dan idea de cambios estructurales en la distribución de la renta, en la composición del gasto o en la carga financiera promedio de los individuos, con un rezago considerable.

Por otra parte, se han desarrollado modelos orientados a medir desigualdad (Gallup, 2010) o diferencias en renta (o bienestar) asociadas a dotaciones específicas como vivienda propia (Frick y Grabka 2002). También han sido comunes los ejercicios orientados a medir convergencia regional en renta (Chasco, 1999), algunos de ellos basados en matrices de transición (Wang, 2004). Al final, se evidencian las dificultades de acceso a información sobre ingresos así como problemas de estimación o imputación, especialmente en los segmentos de renta baja (Kyobe y Daninger, 2005). De esta manera, aunque la información relativa a la renta de los individuos reporta una invaluable utilidad, no es posible realizar

seguimiento permanente sobre la misma ni, por supuesto, ajustar decisiones individuales ni de política, en el corto plazo.

Las instituciones financieras y los burós de crédito, han registrado avances en la estimación de ingreso personal (Brannan, 2010). No obstante, el objetivo particular de sus metodologías es el de evaluar la capacidad individual de clientes actuales o potenciales para atender su portafolio. Por ello, no se revela ningún detalle metodológico y sus resultados no aplican necesariamente a la construcción y seguimiento de indicadores agregados.

En consecuencia, la carencia de información para construir indicadores individuales de carga financiera (o límites de endeudamiento) que puedan ser aplicados a nivel macro no ha sido resuelta. El presente documento propone una metodología para la inferencia del ingreso - mínimo probable - de cada individuo, a partir de información crediticia y datos socio-demográficos disponibles en el buró de crédito CIFIN.

El estimativo de ingreso individual, confrontado con la información crediticia propia de cada titular, permite obtener indicadores ajustados de carga financiera, cuya aplicación puede contribuir al diseño y evaluación de políticas: a nivel micro, por ejemplo, razones como la de desembolsos a ingreso, permiten diferenciar perfiles crediticios o de riesgo en la cartera de un banco, con las consecuentes acciones de mitigación o control. A nivel macro, es posible agregar saldos e ingresos para obtener un indicador global, pero micro-fundamentado, de endeudamiento, que puede contribuir igualmente a la implementación de mecanismos de control, en este caso, monetario o financiero.

El aporte esencial de la metodología radica en el carácter micro de los insumos¹. Al ser aplicado a la base completa de titulares de crédito en Colombia², indicadores como el de carga financiera pueden ser descompuestos por nivel de ingreso, por saldo, por modalidad de crédito, por cantidad de productos, por región y, en general, segmentados de acuerdo con factores específicos que permitan el diseño de mecanismos especializados que focalicen los esfuerzos de política en segmentos críticos.

El presente documento consta de cuatro partes además de esta introducción. El capítulo 2 describe la metodología de estimación del ingreso; los resultados se presentan en el capítulo 3. Una cuarta sección comprende una revisión a algunos indicadores de carga financiera en Colombia. El capítulo 5, concluye.

¹ En el agregado, el indicador de endeudamiento de los individuos puede eventualmente aproximarse a un estimador que relacione la cartera total del sistema sobre la remuneración a los asalariados, por ejemplo.

² Reportados al buró, los cuales comprenden aproximadamente, 16 millones de identificaciones que fueron calificadas con el modelo desarrollado a partir de la muestra de estimación.

2 Metodología para la inferencia de la carga financiera

Tradicionalmente, la carga financiera de los hogares ha sido calculada a través de razones entre agregados, por lo que los mecanismos de estabilización y regulación, tienden a controlar impactos promedio. Sin embargo, las políticas deberían poder focalizarse en sectores con alta exposición o en segmentos vulnerables; por ejemplo, aumentando su efectividad en el primer caso y reduciendo su impacto, en el último. Ello implica, sin embargo, contar con información desagregada suficientemente que permita identificar estas franjas.

Por esta razón, se propone una metodología para calcular carga financiera con base en información micro, con el objeto de evaluar cada uno de sus componentes, para todas y cada una de las unidades de observación. El eje del modelo es la inferencia de ingresos de los individuos. El complemento, es su confrontación con la información crediticia individual de cada titular de crédito. La construcción de los indicadores a nivel macro comprende la agregación de estos datos.

La metodología de inferencia de ingreso, ha sido desarrollada a partir de toda la información crediticia y socio-demográfica reportada al buró de crédito CIFIN, que es relacionada con una base de titulares con datos de ingreso mensual. El modelo se sustenta en intensivos procesos de segmentación de individuos que registran perfiles crediticios similares y reportan, en efecto, niveles de ingreso semejantes. Posteriormente, se entrenan redes neuronales artificiales (RNA) basadas también en variables crediticias. Los resultados son bastante ajustados en los ingresos promedio de cada uno de los 62 segmentos que componen la metodología. No obstante, la volatilidad típica de cada nodo, no resulta suficientemente recogida por las redes.

Por esta razón, se construyó un intervalo de confianza en el cual, por lo menos un 90% de los individuos cuenta efectivamente con un ingreso mayor o igual que el estimado por el modelo. De allí que el resultado final sea el *Ingreso Mínimo Probable* para cada individuo³.

La información utilizada en el desarrollo de la metodología comprende exclusivamente productos consolidados (más de un año en el portafolio de cada titular) con buen manejo (sin morosidad ni comportamientos que impliquen riesgo) además de obligaciones saldadas recientemente de manera voluntaria, sin malas calificaciones por riesgo, morosidad, mal manejo, reestructuraciones o castigos en su historial. Así, sólo hace parte del estimativo de ingreso, aquel portafolio que cada titular evidencia soportar suficientemente⁴.

La relación entre la renta y el perfil crediticio de los titulares, derivado de la información disponible en CIFIN, constituye una aproximación al gasto total de los individuos: una elevada utilización de

³ Esta definición surge del objetivo planteado por el buró de evaluar la carga financiera individual o contribuir a la administración de límites crediticios al interior de entidades que generan crédito.

⁴ Un titular con tres obligaciones, una de ellas en mora de 60 días por ejemplo, sólo soporta de manera efectiva las obligaciones al día por lo que sólo estas hacen parte de su estimativo de gasto. Asimismo, se consideran únicamente obligaciones con más de un año de maduración, con el fin de verificar la capacidad que tiene cada individuo para soportar su portafolio. Ello permite controlar asimismo incrementos inmediatos - y sucesivos - de la renta ante un mayor endeudamiento del titular.

servicios financieros implica una mayor disposición a gastar y viceversa. Por esta razón, el estimativo obtenido de la metodología constituye una expresión (revelada) del gasto individual, que se asume, guarda relación directa con el ingreso.

El motivo por el cual la metodología establece la capacidad de pago revelada desde el gasto, radica en que una mayor renta no necesariamente implica un mayor portafolio (o una mayor utilización) de productos y servicios financieros; en particular, un gasto financiero muy bajo, se relaciona de manera directa con un bajo nivel de gasto, mas no así, con un bajo nivel de ingreso.

2.1 Proceso de segmentación

El desarrollo metodológico parte de la consolidación de 2 millones de registros de titulares de crédito en Colombia con sus respectivos ingresos⁵, recopilados entre 2007 y 2010. Un 20% de estos titulares presenta un ingreso menor o igual que \$500 mil y un 40%, superior a \$1,5 millones, Estos porcentajes difieren de la distribución promedio de la renta en Colombia (60% de la población presenta ingresos bajos) debido a que los titulares de crédito registran en general, mayores ingresos⁶.

La validación del ingreso de cada titular, se realizó a partir de reglas basadas en la información del buró: el cupo total de las tarjetas de crédito, por ejemplo, debe representar máximo, 5 veces el salario reportado; las cuotas asociadas a crédito hipotecario no deben superar la mitad del ingreso⁷; la suma total de las cuotas reportadas no debe exceder un 80% de la renta, entre otros.

Enseguida, un análisis de clúster permitió identificar grupos con características semejantes, que permitían diferenciar niveles de ingreso entre los individuos incluidos en ellos. Este análisis de clúster tiene por fin, agrupar individuos de acuerdo a las características que poseen, agrupando perfiles similares entre ellos (homogeneidad entre individuos del mismo grupo) y perfiles diferentes entre estos grupos (alto grado de heterogeneidad entre los grupos). Existen dos técnicas de análisis clúster, los métodos jerárquicos y no jerárquicos. Los métodos jerárquicos son aquellos que para formar un nuevo clúster, unen o separan alguno ya existente, para dar origen a otro clúster, de forma que se maximice la similitud o se minimice la distancia. Los métodos no jerárquicos clasifican los individuos en k grupos, estudiando todas las particiones de individuos en esos k grupos y eligiendo la mejor partición. La principal ventaja de los métodos jerárquicos consiste en representar el problema en forma de árbol o dendograma, donde se observa fácilmente la solución final.

Este proceso se encuentra soportado por la distribución de los datos mismos. A pesar de relaciones complejas en el agregado, se evidencian grupos de titulares en los cuales la relación entre el ingreso y variables referentes a su perfil crediticio es relativamente uniforme. En la Figura 1, por ejemplo, se relacionan, ingreso, cuota total y cupo en tarjetas de crédito. Como un todo, es prácticamente imposible

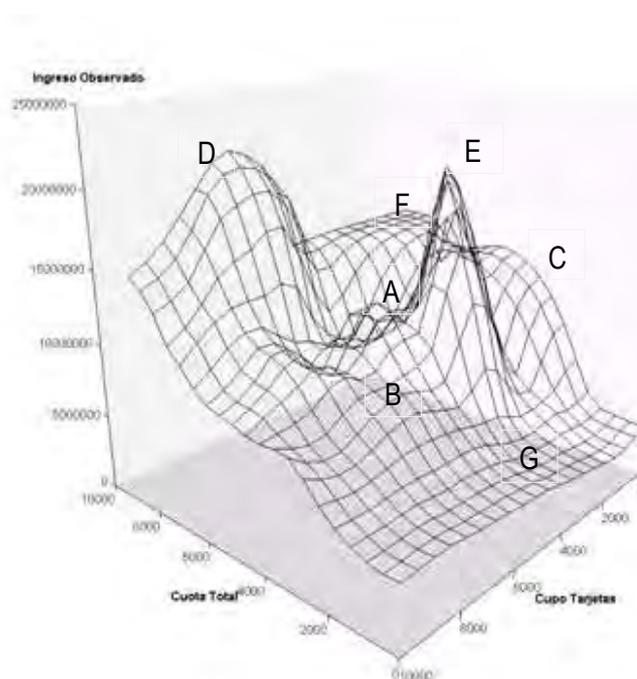
⁵ Esquema de información compartida, exclusivamente para uso estadístico.

⁶ Ello pone de manifiesto la necesidad de indicadores micro-fundamentados para la evaluación de la dinámica del crédito o el diseño de mecanismos de seguimiento y control a nivel agregado.

