"*CREDIT SCORING* PARA LA CARTERA CREDITICIA DE CONSUMO NO REVOLVENTE DE UNA ENTIDAD BANCARIA ESTATAL"

Trabajo de Investigación presentado
para optar al Grado Académico de
Magíster en Finanzas

Presentado por

Srta. Paola Aracelli Roxana Tamayo Medrano

2017

A mis padres por su amor, ejemplo de superación y dedicación incondicional.
A mis hermanos por su amistad y apoyo.
A Alejandra y Enrique por la inconmensurable felicidad que me brindan.

Agradezco a mis profesores de la Maestría de la Universidad del Pacífico por su valioso aporte en mi desarrollo profesional.

Resumen ejecutivo

El presente trabajo de investigación busca entender las variables determinantes de la probabilidad de *default* en la cartera de consumo no revolvente del Banco de la Nación que permita mejoras en la gestión del riesgo crediticio. Se estudiará a través del modelo de regresión logística construido con los desembolsos de la entidad entre el periodo enero a diciembre 2014.

Se obtuvo que variables cualitativas como el departamento en donde se desembolsó el crédito, el tipo de préstamo (sector al que pertenece el trabajador público que solicita el préstamo), la situación laboral del cliente y sexo; así como de las variables intrínsecas de la operación de préstamo como son el plazo, deuda en el sistema financiero además de la variable ingreso del trabajador (que considera lo que en neto ingresa a la cuenta de ahorros en el banco) y antigüedad laboral, permiten obtener un ajuste apropiado para el modelo que estima la probabilidad de *default* del potencial cliente. Para evaluar el poder de discriminatorio se utilizó el estadístico K-S y la curva ROC (*Receiver Operating Characteristics*) y área bajo la curva ROC siendo de 0,7379. Con la validación del modelo se obtuvo que el 61% de los créditos malos observados son estimados por el modelo.

Índice

Índice de tablas

Índice de gráficos

Índice de anexos

Capítulo I. Introducción

Tal como se documenta en Carrera (2011), el crédito es una de las variables claves para entender el mecanismo de transmisión de la política monetaria. Sin embargo, la calidad de cartera debería tener un espacio en dicho análisis. Los créditos generados en el Banco de la Nación (BN) forman parte del crédito agregado pero tienen características que las hacen únicas. El presente trabajo de investigación tiene como objetivo identificar las variables determinantes de la probabilidad de *default* en la cartera de consumo no revolvente del BN que sirva para mejorar la gestión del riesgo de crédito de la cartera de créditos de la entidad.

Actualmente, el BN evalúa al potencial cliente *ex ante* a través de su capacidad de endeudamiento y el cumplimiento de requisitos referidos al tipo de contrato en la entidad pública que labora, situación en el sistema financiero (número de entidades con las que mantiene deuda directa, calificación de riesgos de la Superintendencia de Banca, Seguros y AFPs) y que su cliente mantenga una cuenta de ahorros remunerativa en la institución, lo cual le asegura, hasta cierta medida, la recuperación de sus créditos, evidenciado en su bajo nivel de morosidad[1]; pero, a lo largo de la vida del crédito, el cliente tiene libertad para migrar su cuenta de ahorros hacia otras entidades del sistema, y que, al no tener una cultura de pagos saludable, afecta la recuperación del crédito, impactando en los resultados del BN, tal como se evidencia en que del total de la cartera morosa, el 81%[2] está explicado por aquellos clientes que ya no canalizan sus ingresos a través del banco.

Sin embargo, en la gestión del riesgo crediticio resulta fundamental conocer al potencial deudor que esté acorde con el perfil de riesgos de la empresa, medido a través de la cuantificación de su probabilidad de incumplimiento en el pago de sus obligaciones *ex ante* al otorgamiento del crédito a partir de variables de naturaleza cualitativa y cuantitativa que lo caracterizan.

Es así que se aborda el objeto de la investigación desde los siguientes objetivos específicos: i. estimar un modelo de regresión logística de admisión para obtener la probabilidad de *default* aplicable a la evaluación crediticia de los préstamos de consumo no revolvente del BN; y ii. identificar las variables de naturaleza cualitativa y cuantitativa que determinan el modelo.

[1] Morosidad de 1,53% frente a 2,61% de la banca múltiple a diciembre 2016. Se indica que el BN no compite con la banca múltiple debido a que sus créditos están dirigidas exclusivamente a los trabajadores del sector público, sin embargo, se presenta en comparación con la banca múltiple a manera de referencia.
[2] Dato a junio 2016.

El presente estudio toma interés en las mejoras que se pueden obtener a través de un conocimiento del cliente por la entidad otorgante del crédito y plantea la hipótesis que variables cualitativas como departamento donde se desembolsa el crédito, sector en el que labora el trabajador público, su situación laboral: activo o pensionista y sexo en combinación con variables cuantitativas como plazo del préstamo, antigüedad laboral, ingreso del trabajador, la deuda en el sistema financiero al momento de solicitar el crédito permiten estimar la probabilidad de *default* para la cartera de créditos de consumo no revolvente del BN.

Se propone abordar el tema en tres capítulos hacia adelante, en donde en el primero se realizará el desarrollo del marco teórico, en el segundo capítulo se presentará el planteamiento del problema, en el tercer capítulo se desarrollará la metodología para determinar las variables determinantes del modelo de *credit scoring*, incluyendo la determinación de la muestra, variables utilizadas en la formulación, contrastación empírica y pruebas de poder discriminatorio y validación. Finalmente, se presentarán las conclusiones y recomendaciones.

Capítulo II. Marco teórico

1. Riesgo crediticio

Durante el proceso de otorgamiento de un crédito se producen problemas de información entre el banco y el solicitante; en el presente trabajo de investigación nos referiremos a un solicitante minorista, debido a que el banco conoce menos de la situación actual y proyectos del cliente. Esta asimetría de información genera que se produzcan las siguientes fallas de mercado:

- **Selección adversa:** ocurre previo al otorgamiento del crédito y se produce porque el banco posee mayor probabilidad de otorgar un crédito al solicitante más riesgoso, teniendo en cuenta que este busca con mayor empeño un crédito, lo que hace que su probabilidad de conseguirlo sea mayor.
- **Riesgo moral:** ocurre luego del otorgamiento del crédito y está relacionado con la voluntad de pago del cliente, es decir, una vez que consiguió el financiamiento podría llevar a cabo acciones que satisfagan sus intereses dejando de lado los del banco.

La SBS, en su Resolución 3780-2011 define el riesgo crediticio como «La posibilidad de pérdidas por la incapacidad o falta de voluntad de los deudores, contrapartes, o terceros obligados, para cumplir sus obligaciones contractuales registradas dentro o fuera del balance» (SBS 2011: 2).

Tomando en consideración que detrás del concepto señalado está implícito tanto el riesgo de incumplimiento que se puede aproximar como la probabilidad que se produzca el incumplimiento, así como por el riesgo de mercado, que es la pérdida a la que la entidad bancaria se ve expuesta en caso se produzca el incumplimiento, las entidades bancarias se enfrentan a posibles pérdidas que pueden gestionar con el tratamiento de los factores que intervienen en el riesgo de crédito.

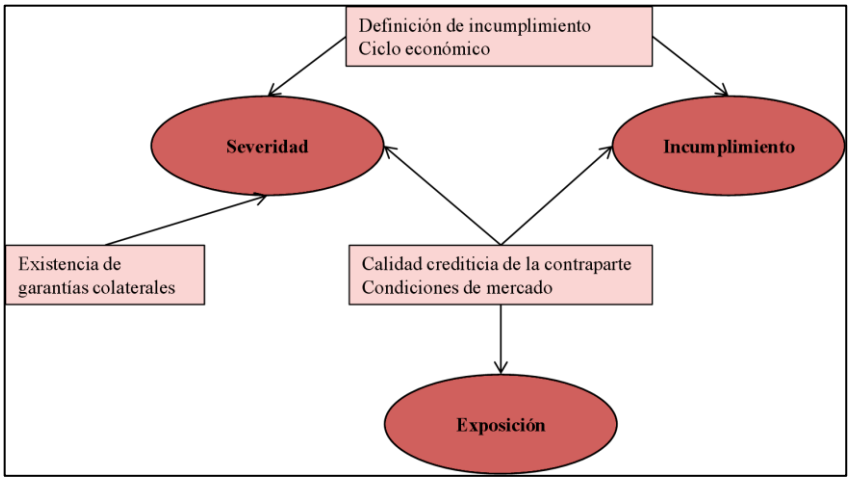
Según Samaniego (2008), los factores determinantes del riesgo de crédito se describen a continuación y se muestran en el gráfico 1.

- **Probabilidad de incumplimiento:** probabilidad que la contraparte no haga frente a sus obligaciones contractuales. En ese punto, resulta determinante la aversión al riesgo de la

entidad bancaria, la definición de incumplimiento que esta realice, entre otros determinantes.

- **Exposición:** valor de mercado de la deuda en el momento del incumplimiento. Variable que depende del instrumento, la calidad crediticia del deudor, entre otros.
- **Severidad:** porcentaje de pérdida que, tras el proceso de recuperación, y dada la pérdida, finalmente se produce. Depende de los costos de recuperación actualizados, la duración del proceso de recupero, entre otros.

Gráfico 1. Factores determinantes del riesgo crediticio



Fuente: Samaniego, 2008.