⁷ Aunque la normatividad propone un límite equivalente a 30% del ingreso mensual para el pago de cuotas hipotecarias, la validación pretende dar un margen sobre cuotas adicionales y reflejar la posibilidad de más de un deudor asociado a cada crédito de vivienda.

encontrar un modelo capaz de interpretar y caracterizar la interacción entre estas tres variables. No obstante, algunos segmentos de clientes permiten aproximar comportamientos más estables. En el segmento “E”, por ejemplo, se encuentra clientes con ingresos superiores, pero cupos y cuotas bajas. En el segmento “D”, los ingresos, las cuotas y los cupos, son altos. El segmento “G” corresponde a cuotas e ingresos bajos.

Figura 1. Operación del proceso de Segmentación de la Base



Así, en la primera etapa se conformaron nueve grupos jerárquicos, cinco de ellos definidos por la modalidad dominante de su portafolio (comercial, hipotecario, consumo, tarjeta de crédito) así como del valor de sus desembolsos o cupos. Tres grupos más fueron caracterizados por información patrimonial y un último grupo recogió los demás casos (Anexos 1 a 6).

Los segmentos que agrupan el menor número de individuos, son : *i.* titulares de créditos hipotecarios superiores a \$50 millones, quienes se caracterizan por ingresos altos (\$6 millones en promedio); *ii.* titulares de créditos de consumo por más de \$20 millones, que registran una renta superior (media igual a \$4,7 millones) y, *iii.* personas que reportan CDT y cuentan con portafolio en sectores diferentes del financiero (ingreso promedio igual a \$1,8 millones).

Los grupos restantes (cartera comercial, cupos altos, cupos medios, inmuebles, vehículos y resto) presentan un número significativo de identificaciones y, asimismo, una alta dispersión interna de la

renta. Por ello, aplicaron procesos adicionales de segmentación mediante árboles de regresión⁸. Esta metodología elige en cada paso una variable que presenta alta correlación con el ingreso y que es común al portafolio de los titulares (valor inicial, cantidad de productos, antigüedad, cuota). Las variables son categorizadas y el proceso se repite hasta obtener una secuencia de agrupaciones o nodos, que reportan niveles de ingreso relativamente cercanos en su interior, y estadísticamente diferentes de los demás grupos.

De esta manera, la amplia dispersión de los ingresos al interior del grupo jerárquico de titulares con cartera comercial (ingreso medio igual a \$4,5 millones), es administrada en el modelo mediante una segmentación en 7 nodos, los cuales se encuentran asociados a variables complementarias del portafolio, tales como el número de cuentas corrientes, el monto desembolsado, la cuota y el saldo actual⁹. Los titulares de menor ingreso (\$1,8 millones¹⁰) se concentran en clientes con desembolsos y cuotas bajas, sin cuenta corriente entre su portafolio. Titulares con cuotas altas - incluso en consumo -, registran ingresos superiores (\$8,7 millones - Anexo 1).

El grupo de individuos con altos cupos en tarjeta de crédito, registra también una alta dispersión en renta (en un 90% de los casos, el ingreso oscila entre \$2,6 millones y \$16 millones). Por ello se procede a segmentar en 6 nodos asociados a variables tales como la sumatoria de los desembolsos y el cupo mismo en tarjetas de crédito. Clientes con desembolsos totales inferiores a \$45 millones, registran el menor ingreso de este grupo (\$4,4 millones). Titulares con los desembolsos más altos, así como sobregiros sobre sus cuentas corrientes, presentan ingreso superior (\$9,8 millones – Anexo 2).

Los titulares con cupos medios en tarjeta de crédito, son separados en 3 nodos que dependen de la cantidad de tarjetas de crédito y el número de cuentas corrientes (Anexo 3). Personas con 2 tarjetas de crédito o menos, se caracterizan por un menor ingreso (\$2,7 millones), mientras titulares con cuentas corrientes presentan el ingreso promedio mas alto del grupo (\$4,5 millones).

Clientes del sector financiero con registro patrimonial sobre un inmueble, son re-agrupados en 9 nodos que dependen de las cuotas, el número de cuentas corrientes, el monto de los desembolsos además de la antigüedad de sus productos en el sector financiero (Anexo 4). Los titulares de estratos 1 y 2 comprenden el conjunto de personas de ingresos más bajos (\$1,2 millones); los clientes que soportan una cuota superior a \$600 mil (en créditos al día únicamente) registran ingresos superiores (\$4,6 millones).

El segmento de clientes con registro patrimonial sobre vehículo se descompone en 7 nodos (Anexo 5), con ingresos promedio que oscilan entre \$ 1,6 millones (clientes sin cuenta corriente, con desembolsos totales inferiores a \$8 millones y utilización de cupos) y \$6,7 millones (titulares con cuota superior a \$925 mil).

⁸ Estos árboles crecen por *detección automática de interacciones chi-cuadrado* (CHAID por sus siglas en inglés).

⁹ El ingreso observado en este nodo oscila entre \$900 mil y \$14 Millones (percentiles 5 y 95).

¹⁰ En adelante, el valor entre paréntesis corresponde al promedio del grupo o segmento correspondiente.

Titulares con historial crediticio limitado o desembolsos bajos son asignados a 25 nodos (Anexo 6). El ingreso promedio entre estos nodos se encuentra entre \$600 mil (clientes sin cuenta corriente, cupo bajo y 1 producto reciente) y \$3,2 millones (titulares de cuenta corriente).

Se obtienen así 9 segmentos y 62 nodos, al interior de los cuales es posible identificar niveles de ingreso significativamente diferentes (la renta promedio oscila entre \$600 mil y \$10 millones). La descomposición en múltiples grupos, permite mitigar la dispersión de los datos al interior de cada nodo y develar asimismo relaciones no observables en el agregado, Todo ello hace posible una aproximación a la renta individual a partir de información contenida en el buró, como se desarrolla enseguida.

2.2 Estimación de la renta mediante regresiones múltiples (RNA)

Para cada uno de los segmentos y nodos definidos en la sección anterior, se estimaron regresiones múltiples basadas en Redes Neuronales Artificiales (RNA). En ellas, modelos lineales ponderan inicialmente las variables de gasto financiero (estandarizadas) que presentan mayor correlación con el ingreso. Un proceso iterativo permite aproximar las relaciones no lineales entre estas variables y la renta.

La importancia de entrenar sistemas de regresiones múltiples en redes neuronales artificiales radica, de una parte, en que la relación entre ingresos y “gasto financiero” no es lineal: individuos con ingresos altos presentan una mayor propensión a la utilización de productos crediticios que los ingresos bajos. También, la dispersión de variables como saldos, cantidad de productos o utilización de cupos varía considerablemente en función de la renta de los individuos. Finalmente, todas las variables que contribuyen a explicar el gasto total presentan una fuerte asociación entre sí.

Se probaron diversas metodologías y, en todos los casos, los resultados por RNA fueron más eficientes, lo que confirma algunas conclusiones previas que se desprenden de análisis comparativos entre métodos de predicción por modelos estadísticos y redes neuronales¹¹, los cuales favorecen a éstas últimas por sus beneficios en términos de ajuste, error de clasificación así como la omisión de supuestos y mejores pronósticos para aplicaciones particulares (HSiao, 2008).

2.2.1 Discusión metodológica

El programa de investigación sobre Redes Neuronales Artificiales (RNA) inició como un intento por simular funciones neuronales biológicas (McCulloch y Pitts, 1943), relacionadas con procesos de aprendizaje y opciones de respuesta. En 1949, Hebb estableció dos principios esenciales – derivados

¹¹ Sólo bajo condiciones idóneas de aplicabilidad en los experimentos es posible que se presente un mejor desempeño en modelos estadísticos tradicionales. Estas condiciones comprenden 7 variables predictivas que correlacionan de forma baja entre sí (ρ_{xy} promedio=.05), pero cada una de ellas correlaciona significativamente con la respuesta (ρ_{xy} promedio=.35). Ello favorece modelos estadísticos convencionales (Pitarque, 1998).

del campo psico-fisiológico - para el desarrollo de esta metodología: *i.* el aprendizaje se localiza en conexiones (sinapsis) entre las neuronas y *ii.* la información se representa en el cerebro humano mediante un conjunto de neuronas activas o inactivas (Stergiou y Siganos, 1998). Estos principios, sintetizan las reglas de aprendizaje aplicadas por los modelos actuales (Ballesteros, 2006).

Con base en estos postulados, se modelaron neuronas como dispositivos binarios con umbrales fijos y funciones lógicas simples como "A o B" y "A y B", que permitieron algunas simulaciones computables (Farley y Clark, 1954; Rochester, et. al., 1956). Desde entonces, la metodología ha evolucionado y sus aplicaciones han trascendido a múltiples campos, incluyendo análisis y predicción de eventos complejos, investigación patológica y clínica, comprensión de imágenes, reconocimiento de patrones de voz y múltiples aplicaciones que involucran grandes cantidades de datos.

La creciente investigación y aplicación de modelos de redes neuronales, se debe principalmente a su carácter no paramétrico, lo cual permite desvincular las estimaciones de los supuestos teóricos requeridos por modelos lineales o análisis multivariado (Smith, 1993), así como a su capacidad para capturar relaciones complejas y no lineales, generalmente ocultas entre variables. En Colombia, la aplicación de modelos RNA para evaluar relaciones (no lineales) entre dinero e inflación, ha permitido establecer asimetrías en los procesos de fijación de precios (Jalil y Misas, 2006; Aristizabal y Misas, 2006) y, de esta forma, generar pronósticos más precisos de inflación. Según el estudio, es posible que asimetrías entre impulsos y respuestas configuren relaciones no lineales que pueden ser capturadas, a través de sus capas ocultas, por los modelos de redes neuronales (Misas, 2002).

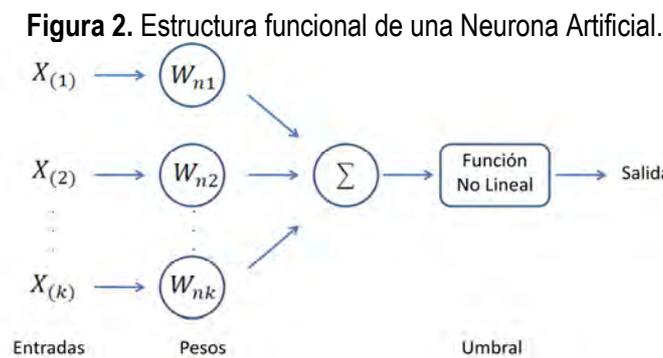
Por otra parte, la aplicación de redes neuronales al pronóstico de la renta individual no ha sido registrada por la literatura. El presente documento realiza una primera aproximación en ese sentido a partir de información crediticia individual reportada al buró de crédito CIFIN. En este caso, la aplicación de esta metodología se encuentra justificada por las asimetrías y no-linealidades que pueden derivarse de diversas fuentes: en primer lugar, los agentes individuales (titulares de los créditos) cuentan con mayor información sobre sus ingresos y gastos; las instituciones financieras sólo conocen la parte de ingreso que puede ser verificada y, en especial, sólo la parte de egresos relacionada con el rubro de gasto financiero. En segundo lugar, la relación entre renta y capacidad de endeudamiento no es lineal, en la medida en que individuos con mayores ingresos cuentan con una capacidad de ahorro más alta y, por ende, disponen de mayor disponible para consumo o inversión. En otras palabras, el portafolio de los titulares crece más que proporcionalmente que su ingreso. Estas características implican la necesidad de caracterizar relaciones complejas entre variables de ingreso y gasto y contemplar esquemas menos restrictivos que los asociados a modelos lineales, por ejemplo, para una aproximación más cercana al ingreso de los individuos a partir de información crediticia. A continuación se explora la metodología de Redes Neuronales Artificiales aplicada en la construcción del modelo, con énfasis en la interpretación de sus resultados.