En ese mismo sentido, la SBS en su Resolución 14354-2009 señala que una empresa del sistema bancario puede postular a utilizar métodos internos para la estimación de su requerimiento de capital por riesgo de crédito para exposiciones minoristas, siempre que cumpla los requerimientos mínimos establecidos por el ente regulador para los parámetros del riesgo, siendo estos parámetros los siguientes:

- **Probabilidad de incumplimiento:** probabilidad de ocurrencia del evento de incumplimiento.
- **Exposición ante el incumplimiento:** es la mejor estimación de la exposición cuando ocurra el evento de incumplimiento. Incluye tanto a las posiciones directas como contingentes.
- **Pérdida dado el incumplimiento:** estimación del porcentaje de la exposición ante el incumplimiento que no será recuperado, producido el evento de incumplimiento. Es la pérdida económica tomando en consideración todos los factores relevantes, incluyendo efectos de descuentos importantes y costos directos e indirectos sustanciales relacionados con el cobro de la exposición.

Así, tenemos que una vez identificados y definidos los factores determinantes del riesgo de la operación que la entidad bancaria enfrenta, resulta necesario que desarrolle una de las etapas más importantes dentro de la gestión del riesgo crediticio, que es el proceso de identificación del deudor, de tal manera que *ex ante* identifique y cuantifique el riesgo de crédito, qué significa y determinar si se encuentra dentro del perfil de riesgos de la empresa y, en función a ello, y a través de herramientas de admisión crediticia, decida si aceptará o rechazará la solicitud del crédito.

Dentro de las herramientas que posibilitan una identificación *ex ante* y variables que determinan la probabilidad de *default* del deudor acorde al perfil de riesgos, se cuenta con modelos de *credit scoring* de admisión permitiendo gestionar adecuadamente el riesgo crediticio.

2. *Credit scoring*

Según Lawrence y Solomon (2002), el *scoring* es una herramienta que sirve para analizar a los solicitantes de créditos (clientes potenciales), así como a los clientes existentes, con la finalidad de predecir su comportamiento futuro, medido a través de la probabilidad de *default* del cliente, basado en las variables que la determinan. Si este se utiliza correctamente, puede proporcionar a los administradores de créditos una cantidad importante de información estadística inyectándole un alto grado de previsibilidad a sus portafolios. Además, Lawrence y Solomon indican que el *credit scoring* utiliza técnicas estadísticas para identificar y clasificar a los clientes potenciales o los clientes actuales de acuerdo a lo atractivo que resulte para la entidad financiera. Esta debe ser definida de antemano por el usuario y puede ser rentabilidad, riesgo, disposición de repago si ha caído en mora, entre otros.

Cabe indicar que existen dos tipos de *scoring*: (i) de aplicación, que es utilizado para potenciales clientes, y (ii) de comportamiento, que es generalmente aplicado para clientes vigentes, según el momento en que se aplican en la evaluación. De otro lado, también se pueden clasificar en dos tipos de *scores*: (i) customizados, siendo desarrollados a partir la propia información de la entidad, y (ii) genéricos, que son desarrollados con gran cantidad de información que vienen de oficinas proveedoras.

Partiendo de estas clasificaciones, el presente trabajo de investigación desarrollará un *scoring* de aplicación customizado.

2.1 Beneficios del *credit scoring*

En ese mismo sentido, los autores enumeran los principales beneficios del *credit scoring*:

- **Evaluación de riesgos objetiva:** elimina la influencia del juicio personal en el proceso de decisión.
- **Proceso costo eficiente:** permite la aprobación rápida de los buenos créditos y el rechazo con la misma rapidez de los créditos que no están dentro del perfil de riesgos de la empresa, de tal forma que permite concentrarse en el análisis detallado de aquellos créditos que se sitúan en un área gris.
- **Control estadístico del portafolio:** el riesgo de cada cliente aceptado puede ser descrito en detalle con gran precisión estadística, permitiendo realizar un *backtesting* de las operaciones.
- **Experimentación controlada:** la empresa puede realizar simulaciones, por ejemplo, respecto a aceptar operaciones de alto riesgo, y estimar el impacto que podría tener en el largo plazo en la rentabilidad de su portafolio.

Dicho de otra manera, Sidiqqi (2006) resume que la entidad financiera, a través de *scoring*, tiene la posibilidad de establecer un proceso de decisión consistente y objetivo, basado en la práctica y derivado de la información propia de sus clientes. Combinado con el conocimiento del negocio, un modelo predictivo proveerá a los administradores de riesgos una eficiencia añadida y control en el proceso de administración de riesgos.

No obstante, se debe mencionar que, en ciertos segmentos de clientes, el juicio humano continúa siendo una herramienta indispensable en la evaluación del otorgamiento de un crédito, por lo cual, muchas veces la técnica del juicio humano coexiste y se complementa con herramientas estadísticas.

2.2 Limitaciones del *credit scoring*

De la misma forma, Lawrence y Solomon (2002) señalan que las mayores limitaciones que presenta la técnica del *credit scoring* son:

- **Proceso de desarrollo intensivo en tiempo:** ante lo cual resulta necesario el apoyo de la administración para la priorización de una implementación.
- **Predictibilidad limitada:** solo identifica la probabilidad de que un cliente sea bueno o malo, no identifica si la cuenta individual será buena o mala.
- **Deterioro a lo largo del tiempo:** considerando que está basado en comportamiento pasado, sus beneficios únicamente perdurarán si el sistema es monitoreado y validado contantemente.

2.3 Metodologías en el desarrollo de *credit scoring*

Existen diferentes modelos de *credit scoring* para la identificación de las variables determinantes de la probabilidad de *default*, clasificadas en modelos paramétricos y no paramétricos:

Tabla 1. Principales técnicas para desarrollar un modelo de *creditscoring*

P/ NP (*)	Método	Técnica principal	Resumen
P	Análisis discriminante	Distancia Mahalanobis	Clasifica casos en grupos predefinidos, minimizando las diferencias dentro del grupo
P	Regresión lineal	Mínimos cuadrados ordinarios	Determinación del modelo para estimar la variable de respuesta continua
P	Regresión logística	Máxima verosimilitud	Determinación del modelo para estimar la variable de respuesta binaria.
NP	Árboles de decisión	RPA	Uso de estructuras de árboles para maximizar las diferencias entre grupos
NP	Redes neuronales	*Multilayerperceptron*	AI *technique*, con resultados con dificultad para interpretar y explicar
NP	Programación lineal	Método simplex	Técnica de investigación operativa, usualmente utilizada para la optimización de la asignación de recursos

(*) P: paramétrico, NP: no paramétrico
Fuente: Anderson, 2007.

Métodos paramétricos

- **Análisis discriminante**

Según De Servigny y Renault (2004), el principal objetivo del análisis discriminante es segregar y clasificar una población heterogénea en subconjuntos homogéneos. El

procedimiento consiste en seleccionar un número C de clases en las cuales se segregará la data. En el caso de un modelo de *credit scoring*, estas pueden ser, por ejemplo: *default* y *non default*. Luego, se busca la combinación lineal de variables explicativas que permiten obtener la mayor distancia entre las dos clases.

Asimismo, De Servigny y Renault (2004) señalan que siguiendo el enfoque de Fisher (1936), se separa en dos clases $\omega_1 y \omega_2$, alas que se denominarán buenos y malos prestatarios. La idea de Fisher es buscar la combinación lineal de las variables explicativas que obtenga la máxima distancia entre las dos clases, maximizando la siguiente función:

$$F = \frac{\left|\omega^T(\mu_1 - \mu_2)\right|^2}{\omega^T \sum \omega} \quad (1)$$

Donde:

ω : vector de pesos

μ_i: media de las variables en la clase i

$\sum$: es la matriz de covarianzas entre clases

En (1), el numerador es la covarianza global y el denominador es la varianza. Según Anderson (2007), el modelo asume que la matriz de varianza/ covarianza es la misma para cada grupo.

El máximo se obtiene diferenciando F respecto al vector de pesos e igualando a cero:

$$\frac{\partial F}{\partial \omega} = 0 \qquad (2)$$

La única solución de (2) es:

$$\omega = \sum{}^{*-1} (\mu_1 - \mu_2) \qquad (3)$$

Una observación x es asignada entonces al grupo ω_1 si $\omega^T x + \alpha > 0$ y al grupo ω_2 si $\omega^T x + \alpha < 0$.

Sin embargo, al maximizar F no se obtiene la variable α, que es el punto de corte que separa los grupos, el cual según la práctica indica que tiene que ser determinado por el usuario.

La aplicación más famosa del análisis discriminante en *credit scoring* es el Z-Score de Altman (1986), que incluía entre sus variables explicativas a ratios financieros: capital de

trabajo/ activos, utilidades retenidas/ activos, EBIT/ activos, ventas/ activos, entre los principales.

Según Anderson (2007), el análisis discriminante presenta debilidades relacionadas a la técnica utilizada, en ese sentido, si partimos de que con frecuencia se utiliza una regresión lineal, este presentará limitaciones respecto a los supuestos realizados para su estimación.

- **Modelo de probabilidad de regresión lineal**

Según Anderson (2007), algunas de las más simples relaciones posibles son las lineales, en donde, conforme un valor se incrementa, la variable dependiente cambia a una tasa conocida y constante, lo que corresponde a un modelo de probabilidad lineal (considerando que la variable dependiente es binaria, como en el caso del *credit scoring*).

En el *credit scoring* se obtiene, desde el modelamiento siguiente, un estimado de la probabilidad de un buen crédito (p(bueno)):

$$P(buen\ crédito)_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + e_i$$

La probabilidad por cada observación i, es la suma de una constante y el producto de una serie de pesos β_j y los valores de la variable X_{ij}, donde las variables toman diferentes valores para cada observación y los pesos difieren por cada variable j (no se considera el término de error e_i).