2.2.1.1 Arquitecturas Neuronales

Se denomina arquitectura neuronal al conjunto de procesadores elementales (neuronas), unidades estructurales (capas de procesamiento) y sus conexiones (sinapsis). El comportamiento de cada red

depende, a su vez, de su estructura de conexiones, del esquema de procesamiento de la información (algoritmo de aprendizaje) y del mecanismo de respuesta (función de activación). El grado de complejidad de las redes, depende de la combinación de estos elementos.

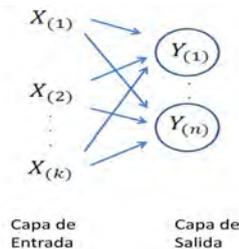
Las neuronas puede ser representadas por unidades de proceso que forman parte de una entidad mayor, la red neuronal. Estos procesadores elementales reciben estímulos (entradas) cuyas ponderaciones () recogen el impacto de estos impulsos para ser transmitido como respuesta final o bien, como entrada para otras neuronas (en cuyo caso, se constituye una red neuronal). La magnitud de la respuesta (variable de salida) depende de la función de activación (figura 2).



Las entradas pueden ser valores enteros, reales o binarios (que equivalen a las señales externas o que envían otras neuronas). Los mecanismos de sinapsis para transmitir la señal comprenden modelos de regresión y sus coeficientes, los cuales tienen lugar en la primera capa de la neurona¹². Estos valores entran en la función de ponderación, que los transforma en una respuesta.

Los conjuntos de neuronas que interactúan conforman capas, las cuales definen dos tipos de arquitectura: monocapa o multicapa. Las redes monocapa corresponden a modelos de regresión múltiple, donde no existe una capa intermedia entre las entradas y la salida (la conexión es directa). En consecuencia, estas redes son asimilables a un modelo de regresión (figura 3).

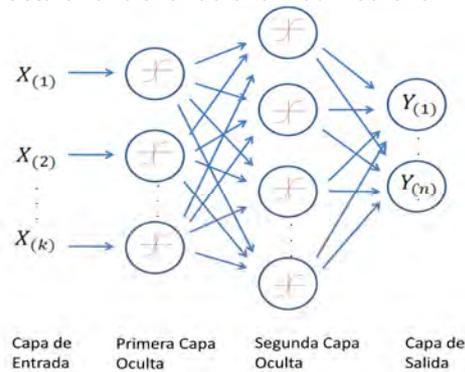
Figura 3. Estructura funcional de una Red Neuronal Artificial Monocapa.



¹² y , equivalen a las señales químicas inhibitorias y excitadoras que se dan en las sinapsis y que inducen a la neurona a cambiar su comportamiento.

Las redes multicapa comprenden una extensión del caso anterior, que incluye una o más capas intermedias entre entradas y salidas. El funcionamiento de una RNA multicapa inicia con la entrada de información (*input*). Durante el entrenamiento se estiman los pesos sobre las variables de entrada para generar, a través de una combinación lineal, una respuesta (*output*). Dicho resultado es transmitido a la siguiente capa, que puede ser de procesamiento intermedio (capa oculta) o de salida (figura 4). La transmisión de información está controlada por su respectiva función de activación.

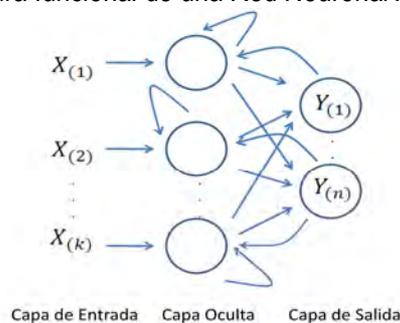
Figura 4. Estructura funcional de una Red Neuronal Artificial Multicapa.



Finalmente, estas redes pueden estar conectadas total o parcialmente. En el primer caso, cada neurona de la capa precedente se encuentra conectada con todas y cada una de las neuronas que confrontan la capa siguiente. En el último caso, alguna (s) neurona (s) puede tener conexión sólo con alguna (s) neurona (s) de la capa posterior.

Ahora bien, según el tipo de conexiones las RNA pueden ser clasificadas en No Recurrentes y Recurrentes. Las redes No Recurrentes se caracterizan por no contar con mecanismos de retroalimentación, de tal manera que no tienen memoria. Las redes Recurrentes, en cambio, pueden presentar conexiones entre neuronas de diferentes capas o entre neuronas de una misma capa, o incluso conexiones en la misma neurona (figura 5). Son estos esquemas complejos de interacción, quienes permiten capturar fenómenos no lineales y asimetrías de información.

Figura 5. Estructura funcional de una Red Neuronal Artificial Recurrente.



2.2.1.2 Métodos de Aprendizaje (Algoritmos de Optimización)

El aprendizaje en RNA consiste en un proceso de ajuste de parámetros orientado a replicar respuestas conocidas frente a estímulos particulares observables. Una vez los parámetros convergen, se considera que la red ha sido entrenada (i.e. ha aprendido).

Los métodos de aprendizaje se pueden agrupar según el objetivo del problema: supervisado, en el caso de pronóstico de una variable definida (repuesta); o no supervisado, como son los problemas de clasificación (las redes se entrenan por relaciones entre variables de entrada mismas). Dentro de las opciones de entrenamiento supervisado se encuentran, aprendizaje por refuerzo y aprendizaje por corrección. El primer caso aplica a problemas de clasificación; el último, la estimación puntual de una respuesta.

La selección del método de optimización o entrenamiento, depende de la estructura del modelo que se propone ajustar. Si bien, los métodos de optimización convergen a los mismos parámetros, esta combinación define la velocidad de convergencia y la intensidad computacional. Por esta razón, es posible seleccionar entre múltiples opciones de iteración:

Levenberg-Marquardt: Soluciones numéricas por mínimos cuadrados para arquitecturas de RNA con un número de parámetros (W_i) no superior a 100. Aplica especialmente a funciones no lineales y presenta mínima dependencia de las condiciones iniciales.

Quasi-Newton: Técnica recomendada para arquitecturas de RNA con un número de parámetros moderado (entre 100 y 500 W_i). El método minimiza la matriz de segundas derivadas hasta alcanzar puntos estacionarios (óptimos locales).

Trust-Region: Algoritmo de optimización mediante aproximaciones cuadráticas a regiones delimitadas de la función objetivo (la región de confianza). Una vez se encuentra un modelo adecuado, se expande la región para iterar nuevamente,

Double Dogleg: Esta técnica combina Quasi-Newton y métodos de *Trust-Region*. El método se desempeña bien donde la función objetivo y el gradiente son de cálculo más rápido que la matriz Hessiana.

Conjugate Gradient: Recomendado para arquitecturas que requieren estimar un alto número de parámetros (mas de 500), por lo que el método es de lenta convergencia.

Otros algoritmos, contemplan tasas de aprendizaje para evaluar el desempeño de las redes. Entre estos, es posible aplicar los siguientes:

Estándar Back-Prop: A pesar de ser la técnica más popular de entrenamiento, se caracteriza por ser lenta, poco confiable y con ajustes en la tasa de aprendizaje que pueden resultar tediosos. Adicionalmente, la técnica puede hallar mínimos locales mas no encontrar mínimos globales.

Incremental Back-Prop: En este método, los pesos son incrementados después de leer cada observación a diferencia del método *Estandar Back-Prop* que lee todos los datos y después ajusta los pesos. El ajuste a los parámetros mediante ensayo y error involucra un mecanismo de aprendizaje que puede consumir un tiempo de cálculo considerable.

Quickprop: Este algoritmo usa la misma línea de búsqueda que el gradiente descendiente para lograr convergencia. La técnica converge rápido pero requiere más iteraciones que las técnicas gradiente conjugado. De otro lado, la convergencia depende de los parámetros iniciales.

RPROP: De iteraciones rápidas pero numerosas. Usa diferentes tasas de aprendizaje para cada estimación de los pesos, donde la tasa de aprendizaje es delimitada dentro de un intervalo específico. Este método es más estable que las técnicas “Prop” mencionadas anteriormente.

2.2.1.3 Función de Activación

La función de activación juega un papel crítico en RNA: normalizar los pesos estimados en la capa oculta o la capa de salida (con una distribución centrada en cero) con el objetivo de asegurar convergencia a los parámetros correctos.

Estas funciones son de tipo lineal (las más sencillas), escalón (para redes binarias) e hiperbólicas o tangenciales. Entre estas últimas, la función más común es la logística, definida como $f(x) = \frac{1}{1+e^{-x}}$ que redefine el rango de la variable objetivo desde el intervalo $(-\infty, \infty)$ hacia intervalo $[0, 1]$. También es común la aplicación de la tangente hiperbólica, definida como $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, que normaliza la respuesta dentro del intervalo $[-1, 1]$ con alta velocidad de convergencia para los procesos de optimización (comparado con funciones logísticas) especialmente en tamaños de muestra, grandes.

Otras funciones de activación hiperbólicas mapean también valores reales en intervalos unitarios, con el fin de normalizar las respuestas parciales obtenidas de cada unidad de procesamiento, dependiendo de los insumos y las respuestas (continuos, discretos). Entre ellos se encuentran los siguientes (Matignon, 2005):

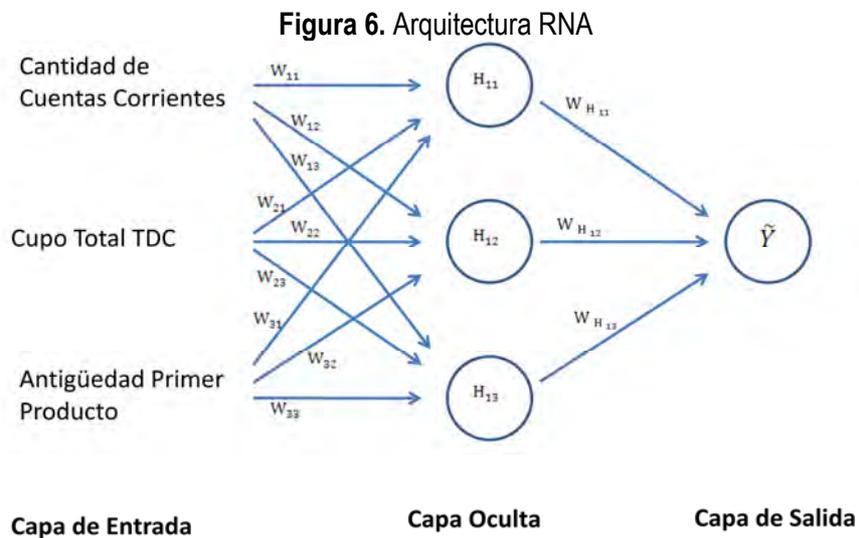
<i>Elliot</i>	$f(x) = \frac{1}{2} \left(\frac{1+x}{1+ x } \right)$	<i>Gausiano</i>	$f(x) = e^{-\frac{x^2}{2}}$
<i>Identidad</i>	$f(x) = x$	<i>Exponencial</i>	$f(x) = e^x$
<i>Raíz cuadrada</i>	$f(x) = \sqrt{x}$	<i>Recíproco</i>	$f(x) = \frac{1}{x}$

En el presente documento, se optó por una arquitectura Multicapa - con una capa de entrada, una capa oculta y una capa de salida – para cada uno de los segmentos definidos previamente. Los impulsos comprenden toda la información crediticia relativa al portafolio de cada titular como puede ser, la

cantidad de productos, los montos desembolsados, los porcentajes de pago, el servicio de las deudas, el valor de los cupos, etc. La capa oculta en cada una de las redes estimadas, incorpora entre 3 y 5 neuronas en una estructura totalmente conectada. La capa de salida, genera una respuesta única que corresponde a los ingresos estimados de cada individuo.