La principal deficiencia del modelo de regresión lineal es el relativo al cumplimiento de la distribución normal de los términos de error y la homocedasticidad, porque el resultado solo tiene dos posibles valores, 0 y 1.

Es así que Anderson (2007), señala que las principales desventajas del modelo de regresión lineal son los supuestos que realiza: (i) linealidad, (ii) homocedasticidad, (iii) distribución normal de los errores, lo que implica que la variable independiente es continua y distribuida normalmente, (iv) independencia del término de error, (v) aditividad, (vi) ausencia de correlación entre las variables independientes, y (vii) uso de variables relevantes.

- **Modelo de regresión logística**

Según Siddiqi (2006), la regresión logística es una técnica común utilizada en el desarrollo de *scorecards* en la industria financiera, en donde la variable dependiente es binaria. Su construcción requiere realizar los siguientes supuestos: (i) variable dependiente categórica, (ii) relación lineal sobre la función log *odds*, (iii) independencia del término de error, (iv) variables independientes no correlacionadas; y (v) uso de variables relevantes.

Asimismo, Anderson (2007) indica que la regresión logística es la técnica con mayor aceptación para elegir para el desarrollo del modelo de *credit scoring*, en particular porque: (i) es específicamente diseñado para una variable dependiente binaria, (ii) el *score* que el modelo estima es fácilmente convertida en la probabilidad de *default* del crédito.

Sean ρ variables independientes y la probabilidad condicional de que el resultado este presente sea $\Pr(Y = 1|x = \pi(x)$, el modelo de regresión logística está dado por:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_{1+\cdots+}\beta_p x_p$$

Donde:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

Asumimos que tenemos una muestra de n observaciones independientes $(x_i, y_i), i = 1, 2, \ldots, n$. El modelo requiere que se obtengan los coeficientes de las variables independientes dado por el vector β'=$\beta_0, \beta_1, \ldots, \beta_p$, utilizando la máxima verosimilitud, por lo que, existirán p+1 ecuaciones de verosimilitud que se obtendrán diferenciando la función log de verosimilitud respecto a los p+1 coeficientes, como se muestra:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

y

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

Para j=1,2,…,p

Para la obtención de los coeficientes se recurre a los diferentes *softwares* estadísticos que se tiene a disposición.

Las unidades de los parámetros obtenidos van a depender de la unidad de la variable de entrada y necesitan ser estandarizados para facilitar el análisis. Según Siddiqi (2006), se puede pasar por alto la unidad de entrada y realizar la regresión contra el *Weight of Evidence* (WOE) de cada agrupación creada; asimismo, esta metodología permite capturar las diferencias entre grupos creados en las variables dependientes, teniendo en cuenta la tendencia y la escala de la relación entre un grupo y otro.
El WOE está definido por la siguiente fórmula:

$$WOE = \ln\left(\frac{Distr\ buenos}{Distr\ malos}\right)$$

Adicionalmente, Siddiqi (2006) indica que un WOE negativo indica que el atributo particular está aislando una proporción más alta de malos que de buenos y se busca que estos sean lo suficientemente diferentes entre las agrupaciones de los atributos, es decir, el agrupamiento debe realizarse con el objetivo de maximizar la diferencia entre buenos y malos créditos, siendo más importante la diferencia entre el WOE de los grupos para establecer la diferenciación. A mayores diferencias entre grupos, mayor es la habilidad predictiva de la variable.

Métodos paramétricos

Según Anderson (2007), los métodos no paramétricos tienen como principal desventaja la falta de transparencia, así como una sobrealimentación.

- **Arboles de decisión**

Anderson (2007) indica que es una herramienta gráfica, con una estructura de cajas y líneas de rama o raíz, utilizada para mostrar posibles giros de eventos que pueden o no ser controlables; cada rama representa las opciones disponibles para un tomador de decisiones. Los árboles de decisión también se utilizan para la visualización de datos en problemas de clasificación y predicción. La forma más primitiva es un tipo de sistema experto, donde

personas con experiencia práctica definen un conjunto de reglas. Todavía se aplica en diagnósticos médicos y donde no hay datos suficientes para hacer un análisis empírico pudiéndose derivar formularios más avanzados basados en el análisis de datos.

La técnica primaria utilizada se denomina RPA, la cual consiste en determinar las ramas que tendrá el árbol de decisión, a través de intentos repetidos para encontrar la mejor división posible. Específicamente, la regla de división es compleja debido a que busca dividir la población en diferentes grupos homogéneos y grupos mutuamente excluyentes. El objetivo es minimizar la distancia entre los miembros de un grupo (similares tasas de *default*), y maximizar la distancia entre grupos (diferentes tasas de *default*).

En general, el RPA no es un modelo adecuado para modelos predictivos, pero existen ciertas instancias donde pueden ser considerados. Por ejemplo, cuando la data disponible para desarrollar un *scorecard* es limitado, como para un nuevo producto.

A pesar de las deficiencias, los RPA siguen siendo herramientas poderosas para su uso en el negocio bancario. Se utilizan mejor para la exploración rápida y sucia de datos, ya sea para obtener información sobre los datos, describir los datos para el negocio, identificar las variables predictivas clave, identificar divisiones *scorecards* o actuar como un punto de referencia para otros modelos.

- **Redes neuronales**

Anderson (2007) las define como redes de elementos computacionales que responden a las variables de entrada, y que aprenden y se adaptan al medio. Tiene el objetivo de imitar el trabajo del cerebro humano, especialmente en lo referente a auto organización y aprendizaje. A diferencia de otras técnicas que siguen procedimientos con fórmulas, las redes neuronales son entrenadas a través de la repetición de ejemplos (Chorafas 1990). El resultado es similar a un árbol de decisión, excepto que es más detallado con reglas de decisión más complejas.

Las ventajas de las redes neuronales son: (i) procesa grandes cantidades de información; (ii) descubre patrones y sigue las relaciones de la data, especialmente las interacciones; (iii) lidia con las relaciones no lineales en la data, y (iv) se entrena basado en las diferencias observadas y los resultados actuales. Sin embargo, también presenta una serie de desventajas o problemas: (i) requiere gran cantidad de iteraciones antes que se obtenga el

modelo final; (ii) es caro de implementar y mantener, especialmente lo referido a entrenamiento, que permita que se adapte a las circunstancias cambiantes; (iii) no es transparente, las relaciones detectadas por el modelo son difíciles de interpretar, y (iv) tiene significante probabilidad de sobrealimentación.

Las redes neuronales no tienen buena adaptación para ambientes donde la decisión lógica debe ser entendida, como es el caso del *scoring* de aplicación de créditos de consumo, donde las compañías podrían explicar las razones de la decisión de no conceder el crédito al potencial cliente, o cuando el negocio exige cierta comprensión del proceso subyacente. Sin embargo, puede ser bien adaptado donde la precisión y predicciones adaptativas son críticas y la transparencia es secundaria.

De acuerdo con Thomas (2000), su primer uso es en áreas donde se cuenta con menor cantidad de datos, como un *scoring* para corporaciones o es *scoring* de fraude. Tan rápido como los prestamistas identifican fraude, y colocan mecanismos de control, el *modus operandi* es cambiado y son encontradas nuevas debilidades. Las redes neuronales tienen la habilidad de adaptarse a esas circunstancias cambiantes, pero requiere monitoreo y reentrenamiento a lo largo de su uso.

- **Programación lineal**

Anderson (2007) indica que el objetivo original de la programación lineal (PL) estaba orientada al apoyo en los problemas de asignación de recursos, como una amplia generalización, la programación lineal es un medio para resolver problemas de asignación de recursos que tienen restricciones. Para el *credit scoring*, funcionaría resolviendo los valores de β en un problema que se presenta en la forma:

Minimizar $\sum e_i^2$ sujeto a:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + e_i$$

$\beta_1 < \beta_3$

$\beta_2 > 0$, etcétera

En otras palabras, intenta elaborar una ecuación de regresión que minimice un cierto término de error, que puede variar, al tiempo que garantiza que las asignaciones de puntos individuales caen dentro de las limitaciones dadas.

La principal ventaja de la PL es que el desarrollador del *scorecard* tiene un mayor control sobre las puntuaciones finales, al ser capaz de incluir sesgos requeridos en las sentencias "sujeto a". Si bien es técnicamente posible utilizar esta técnica para la calificación crediticia, rara vez se usa en la práctica. El rendimiento real de los modelos resultantes puede ser suficiente, pero los prestamistas pueden ser evaluados con mayor ajuste con la aplicación de otros modelos.

Según De Servigny y Renault (2004), varios factores son importantes a la hora de elegir una clase particular de modelo sobre otra:

- **Desempeño:** un modelo más complejo debería proveer una mejora significante en la detección o clasificación de los clientes con más riesgos.
- **Disponibilidad de datos y calidad:** muchos modelos se desempeñan bien en un ambiente de prueba, pero no lidia bien con los problemas prácticos como valores faltantes o valores atípicos (*outliers*). Un modelo más simple, pero más robusto, puede ser preferible a un sistema de última generación en conjuntos de datos desiguales.
- **Entendido por los usuarios:** los usuarios del modelo deben entender perfectamente el funcionamiento del modelo y los principales factores que determinan su resultado. De lo contrario, no podrán detectar sesgos sistemáticos ni comprender los límites.
- **La robustez del modelo ante datos nuevos:** el modelo debe pasar por las diferentes pruebas que aseguren estabilidad, revisando que el modelo esté detectando un óptimo global y no uno local.
- **Tiempo requerido para calibrar o recalibrar el modelo:** es importante, dependiendo de los cambios que se den en la población en la cual se aplicará, se recomienda una revisión anual al respecto.