La función de activación utilizada es tangente hiperbólica, seleccionada por el tamaño de muestra y el tipo de variable respuesta (continua). El algoritmo aplicado fue el de Levenberg-Marquardt, que minimiza la suma de cuadrados del error y evalúa múltiples alternativas para alcanzar la convergencia de parámetros. Este algoritmo ofrece también ventajas cuando los parámetros estimados se encuentran altamente correlacionados.

A continuación se ilustra la operación de esta arquitectura: una red perceptron multicapa con tres variables, una capa oculta con tres nodos, la función de activación es tangente hiperbólica y una de capa de salida con variable respuesta continua (figura 6).



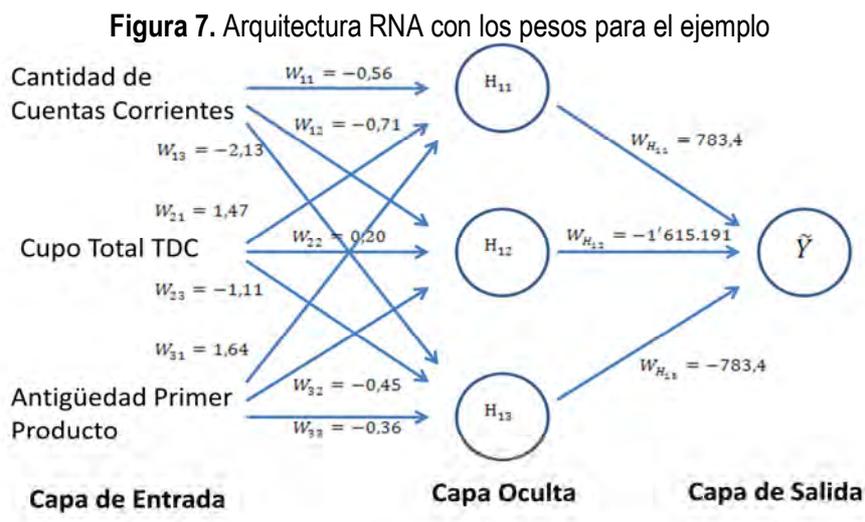
El pronóstico está dado por

Donde,

En (2), se tiene además,

De esta manera, las variables de entrada son normalizadas y, a partir de esta transformación, se estiman las ponderaciones (entrenamiento de la RNA). En la capa oculta se aplica una transformación no lineal (tangente hiperbólica) sobre las respuestas obtenidas en la primera etapa y, en la capa de salida se estiman nuevos parámetros de manera lineal, para obtener una salida. El proceso se repite hasta minimizar la diferencia entre los valores observados y los valores estimados de la variable respuesta (convergencia de los parámetros).

Con fines de exposición, se estimó una versión simplificada de una red con tres variables dentro del segmento de Tarjetahabientes con cupos totales inferiores a \$ 5 millones (figura 7).



La primera neurona asigna el mayor coeficiente (1,64) a la antigüedad del titular (valor estandarizado), seguido por la suma de sus cupos en tarjetas de crédito (1,47). La ponderación más baja aplica a la titularidad sobre cuentas corrientes (-0.56). Esta combinación lineal constituye la contribución optimizada de esta primera neurona a la salida de la red. Las demás neuronas estiman sus parámetros para generar sus propias contribuciones. Estos insumos son evaluados y ponderados en la capa oculta para establecer un esquema interconectado de pesos, que lleva a un estimativo final del ingreso de cada individuo.

Ejemplo 1. *Aplicación de la versión simplificada para el segmento de clientes con cupos bajos sobre un titular.* El ingreso medio para este segmento es igual a \$3,5 millones. Los valores estandarizados de variables insumo para una identificación particular son los siguientes:

$$\text{Cuentas Corrientes} = 0 \quad Z_1 = -2,48$$

$$\text{Cupo Total TDC} = 2.000.000 \quad Z_2 = -0,29$$

$$\text{Antigüedad Primer Producto} = 2 \quad Z_3 = -0,92$$

A partir de estas entradas, de (2) se tiene:

$$H_{11} = \text{TanHip}\{-0,68 + (-0,56 * Z_1 + 1,47 * Z_2 + 1,64 * Z_3)\} = -0,83$$

$$H_{12} = \text{TanHip}\{2,74 + (-0,71 * Z_1 + 0,20 * Z_2 - 0,45 * Z_3)\} = 0,99$$

$$H_{13} = \text{TanHip}\{-0,94 + (-2,13 * Z_1 - 1,11 * Z_2 - 0,36 * Z_3)\} = 0,99$$

Luego, de (1) y reemplazando los pesos obtenidos del entrenamiento se tiene,

$$\tilde{Y} = 5.082.416 + 783.453 * H_{11} - 1.615.191 * H_{12} - 783.453 * H_{13} = \mathbf{2.026.243}$$

De esta forma, el pronóstico de ingreso para una persona sin cuenta corriente, con cupo total en TDC igual a \$ 2 millones y con una antigüedad de 2 meses en el mercado, es de \$2.026.243.

Ejemplo 2. *Aplicación del mismo modelo a una identificación diferente.* En este caso, el individuo registra 2 cuentas corrientes, cupo total en Tarjeta de Crédito (TDC) igual a \$ 4 millones y antigüedad de 3,8 años en el mercado. Las variables estandarizadas para este caso, son:

$$\text{Cuentas Corrientes} = 2 \quad Z_1 = 1,71$$

$$\text{Cupo Total TDC} = 4.000.000 \quad Z_2 = 0,89$$

$$\text{Antigüedad Primer Producto} = 46 \quad Z_3 = 0,98$$

Luego, de (2) se tiene,

$$H_{11} = \text{TanHip}\{-0,68 + (-0,56 * Z_1 + 1,47 * Z_2 + 1,64 * Z_3)\} = 0,85$$

$$H_{12} = \text{TanHip}\{2,74 + (-0,71 * Z_1 + 0,20 * Z_2 - 0,45 * Z_3)\} = 0,70$$

$$H_{13} = \text{TanHip}\{-0,94 + (-2,13 * Z_1 - 1,11 * Z_2 - 0,36 * Z_3)\} = -0,99$$

Al aplicar (1), el pronóstico de ingreso para este caso es de \$5.389.417,

$$\tilde{Y} = 5.082.416 + 783.453 * H_{11} - 1.615.191 * H_{12} - 783.453 * H_{13} = \mathbf{5.389.417}$$

La comparación entre estos dos casos evidencia también el papel de la función de activación (tangente hiperbólica) en el último paso del cálculo. Aún cuando los coeficientes en la capa de salida son los mismos, la forma en que se incorporan los insumos conlleva a potencializar la respuesta de alguna (s) unidad (es) de procesamiento y, por ende, a modificar la respuesta.

Ahora bien, el modelo sugiere que cuanto mayor sea el tamaño del portafolio de un titular, mayor es también la probabilidad de ser asignado a un mayor nivel de ingreso, lo cual puede generar cambios repentinos y grandes en la respuesta ante modificaciones en los impulsos. Para mejorar las condiciones de estabilidad del pronóstico, en la construcción del modelo se definieron dos condiciones ad-hoc (insumos homogéneos): la antigüedad mínima de los productos que hacen parte del portafolio de un titular es de 1 año y, sólo hace parte de las variables de entrada, la información relativa a los productos que el individuo ha mantenido al día. De esta manera, sólo hace parte de la estimación de ingresos, el portafolio que cada individuo ha demostrado soportar efectivamente.

2.2.2 Desempeño

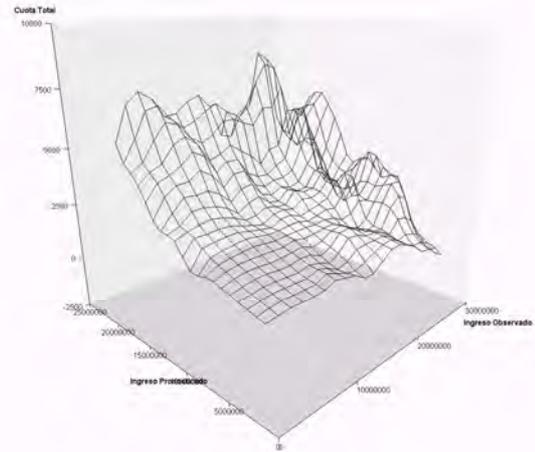
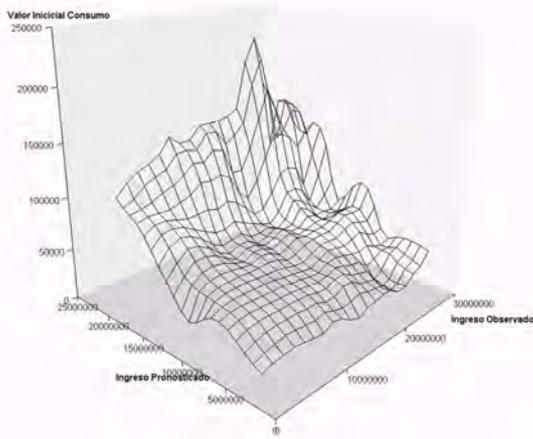
La metodología permite establecer relaciones bien definidas entre el ingreso observado, el ingreso estimado y variables relativas al gasto financiero. En el segmento de cupos altos, por ejemplo, se presenta una asociación directa entre renta observada, ingreso estimado y los desembolsos en consumo (Figura 8) panel A, o las cuotas totales (panel B). Se evidencia asimismo que la relación no es lineal; los desembolsos y las cuotas aumentan en proporción mayor que la renta¹³.

¹³ Ello refleja entre otros, un hecho estilizado en la colocación de cartera – que se confirma en la Encuesta de Ingresos y Gastos DANE: a mayor ingreso, mayor capacidad e ahorro y, por ende, mayor capacidad e endeudamiento.

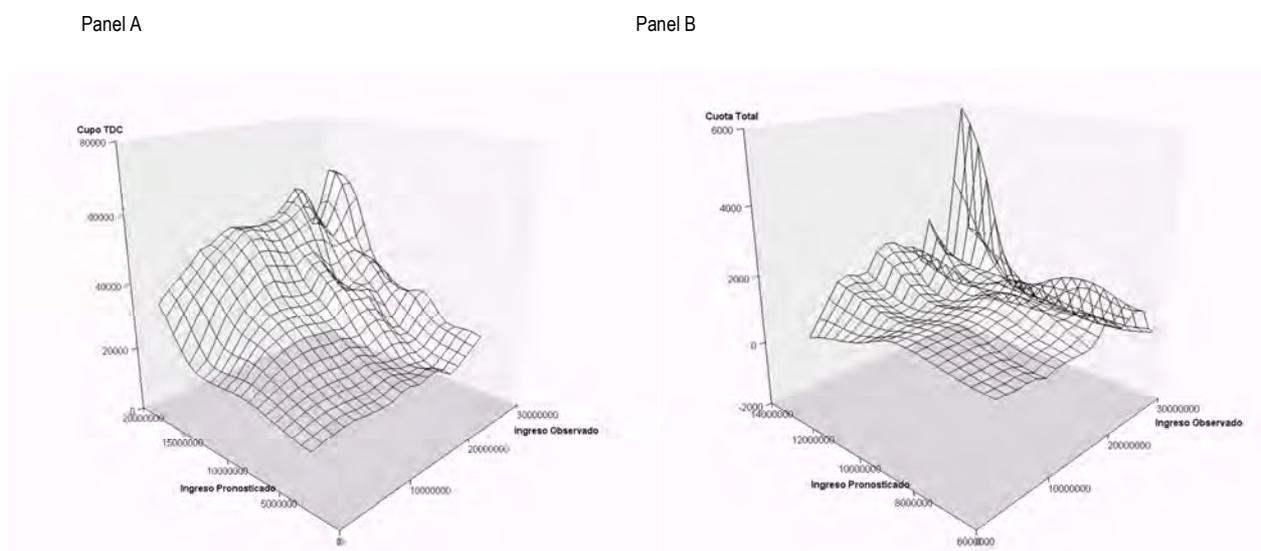
Figura 8. Tramas de superficie para Ingreso observado, probable y Valor Inicial Consumo / Cuota Total

Panel A

Panel B



El crecimiento exponencial de valor de las cuotas frente a la renta se puede evidenciar en otros nodos, como el correspondiente a desembolsos altos (Figura 9). La variación de los cupos asignados en tarjetas de crédito resulta, en cambio, proporcional al ingreso.

Figura 9. Tramas de superficie para Ingreso observado, probable y Cupo - Cuota Total

Estos resultados tienen profundas implicaciones para la comprensión de la renta de los individuos desde el gasto y justifican asimismo la metodología implementada por Asobancaria – CIFIN: la relación entre el gasto financiero y la renta no es lineal y varía en función, tanto del ingreso como del perfil crediticio de cada individuo.

La Tabla 1 resume los principales indicadores de desempeño de la metodología para cada nodo. Las estimaciones son consistentes en todos los casos; no obstante, los resultados son eficientes sólo en algunos. En términos generales, la función de distribución de ingreso estimada tiende a concentrarse en la media, mientras la observada es mucho más plana.

En los rangos de mayor ingreso, la metodología acierta en el promedio y se aproxima también a la varianza. Esto se debe a que la volatilidad de la renta es menor en estos segmentos, pero la estimación es más eficiente en virtud de un mayor poblamiento (y homogeneidad) de variables explicativas. En términos del modelo, las observaciones se agrupan en nodos bien definidos (baja dispersión) y se reporta una mayor cantidad de información financiera para el entrenamiento de las redes (clientes con mayores portafolios). En el nodo 57, por ejemplo, se obtiene una estimación ajustada de la media y, en la medida en que la volatilidad observada de la renta es baja, los coeficientes de variación estimados son también adecuados (39% estimado frente a 48% observado).

En segmentos que implican una menor utilización de productos y servicios financieros, la dispersión observada es mucho mayor y la información se encuentra restringida. En estos casos, el desempeño de la herramienta resulta diezmado y su capacidad para aproximar las variaciones en el ingreso, disminuye. En el nodo 38, por ejemplo, el ingreso promedio es igual a \$1,1 millones tanto observado como estimado. El coeficiente de variación en cambio, es igual a 58% en el observado y de 24% en el estimado.

Tabla 1. Resumen estadísticos por nodo final

Nodo	N	Promedio Observado	Coefficiente Variación Observado	Promedio Pronóstico	Coefficiente Variación Pronóstico	RMSE
33	21.827	801.748	39%	791.674	12%	267.103
34	22.466	909.972	41%	909.142	14%	317.226
40	22.108	919.090	49%	912.450	14%	367.246
41	23.693	952.065	44%	953.516	19%	358.002
25	28.114	975.241	43%	957.843	14%	354.696
36	26.514	983.125	41%	975.199	13%	344.591
24	39.713	1.005.406	42%	1.007.526	15%	359.360
38	44.290	1.059.893	58%	1.058.114	24%	489.594
46	21.824	1.119.319	48%	1.110.574	14%	452.917
47	21.946	1.231.903	48%	1.228.650	17%	500.630
72	18.365	1.232.995	48%	1.236.282	32%	464.033
27	39.710	1.239.809	46%	1.234.710	21%	485.560
45	20.355	1.242.648	48%	1.234.303	21%	487.576
39	37.724	1.259.738	65%	1.282.820	21%	655.006
30	23.107	1.311.069	44%	1.313.220	18%	470.103
44	22.823	1.325.368	55%	1.322.008	21%	589.634
22	20.189	1.449.516	67%	1.433.752	14%	794.049
43	22.608	1.483.910	63%	1.468.832	20%	752.980
12	30.496	1.533.844	41%	1.532.758	20%	513.839
42	22.191	1.552.112	65%	1.520.530	19%	799.733
73	10.544	1.579.947	59%	1.542.856	16%	763.553
56	5.872	1.591.164	57%	1.594.796	31%	738.538
21	32.900	1.641.478	58%	1.629.809	14%	772.089
77	17.418	1.800.371	68%	1.817.178	8%	1.026.316
81	13.982	1.813.759	65%	1.823.800	32%	927.046
67	5.142	1.832.971	63%	1.839.896	36%	929.634
74	17.319	1.862.478	55%	1.867.629	29%	827.158
55	6.415	1.984.343	60%	1.959.983	26%	963.893
7	22.420	2.068.490	55%	2.064.896	28%	938.264
11	23.107	2.086.048	50%	2.067.012	21%	853.842
78	10.019	2.133.579	63%	2.131.904	23%	1.100.332
75	16.368	2.144.626	52%	2.229.421	32%	899.346
66	5.180	2.240.232	57%	2.247.926	29%	1.070.833
14	20.270	2.291.149	40%	2.289.319	27%	714.138
76	10.784	2.306.765	46%	2.290.251	28%	816.003

Nodo	N	Promedio Observado	Coefficiente Variación Observado	Promedio Pronostico	Coefficiente Variación Pronostico	RMSE
53	10.263	2.480.940	65%	2.478.739	20%	1.296.463
13	23.186	2.549.396	44%	2.546.911	29%	889.659
49	8.181	2.587.335	52%	2.605.853	23%	1.094.926
54	6.239	2.615.656	48%	2.594.053	28%	939.123
4	17.885	2.769.283	41%	2.835.484	24%	909.202
68	5.473	2.966.752	54%	2.999.901	20%	1.320.240
2	30.723	3.240.424	62%	3.220.246	28%	1.614.403
3	16.086	3.383.867	46%	3.436.751	23%	1.277.815
50	5.841	3.404.721	43%	3.359.642	18%	1.261.334
71	11.268	3.682.369	64%	3.506.064	18%	1.940.542
51	5.059	4.035.374	55%	3.989.797	26%	1.863.598
69	6.055	4.099.311	51%	4.113.785	30%	1.679.461
52	5.167	4.351.075	61%	4.305.295	22%	2.227.198
60	8.805	4.429.067	36%	4.353.899	13%	1.368.521
1	13.958	4.595.868	54%	4.650.366	38%	1.838.209
70	12.212	4.668.435	48%	4.666.190	40%	1.638.590
80	17.619	4.793.911	51%	4.779.459	41%	1.740.379
63	9.074	5.163.904	60%	5.276.021	32%	2.445.000
59	7.403	5.533.311	46%	5.461.245	21%	2.118.966
79	8.164	6.071.939	48%	6.079.344	39%	2.071.358
61	5.720	6.240.824	40%	6.164.242	18%	2.145.385
64	5.129	6.544.643	60%	6.555.279	38%	2.974.985
48	5.855	6.722.923	47%	6.764.795	33%	2.320.303
57	6.426	8.045.679	48%	7.997.838	39%	2.473.793
62	6.048	8.437.240	39%	8.456.032	22%	2.773.641
65	8.107	8.732.924	54%	8.718.336	35%	3.485.990
58	5.563	9.804.355	47%	10.016.954	32%	3.219.663

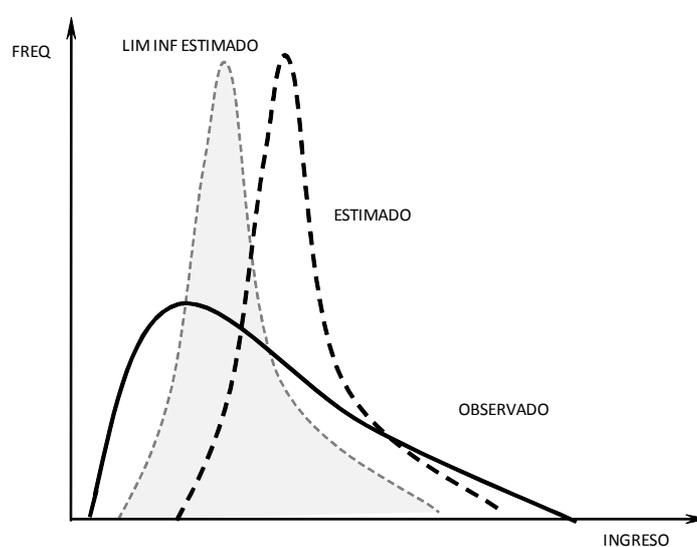
En general, grupos homogéneos (y con información suficiente) permiten estimaciones más eficientes. En segmentos con menor información (por ejemplo, baja utilización de productos y servicios financieros), el modelo enfrenta importantes restricciones. Ello propone dos implicaciones importantes para el modelo: en primer lugar, se justifican sendos procesos de segmentación previos al entrenamiento de modelos, en la medida en que ello contribuye a la conformación de grupos con menor volatilidad; en segundo lugar, los resultados son más confiables para clientes con un historial crediticio consolidado, en la medida en que ello permite establecer relaciones concretas entre gasto financiero y renta individual.

2.2.3 Ingreso Mínimo Probable

El hecho que la metodología garantice en el agregado una alta conformidad entre el ingreso promedio, observado e inferido, pero presente diferencias importantes en volatilidad, hace que se registren desviaciones considerables en el estimativo de renta a nivel individual, lo que implica un alto riesgo de sobreestimación de la capacidad de pago de los titulares¹⁴.

Con el fin de controlar esta ineficiencia y elevar la confiabilidad de los indicadores construidos a partir de dicha información, se procedió a conformar en cada nodo, un intervalo que consolidara el 90% de las observaciones, para seleccionar el límite inferior como la renta estimada (ajustada). Ello equivale a trasladar la distribución estimada hacia la izquierda, reduciendo el número de observaciones que, en cada nodo, habrían resultado por encima del nivel observado (el factor aplicado garantizara que al menos un 90% de las observaciones registre un ingreso observado mayor o igual que el inferido – Figura 10).