Tomando en cuenta lo mencionado, el presente trabajo de investigación buscará contrastar su hipótesis desarrollando un modelo de regresión logística.

Capítulo III. Planteamiento del problema

1. Antecedentes generales

El Banco de la Nación es una empresa estatal con potestades públicas, integrante del sector Economía y Finanzas de la República del Perú, bajo operación autónoma en términos económicos, financieros y administrativos, creada mediante Ley 16000 en el año 1966. Se encuentra regulado por su Estatuto aprobado mediante Decreto Supremo 07-94-EF, el Decreto Legislativo que promueve la Eficiencia de la Actividad Empresarial del Estado aprobado mediante Decreto Legislativo 1031 y supletoriamente por Ley 26702, Ley General del Sistema Financiero y del Sistema de Seguros y Orgánica de la Superintendencia de Banca y Seguros.

De acuerdo con su Estatuto, el Banco está facultado para realizar diversas funciones, ninguna de las cuales en forma exclusiva respecto de las empresas y entidades del sistema financiero, dentro de las que indica: «Otorgar una línea de crédito única a los trabajadores y pensionistas del sector público que, por motivo de sus ingresos, posean cuentas de ahorro en el Banco de la Nación. Dicha línea de crédito podrá ser asignada por el beneficiario para su uso mediante préstamos y/o como línea de una tarjeta de crédito. Estas operaciones se harán de acuerdo a un programa anual aprobado por el Ministerio de Economía y Finanzas que podrá ser revisado anualmente».

Dentro de sus estados financieros, al cierre del año 2016, el activo del Banco alcanzó S/ 28.500 millones, disminuyendo en S/ 1.051 millones (-4 %) con relación al 2015. El pasivo también disminuyó alcanzando S/ 26.127 millones, cifra que resulta S/ 1.629 millones menor a la registrada el año anterior, dicha reducción fue mitigada por el incremento, debido a la emisión de bonos subordinados por S/ 250 millones, en el marco del programa de fortalecimiento patrimonial aprobado en el año 2016.

En el mismo periodo, el patrimonio alcanzó S/ 2.373 millones, aumentando en S/ 578 millones con relación al 2015. Esto se explica por el incremento del capital social en S/ 200 millones (capitalización de parte de las utilidades del año fiscal) y por el resultado neto del ejercicio (S/ 191 millones). Asimismo, muestra indicadores financieros de capital, calidad de activos, manejo administrativo, rentabilidad y liquidez, saludables.

Tabla 2. Principales indicadores del BN

Resumen	Unidad	2012	2013	2014	2015	2016
Información del estado de situación financiera y estado de ganancias y pérdidas						
Créditos de consumo total no revolvente	Millones de S/	2 852	3 194	3 354	3 410	3 460
Total activo	Millones de S/	24 179	27 020	28 284	29 550	28 500
Créditos consumo/ total activo	%	11.8	11.8	11.9	11.5	12.1
Total pasivo	Millones de S/	22 116	25 229	26 323	27 755	26 127
Patrimonio	Millones de S/	2 063	1 792	1 961	1 795	2 373
Resultado neto	Millones de S/	669	589	705	685	876
Principales indicadores financieros						
1. Capital						
Ratio de capital global	%	19.2	15.8	13.9	13.3	19.4
Pasivo total / cap. social y reservas	N° veces	16.4	18.7	19.5	20.6	16.9
2. Calidad de activos (*asset*)						
Cartera atrasada / créditos directos	%	0.6	0.5	0.5	0.6	0.7
Provisiones / créditos directos	%	3.1	2.8	2.8	2.5	2.7
3. Manejo administrativo (*management*)						
Gastos administrativos / ingresos totales	%	42.05	45.7	43.3	42.6	38.1
Gastos adm. (sin jub.) / ingresos totales	%	36.4	40.4	37.2	36.6	32.8
Créditos directos / personal	Miles de S/	1,359	1,601	1,753	2,330	2,181
Depósitos / oficinas	Miles de S/	34,365	37,924	37,945	40,213	36,667
4. Rentabilidad (*earnings*)						
ROE (utilidad anualizada / patr. prom.)	%	35.6	33.8	40.0	37.3	41.2
ROA (utilidad anualizada / activo prom.)	%	2.9	2.4	2.6	2.5	3.2
Ingresos financieros / ingresos totales	%	62.8	65.7	64.6	63.7	64.8
Gastos financieros / ingresos financieros	%	5.9	7.0	3.8	10.7	10.2
5. Liquidez (*liquidity*)						
Liquidez MN	%	85.4	80.5	86.5	72.5	72.7
Liquidez ME	%	119.9	170	215.7	144.1	152.3

Fuente: Banco de la Nación, 2017.