Figura 10. Desplazamiento de la distribución del ingreso estimado.



Debido a este ajuste, y a la salvedad que se ha reiterado en el presente documento en el sentido de una aproximación a la renta desde el gasto (en este caso, financiero), la metodología reporta al final, un Ingreso Mínimo Probable. De esta manera, el ingreso mínimo probable subestima generalmente al

¹⁴ El riesgo a subestimar es igualmente probable, mas se tiende a asumir este riesgo y no el de sobreestimación, ya que la metodología aplica a decisiones de otorgamiento de créditos a nivel individual, donde éste último sesgo puede llevar a sobreendeudamiento y riesgo de impago de los titulares.

observado en un grado que varía para cada segmento. En consecuencia, el grado de subestimación depende del nivel de la renta y de la cantidad de información crediticia de los titulares.

3 Resultados

A continuación se presenta la distribución del ingreso observado¹⁵ y el ingreso pronosticado por el modelo. De acuerdo con la metodología se espera que, en un 3% de los casos, el ingreso mínimo probable exceda significativamente (sobreestime en más de un 40%) el ingreso observado. Igualmente se espera que en un 25% de los casos, el estimado implique una subestimación severa (menor o igual que el 60% del observado – Tabla 2).

Tabla 2. Ingreso Observado vs Ingreso Mínimo Probable

Ingreso Observado	Ingresos Pronosticado					Total
	<= 700.000	700.000 - 1.000.000	1.000.000 - 1.700.000	1.700.000 - 3.000.000	> 3.000.000	
<= 700.000	192.831	27.212	4.338	376	7	224.764
700.000 - 1.000.000	121.535	45.032	12.486	834	21	179.908
1.000.000 - 1.700.000	91.433	62.544	52.082	8.069	94	214.222
1.700.000 - 3.000.000	43.021	43.337	68.239	49.368	4.124	208.089
> 3.000.000	14.578	16.397	34.116	54.227	64.023	183.341
Total	463.398	194.522	171.261	112.874	68.269	1.010.324

De otra parte, la distribución resultante del ingreso mínimo probable se aproxima a lo esperado: el sector financiero acumula individuos con ingreso (mínimo probable) superior a \$1 millón. Las personas con ingresos más bajos, acceden al crédito por medio del sector real (Tabla 3).

Tabla 3. Ingreso Mínimo Probable S Financiero y S Real

¹⁵ Corresponde a la muestra de Ingresos para el desarrollo del modelo, remitida por diferentes entidades del sector financiero Colombiano.

INGRESO MÍNIMO PROBABLE	Titulares S Financiero		Titulares S Real	
		%		%
< \$ 470.000	333,117	4.5%	778,094	10.2%
\$ 470.000 - \$ 530.000	451,021	6.1%	681,462	8.9%
\$ 530.000 - \$ 600.000	537,869	7.2%	732,992	9.6%
\$ 600.000 - \$ 660.000	556,228	7.5%	866,814	11.3%
\$ 660.000 - \$ 710.000	519,073	7.0%	824,033	10.8%
\$ 710.000 - \$ 800.000	652,441	8.8%	899,073	11.7%
\$ 800.000 - \$ 940.000	671,186	9.0%	637,530	8.3%
\$ 940.000 - \$ 1.400.000	1,115,594	15.0%	692,146	9.0%
\$ 1.400.000 - \$ 2.550.000	1,271,567	17.1%	717,894	9.4%
> 2.550.000	1,328,927	17.9%	827,345	10.8%
TOTAL	7,437,023	100%	7,657,383	100%

El ingreso medio de los clientes del sector financiero supera el de los usuarios de productos y servicios crediticios en el sector real, debido a dos factores: *i.* el acceso a servicios financieros ha estado efectivamente orientado hacia clientes con mayores niveles de ingreso y capacidad de pago y *ii.* la metodología aproxima el ingreso mediante el gasto financiero, que en el sector financiero es mayor.

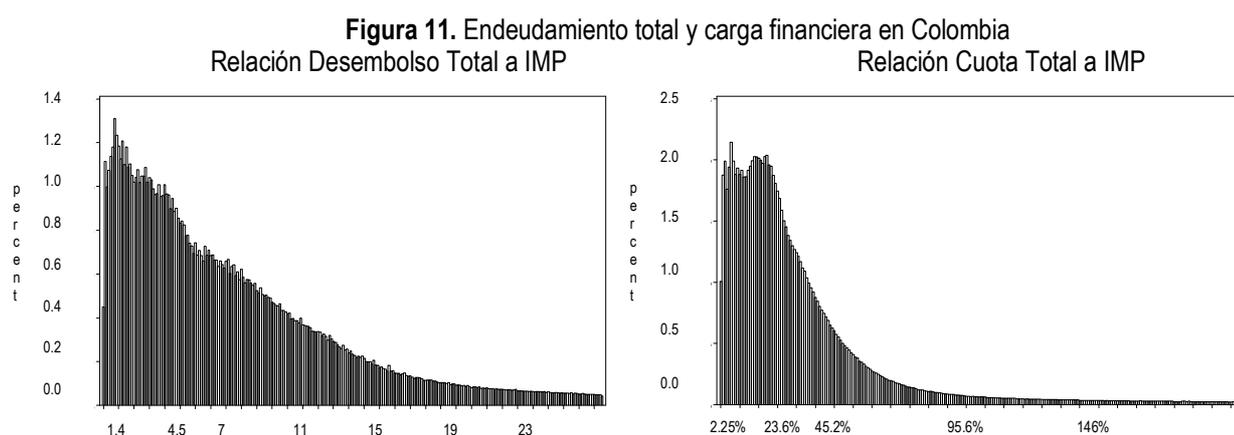
4 Endeudamiento, capacidad y carga financiera

Indicadores de capacidad de pago, reflejan la facultad que tiene un individuo (o sistema) para soportar su nivel actual de endeudamiento y su sostenibilidad en el largo plazo. De este tipo de indicadores se desprenden decisiones en lo micro y en lo macro, que van desde la adquisición de deuda adicional para la compra de un nuevo sofá, hasta modificaciones de la postura monetaria de un país ante posibles desbordes en el gasto financiero de los hogares que puedan afectar la estabilidad del sistema. Pueden considerarse, al menos, tres tipos de razones: endeudamiento, carga financiera y capacidad de endeudamiento.

El endeudamiento refleja el tamaño del portafolio como proporción del ingreso personal, y se calcula como la relación saldo (o desembolso) a ingreso (mínimo probable). Este indicador puede estar asociado a desembolsos en consumo, a los cupos en tarjetas de crédito o a los saldos en vivienda, por ejemplo. De acuerdo con el modelo, un 50% de la población en Colombia¹⁶ tiene un portafolio cuyo valor inicial representa hasta 5,7 veces su ingreso (mínimo probable - IMP) y en un 95% de los casos, este indicador es inferior a 19.

¹⁶ Reportada a la base de datos Asobancaria – CIFIN.

La carga financiera mide la incidencia que tienen los pagos que regularmente realiza un individuo por concepto del servicio de sus obligaciones sobre su flujo de ingreso. Se define como la relación cuota a ingreso y puede ser también especializado por modalidad, línea de crédito o portafolio del individuo. Esta razón se encuentra, en promedio, en 28% (Figura 11)¹⁷. De acuerdo con el modelo se obtiene además que un 10% de los titulares de crédito en Colombia, puede presentar una carga excesiva (que representaría más de 65% de la renta mensual).



La capacidad de endeudamiento comprende el gasto financiero límite que presenta un individuo, de acuerdo con su ingreso disponible. El ingreso disponible es definido como la renta después de gastos, personal y familiar (alimentación, vivienda, impuestos, entre otros). El gasto depende de variables socio-demográficas como el tamaño del hogar, el tipo de vivienda o el nivel educativo del individuo. No obstante, esta información sólo se encuentra disponible de manera agregada a partir de la Encuesta de Ingresos y Gastos DANE (2006 – 2007)¹⁸. En consecuencia, la metodología asigna a cada individuo un porcentaje promedio dependiendo de su nivel de renta (Tabla 4)¹⁹. Este porcentaje es aplicado al ingreso mínimo probable para obtener el Ingreso Disponible.

Tabla 4. Resumen Encuesta Nacional de Ingresos y Gastos 2006 - 2007

Gasto según Nivel de Ingreso Unidad de Gasto – Promedio por Ocupado

¹⁷ Información Correspondiente al Mes de Septiembre de 2010.

¹⁸ La Encuesta investiga la estructura de gasto de los hogares de acuerdo con niveles de ingreso agregado de sus miembros. Las consultas se encuentran clasificadas (a nivel agregado) según fuentes de ingreso, ocupación, nivel Educativo, categoría ocupacional, número de ocupados de la Unidad de Gasto, perceptores de ingreso. El gasto promedio por persona ocupada por franja de ingreso, es estimada a partir de los datos agregados de tamaño medio del hogar, tasa de participación y tasa de ocupación por rango de ingreso.

¹⁹ Los gastos excluyen el rubro de gasto financiero que es obtenido, a nivel de individuo, de la información reportada a la base de Datos de Asobancaria - CIFIN.

Niveles de Ingreso de la Unidad de Gasto	Total Hogares	Total Personas	Estimado Número de Ocupados	Ingreso Total Promedio x Ocupado	Gasto Total Promedio x Ocupado *	Excedente G Financiero y Otros	Porcentaje Neto Disponible Aplicable
0 - 1 S.M.L.V	2.570.210	8.312.360	3.302.512	307.261	322.773	(15.512)	-5%
1 - 2 S.M.L.V	3.219.584	12.383.253	4.919.883	544.623	445.429	99.194	18%
2 - 3 S.M.L.V	1.778.797	7.169.841	2.848.587	842.302	602.424	239.878	28%
3 - 4 S.M.L.V	1.066.484	4.528.967	1.799.365	1.092.210	716.421	375.789	34%
4 - 5 S.M.L.V	686.689	2.867.455	1.139.244	1.407.990	861.099	546.891	39%
5 - 6 S.M.L.V	469.594	1.933.669	768.249	1.724.591	1.034.754	689.837	40%
6 - 7 S.M.L.V	325.301	1.348.404	535.723	2.011.157	1.188.061	823.096	41%
7 - 8 S.M.L.V	234.349	948.061	351.524	2.557.707	1.500.064	1.057.643	41%
8 - 9 S.M.L.V	139.869	548.823	209.804	2.904.081	1.713.408	1.190.673	41%
9 - 10 S.M.L.V	131.120	501.647	169.921	3.738.526	2.205.730	1.532.796	41%
10 - 14 S.M.L.V	156.856	569.064	256.960	5.379.542	2.528.385	2.851.157	53%
> 15 S.M.L.V	365.997	1.327.782	306.876	7.272.404	3.413.475	3.858.929	53%
TOTAL	11.144.850	42.439.326	16.608.648				

FUENTE; DANE – Encuesta Nacional de Ingresos y Gastos 2006 – 2007.

*Corresponde a (Ingreso Total del Hogar / Personas Ocupadas por Hogar).

Nota. El ingreso total incluye la imputación de gasto de vivienda realizada por el DANE en la encuesta nacional de Ingresos y Gastos 2006 – 2007.

Posteriormente se suman todas las cuotas asociadas a cada una de las obligaciones financieras vigentes que ostenta cada titular. Este valor es descontado del ingreso disponible, para obtener la capacidad de endeudamiento individual. Formalmente,

$$Capacidad\ Edeudamiento = Disponible - Total\ Cuotas = \bar{Y} * \%DANE - Total\ Cuotas$$

Donde %DANE se refiere al porcentaje de gasto (personal y familiar) promedio, asociado al nivel de ingreso inferido. El cálculo de las cuotas depende de la naturaleza del pasivo del titular: en fijos, las cuotas corresponden a la sumatoria de los pagos mensuales reportados por cada entidad acreedora. En cupos, estas son calculadas a partir del cupo total, independientemente de su utilización o saldo²⁰.

A la luz de la presente metodología, este cálculo refleja la relación entre capacidad de gasto revelada y carga financiera actual. Un valor positivo implica, con elevados niveles de confianza (90%), que el titular ha demostrado una capacidad de gasto superior a su carga financiera corriente. Un valor negativo indica que las cuotas exceden la capacidad de pago revelada o, lo que es lo mismo, probable sobreendeudamiento de la contraparte.

De 13,4 millones de titulares con información en Julio y Septiembre de 2010, un 11,1% puede evidenciar excesos de endeudamiento (capacidad de endeudamiento negativa) y un 11,9% adicional, está muy cerca de él (capacidad de endeudamiento muy baja - Tabla 5).

²⁰ En este sentido, el valor de la cuota reportada es aquel que el titular pudiera registrar en el evento en que utilizara la totalidad de su disponible.

Tabla 5. Evolución de la Capacidad de Endeudamiento (% Ingreso Mínimo Probable)
- Cantidad de Titulares

Razón Capacidad / (IMP)	Capacidad de Endeudamiento			
	Julio 2010	% Titulares	Septiembre 2010	% Titulares
< - 10 %	1.530.662	11,40%	1.489.366	11,09%
- 10 % - 10 %	1.606.644	11,97%	1.603.593	11,94%
10 % - 20 %	1.576.552	11,74%	1.585.614	11,81%
20 % - 25 %	1.385.238	10,32%	1.399.934	10,43%
25 % - 30 %	2.454.385	18,28%	2.465.125	18,36%
30 % - 33 %	1.902.588	14,17%	1.913.791	14,25%
33 % - 37 %	1.811.152	13,49%	1.817.039	13,53%
37 % - 40 %	595.279	4,43%	595.033	4,43%
40 % - 42 %	143.290	1,07%	143.016	1,07%
> 42 %	421.674	3,14%	414.953	3,09%
TOTAL	13.427.464		13.427.464	

Aunque este porcentaje permanece relativamente estable en el corto plazo, el seguimiento a esta variable en el nivel agregado así como la gestión sobre estos titulares en el nivel micro, constituyen la premisa para preservar la estabilidad financiera a partir, en este caso, de un oportuno control al endeudamiento de los usuarios del crédito. El principal valor de la metodología presentada es aportar a esta tarea.

Por supuesto, es importante tener presente que se trata de un estimativo (con error estadístico) que opta además por un límite inferior (que subestima la verdadera renta). Por ello, más que los valores absolutos asociados a los límites de endeudamiento, la aplicabilidad del análisis basado en esta metodología radica especialmente en el seguimiento a su comportamiento en el tiempo, que permita identificar tendencias y cambios en el corto plazo, contemplando la totalidad de la cartera, a partir de datos a nivel de individuo.

5 Conclusión

Excesos de endeudamiento pueden ocasionar deterioros de cartera, en la medida en que los balances individuales pueden resultar insostenibles. De presentarse excesos de endeudamiento para un número creciente de personas, es posible que sus efectos se transmitan a todo el sistema. Por esta razón, el monitoreo robusto y permanente a indicadores de endeudamiento contribuyen a un entorno financiero controlado y un crecimiento del crédito, saludable.

Como contribución a este sistema de monitoreo, Asobancaria – CIFIN ha diseñado una metodología para la inferencia de ingresos a partir de información crediticia de los individuos que contribuye a la construcción de indicadores micro-fundamentados sobre límites de endeudamiento. La metodología ha permitido el cálculo de carga financiera para 16 millones de titulares.

De acuerdo con el modelo, el portafolio (inicial) medio de la población en Colombia representa 5,7 veces su ingreso. La carga financiera, definida como la relación cuota a ingreso se encuentra, en promedio, en 28%. Cerca de 2 millones de personas registran una baja carga financiera pero aproximadamente un 10% de la población, puede presentar niveles excesivos de endeudamiento.

6 REFERENCIAS

ANNELI, Juntto. The comparability of imputed rent in EU-SILC 2007– differences in variable definitions and methods concerning institutional housing. Luxemburgo; Eurostat Methodologies and Working Papers, 2010

ARISTIZÁBAL, M. MISAS, M. Evaluación asimétrica de una red neuronal artificial: Una aplicación al caso de la inflación en Colombia. Borradores de economía. Banco de la Republica. Bogotá. 2006.

BALLESTEROS, A. Neural Networks Framework, Trabajo de grado, Departamento de ingeniería. University of Malaga, Spain. 2010.

BENIN, S. RANDRIAMAMONJY, J. Estimating Household Income to Monitor and Evaluate Public Investment Programs in Sub-Saharan Africa. IFPRI Discussion Papers 771. 2008.

BRANNAN, Johnston. Right Income Model Is Key. Experian, American Banker. New York. 2010.

CLEVELAND, Robert. Alternative Income Estimates in the United States. U.S. Department of Commerce, Economics and Statistics Administration, Census Bureau. 2005.

CHASCO, Yrigoyen. An Estimation of Disposable Personal Income of the Spanish Municipalities in 1997. Dublín; European Regional Science Association (ERSA), 1999.

CHOTIKAPANICH, D, GRIFFITHS, W, RAO, D. Estimating and Combining National Income Distributions using Limited Data. Journal of Business & Economic Statistics, American Statistical Association, vol. 25, pages 97-109, 2007.

COOPER, Daniel. Imputing Household Spending in the Panel Study of Income Dynamics: A Comparison of Approaches. Working papers, Federal Reserve Bank of Boston, 2010

Departamento Nacional de Estadística - DANE. Encuesta de Ingresos y Gastos 2006 – 2007 (Resumen online:

http://www.dane.gov.co/daneweb_V09/index.php?option=com_content&view=article&id=114&Itemid=66)

DEY, Shubhamoy. Modeling the combined effects of credit limit management and pricing actions on profitability of credit card operations. International journal of business and management Vol 5 No. 4, 2010.

FAY, R. TRAIN, G. Aspects of Survey and Model-based Postcensal Estimation of Income and Poverty Characteristics for States and Counties. American Statistical Association Conference Annual Meeting. Orlando, 1995.

<http://www.census.gov/did/www/saipe/publications/files/FayTrain95.pdf>

- FERGUSON, A. TANDON, E. GAKIDOU y MURRAY. Estimating Permanent Income Using Indicator Variables. World Health Organization, Evidence and Information for Policy Cluster. Geneva. 2003.
- FISHER, Jonathan D. Income Imputation and the Analysis of Consumer Expenditure Data. Bureau of Labour Statistics, Working Paper 394, 2006.
- FISHER, Robin. Methods Used For Small Area Poverty And Income Estimation. US Bureau of the Census, Washington, DC. 1996.
- FLETCHER, A. HAMILTON, T. Scoring and Revenue Estimation. Briefing Paper No. 5. Harvard Law School. 2008.
- FRICK, J, GRABKA, M. The Personal Distribution of Income and Imputed Rent. A Cross-National Comparison for the UK, West Germany, and the USA. German Institute for economic research. Discussion Papers 271. Berlin, 2002.
- GALLUP, John. Consistent Estimation of Income Inequality. Portland State University. Portland. 2010. http://www.mth.pdx.edu/~fountain/seminar/consistent_inequality.pdf
- GERTEL, et al. An Estimation of Personal income Distribution Using the Dagum Model With an Application to Cordoba 1992-2000. Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, 2001. <http://cdi.mecon.gov.ar/biblio/docelec/utdt/pob2-5.pdf>
- HSIAO-TIEN, Pao. A comparison of neural network and multiple regression analysis in modeling capital structure. Department of Management Science, National Chiao Tung University. Taiwan, 2008.
- KITOV, Ivan. Mechanical Model of Personal Income Distribution. Society for the study of economic inequality. Working Paper Series 110. 2009.
- KONOVALIKHIN, M. KULIK, V. SERGIENKO, D. BALAEVA, O. The scaling model for credit limit management. World congress on engineering, Londres, 2008.
- KYOBE, A. DANNINGER, S. Revenue Forecasting - How is it done? Results from a Survey of Low-Income Countries. IMF Working Paper WP/05/24. 2005.
- LEIPPOLD, M. EBNOETHER, S. VANINI, P. Optimal Credit Limit Management. National Centre of competence in Research – Financial valuation and Risk Management. Working Paper Series 72. 2003.
- MANDAL, B. STASNY, E. Imputing Missing Income Data and Weighting Data with Imputed Income. Business Meeting, American Statistical Association, Ohio State University. August, 2003. <http://www.amstat.org/sections/srms/proceedings/y2004/files/Jsm2004-000173.pdf>

- MATIGNON, Randall. Data Mining using SAS Enterprise Miner: Model Nodes. Wiley. New Jersey, 2007.
- MATIGNON, Randall. Neural Network Modeling using SAS Enterprise Miner: Neural Network Architecture. Lexington, Authorhouse. 2010.
- MONTAÑO, Juan J. Redes neuronales artificiales: abriendo la caja negra. Tesis Doctoral. Universitat de les Illes Balears, Facultat de Psicologia. Palma de Mallorca. 2002.
- MORALES, I. Comparación Teórico Práctica entre Modelos Estadísticos y el Perceptrón Multicapa. Trabajo de grado, Pontificia Universidad de Valparaíso, Facultad de ciencias, Instituto de Estadística. Valparaíso. 2010.
- MUNIR A. MISAS M. Evaluación de pronósticos del tipo de cambio utilizando redes neuronales y funciones de pérdida asimétricas. Revista Colombiana de Estadística. Bogotá. 2007.
- PIA, M y FESER, Victoria. Robust estimation of personal income distribution models Distributional Analysis Research Programmer, Discussion Paper. 1993.
- PITARQUE, A. ROY, J. RUIZ, J. Redes neurales vs modelos estadísticos: Simulaciones sobre tareas de predicción y clasificación. International Journal of methodology and experimental Psychology. Universidad de Valencia. Valencia. 1998.
- SERRANO, A. SORIA, E. MARTIN, J. Redes Neuronales Artificiales. Guía Curso 2009-2010, Escuela Técnica superior de ingeniería, Departamento de ingeniería y electrónica. Valencia. 2009.
- SCHENKER, et al. Multiple Imputation of Missing Income Data in the National Health Interview Survey Nathaniel. Journal of the American Statistical Association, 2006.
- SMITH, M. Neural Networks for Statistical Modelling. Van Nostrand Reinhold. New York. 1993.
- STERGIOU, C, SIGANOS D. Neural Networks. Imperial College London. Surprise 96 Journal. Londres, 1998.
- STIGLITZ, Joseph. Caída Libre. Taurus. Madrid, 2010.
- WANG, Yibing. A Nonparametric Analysis of the Personal Income Distribution across the Provinces and States in the U.S. And Canada. Regional and Sectoral Economic Studies, AEEADE. Vol 4-1, 2004.

ANEXO 1. Composición y orden de las variables de discriminación para el segmento Desembolso de Crédito Comercial.

Crédito Comercial					
Cuota < 1.606			Cuota >= 1.606		
Cuentas Corrientes = 0			Cuentas Corrientes > 0	Cuota Total < 963	Cuota Total >= 963
Valor Inicial Total < 10.807		Valor Inicial Total >= 10.807 y < 21.850	Valor Inicial Total >= 21.850		
Saldo Comercial = 0	Saldo Comercial > 0				

ANEXO 2. Composición y orden de las variables de discriminación para la Subpoblacion Cupo TDC Superior a 10 Millones.

Cupo TDC Superior a 10 Millones				
Valor Inicial Total < 44847			Valor Inicial Total >= 44.847	
Cupo TDC < 14299	Cupo TDC >= 14299 < 18000	Cupo TDC >= 18.000	Días Sobregiro < 28	Días Sobregiro >= 28
Cuentas Corrientes < 5	Cuentas Corrientes >= 5			

ANEXO 3. Composición y orden de las variables de discriminación para la Subpoblacion Cupo TDC 5 Millones hasta 10 Millones.

Cupo TDC 5 Millones hasta 10 Millones			
Cuentas Corrientes = 0		Cuentas Corrientes > 0	
Cantidad TDC <= 2	Cantidad TDC > 2		

ANEXO 4. Composición y orden de las variables de discriminación para la Subpoblacion Registro patrimonial de Inmuebles.

Registro patrimonial de Inmuebles					
Cuota < 594					Cuota >= 594
Cuentas Corrientes = 0					Cuentas Corrientes > 0
Estrato 1 y 2	Valor Inicial Total = 0	Valor Inicial Total > 0 y <= 1500	Valor Inicial Total > 14000	Valor Inicial Total > 1500 y <= 5000	Valor Inicial Total > 5000 y <= 14000

Tiempo Apertura SF <= 1000 Tiempo Apertura SF > 1000

ANEXO 5. Composición y orden de las variables de discriminación para la Subpoblación Registro patrimonial de Vehículos.

Registro patrimonial de Vehículo					
Cuota < 925				Cuota >= 925	
Cuentas Corrientes = 0			Cuentas Corrientes > 0		
Valor inicial Total = 0	Valor inicial Total > 7849	Valor inicial Total > 0 y <= 7849		Variación TDC (3) < 64	Variación TDC (3) >= 64
		Saldo TDC (12) < 239	Saldo TDC (12) >= 239		

ANEXO 6. Composición y orden de las variables de discriminación para la Subpoblacion bajos montos y poca historia.

Bajos Montos y poca Historia *											
Cuentas Corrientes = 0									Cuentas Corrientes > 0		
Cuota <= 359					Cuota > 359						
Cupo Max <= 2700						Cupo Max > 2700	Cuota < 273	Cuota >= 273			
Saldo Total (12) <= 0				Saldo TDC (12) > 0				Cantidad Productos No rot <= 2	Cantidad Productos No rot > 2	Saldo TDC = 0	Saldo TDC > 0
Cantidad Entidades= 0			Cantidad Entidades> 0		Cupo TDC < 649			Cupo TDC >= 649			
Antigüedad primer Productos <= 1035		Antigüedad primer Productos > 1035		Valor Mora Productos Atraso = 0	Valor Mora Productos Atraso > 0	Antigüedad 1er Productos > 1035	Antigüedad 1er Productos <= 1035	Cupo < 925	Cupo >= 1829	Cupo Max >= 925 y < 1829	
Cuota TDC = 0	Cuota TDC > 0 y <= 152	Cuota TDC > 152	Cuota Consumo = 0	Cuota Consumo > 0			Productos Total <= 1	Productos Total > 1	Cantidad TDC <= 1	Cantidad TDC > 1	

*El árbol de regresión para esta Subpoblacion fue el de mayor profundidad y con mayor cantidad de variables debido a el numero de registros que lo conforman (760.000).

ANEXO 7. Consistencia entre Ingreso Observado e Ingreso Mínimo Probable

Nodo	Ingreso Mínimo Probable (IMP)			Acierto	Razón (IMP / Ingreso Observado)			
	n	Promedio	Desviación		< 0,6	0,6 - 1,2	1,2 - 1,4	> 1,4
1	13.796	2.546.589	1.048.492	89%	39%	54%	4%	3%
2	30.464	1.446.883	419.132	84%	49%	39%	4%	8%
3	15.946	1.886.221	469.050	89%	38%	56%	4%	3%
4	17.741	1.697.235	468.259	90%	30%	65%	3%	2%
7	22.343	1.031.251	298.498	86%	46%	44%	4%	6%
11	22.985	1.029.881	212.707	89%	47%	47%	4%	2%
12	30.430	918.944	196.011	92%	32%	65%	3%	0%
13	23.127	1.527.231	469.851	93%	31%	66%	3%	1%
14	20.246	1.601.334	454.907	85%	21%	72%	5%	1%
21	32.745	733.225	103.005	84%	49%	39%	6%	6%
22	20.126	645.254	93.524	76%	45%	38%	10%	7%
24	39.704	554.225	82.518	92%	34%	61%	3%	1%
25	28.092	526.344	72.047	93%	39%	58%	3%	1%
27	39.609	677.779	144.870	89%	40%	54%	4%	3%
30	23.071	721.696	130.790	90%	39%	56%	4%	1%
33	21.822	475.107	60.583	94%	32%	66%	2%	0%
34	22.460	500.101	71.856	93%	39%	58%	2%	1%
36	26.496	584.099	75.859	88%	32%	63%	4%	1%
38	44.239	528.741	128.695	88%	37%	55%	5%	3%
39	37.669	577.034	124.348	81%	42%	45%	10%	3%
40	22.098	501.600	73.930	91%	33%	63%	3%	1%
41	23.665	523.757	100.885	91%	39%	57%	4%	1%
42	22.149	684.536	130.818	80%	45%	41%	7%	7%
43	22.557	661.050	138.037	84%	46%	43%	6%	4%
44	22.772	661.003	139.080	83%	42%	48%	7%	4%
45	20.334	678.525	144.039	86%	37%	54%	5%	3%
46	21.800	611.030	88.738	85%	37%	55%	6%	2%
47	21.884	613.177	102.489	89%	45%	49%	5%	2%
48	5.832	4.733.183	1.690.046	83%	22%	69%	6%	3%
49	8.140	1.302.292	306.803	87%	47%	43%	5%	5%
50	5.805	1.845.306	326.002	91%	43%	53%	3%	1%
51	5.026	1.989.027	519.256	86%	47%	44%	4%	5%
52	5.132	1.720.373	403.956	88%	57%	34%	3%	6%
53	10.234	991.602	203.562	83%	54%	31%	5%	10%
54	6.148	1.426.009	403.933	91%	40%	54%	3%	2%
55	6.394	882.158	229.722	86%	52%	38%	5%	5%
56	5.862	876.315	277.856	82%	39%	47%	7%	7%
57	6.397	5.979.013	2.372.768	84%	18%	74%	6%	2%
58	5.524	7.002.676	2.279.997	83%	22%	69%	6%	3%
59	7.335	2.996.980	648.437	92%	41%	56%	3%	1%
60	8.749	2.608.184	352.513	96%	34%	65%	1%	0%
61	5.701	4.311.134	794.935	83%	24%	67%	6%	3%
62	6.035	5.917.812	1.317.035	86%	23%	71%	5%	1%
63	8.997	2.360.911	749.817	83%	51%	39%	4%	6%
64	5.084	3.604.828	1.420.238	84%	43%	44%	4%	8%
65	8.064	5.205.965	1.876.648	86%	35%	56%	5%	4%

Nodo	Ingreso Mínimo Probable (IMP)			Acierto	Razón (IMP / Ingreso Observado)			
	n	Promedio	Desviación		< 0,6	0,6 - 1,2	1,2 - 1,4	> 1,4
66	5.169	1.234.994	374.151	77%	39%	44%	6%	11%
67	5.137	727.807	262.084	89%	59%	35%	2%	4%
68	5.463	1.641.612	326.362	79%	39%	45%	6%	10%
69	6.012	2.250.677	683.244	85%	42%	47%	4%	6%
70	12.152	3.265.688	1.441.732	84%	22%	69%	6%	3%
71	11.210	1.753.959	310.043	79%	45%	40%	5%	11%
72	18.339	679.319	225.288	89%	39%	55%	4%	2%
73	10.508	616.241	98.595	91%	57%	38%	3%	2%
74	17.275	838.786	243.931	89%	54%	40%	4%	3%
75	16.325	1.225.479	405.218	85%	37%	52%	5%	6%
76	10.732	1.256.975	360.728	92%	42%	54%	3%	1%
77	17.396	726.812	56.357	78%	52%	30%	7%	11%
78	9.995	853.424	198.744	83%	56%	31%	5%	8%
79	8.126	4.247.987	1.737.988	83%	23%	66%	7%	4%
80	17.468	2.866.574	1.279.463	95%	32%	66%	2%	1%
81	13.953	820.016	267.398	84%	50%	38%	5%	6%
Total	1.016.089	1.207.167	1.233.074	87%	40%	52%	5%	3%

ANEXO 8. Equivalencia en la terminología estadística y de redes neuronales.

Terminología estadística	Terminología de redes neuronales
Observación	Patrón
Muestra	Datos de entrenamiento
Muestra de validación	Datos de validación, test
Variables explicativas	Variables de entrada
Variable de respuesta	Variable de salida
Modelo	Arquitectura
Residual	Error
Error aleatorio	Ruido
Estimación	Entrenamiento, aprendizaje
Interpolación	Generalización
Interacción	Conexión funcional
Coefficientes	Pesos de conexión
Constante	Peso umbral
Regresión y análisis discriminante	Aprendizaje supervisado o heteroasociación
Reducción de datos	Aprendizaje no supervisado o autoasociación
Análisis de cluster	Aprendizaje competitivo

Fuente: Tesis Doctoral Redes Neuronales Artificiales aplicadas al análisis de datos. Juan José Montaña Moreno (2002).

ANEXO 9. Equivalencia entre modelos estadísticos y modelos de red neuronal.

Modelo estadístico	Modelo de red neuronal
Regresión lineal múltiple	Perceptrón simple con función lineal
Regresión logística	Perceptrón simple con función logística
Función discriminante lineal	Perceptrón simple con función umbral
Regresión no lineal múltiple	Perceptrón multicapa con función lineal en la salida
Función discriminante no lineal	Perceptrón multicapa con función logística en la salida
Análisis de componentes principales	Regla de Oja, Perceptrón multicapa autoasociativo
Análisis de clusters	Mapas auto organizados de Kohonen
K vecinos más cercanos	Learning Vector Quantization (LVQ)
Regresión kernel	Funciones de Base Radial (RBF)

Fuente: Tesis Doctoral Redes Neuronales Artificiales aplicadas al análisis de datos. Juan José Montaña Moreno (2002).