Respecto a los préstamos de consumo no revolvente[3], en el año 2001lanzó al mercado el préstamo de consumo, denominado Préstamos Multired, el cual está dirigido exclusivamente a trabajadores del sector público que, por motivos del depósito de sus remuneraciones o pensiones, mantengan una cuenta de ahorros en el Banco, teniendo el carácter de no revolvente y que estos depósitos permitan el repago de las facilidades otorgadas.

Principales características del producto:

- El préstamo Multired ofrece las modalidades de: (i) Convenio, dirigido a un sector específico con el que el Banco suscribe un convenio;(ii) Clásico, el cual requiere un garante, y (iii) Comercial, que está dirigido a financiar la adquisición de productos o servicios con

[3] Según Resolución S.B.S. 11356 – 2008, los créditos no revolventes son aquellos créditos otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

empresas proveedoras autorizadas que firmen un acuerdo comercial a favor de clientes del BN.

- El Banco realiza el cobro automático con cargo a la cuenta de ahorros del cliente, esta característica lo diferencia del préstamo descuento por planilla, el cual realiza el cobro directamente de la planilla del trabajador.

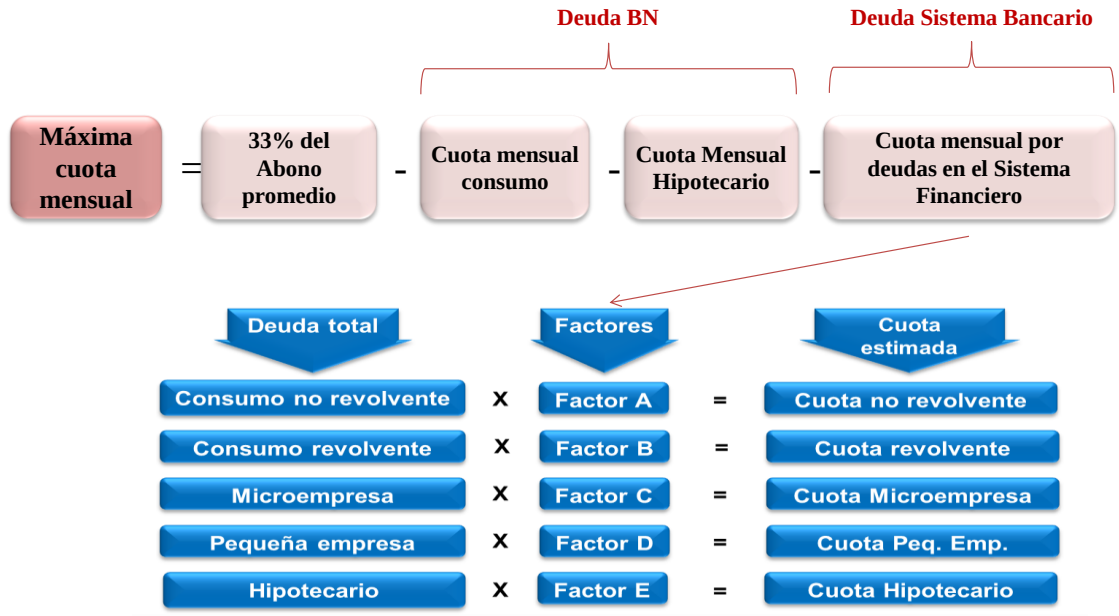
Requisitos para la admisión:

- Ser trabajador con contrato a plazo indeterminado, determinado o con Contrato Administrativo de Servicios (CAS); para los dos últimos requiere la presentación de garante, independientemente de la modalidad del crédito.
- Mantener deuda directa con cuatro entidades del sistema financiero como máximo, incluyendo al BN.
- El cliente debe estar con calificación Normal en la Central de Riesgo de la SBS.

En caso el cliente cumpla con lo mencionado, el modelo de negocios cuenta con un cálculo matemático de la capacidad de endeudamiento del cliente a través de la máxima cuota a la que el cliente puede acceder, la cual, en términos generales, consiste en lo siguiente:

- La capacidad de pago está en función del promedio de ingreso mensual del cliente y de su nivel de endeudamiento, considerando las deudas con otras entidades del sistema financiero. Asimismo, incorpora medidas para evitar el riesgo de sobreendeudamiento.
- Considera el 33% de afectación de los ingresos mensuales, deduciendo las deudas del cliente en el sistema financiero, así como los créditos vigentes que el cliente mantenga en el Banco de la Nación.
- Aplicando la siguiente fórmula se obtiene la máxima cuota mensual a la que el cliente puede acceder y cual es llevada a importe a desembolsar a través del valor actual de la cuota mensual obtenida de acuerdo al plazo que opte el cliente.

Gráfico 2. Cálculo de la capacidad de endeudamiento para un cliente BN



Fuente: Banco de la Nación, 2017.

- La capacidad de endeudamiento está limitada, además, por el monto máximo a la que puede acceder un cliente según la edad del mismo (asociado a su vez, a las coberturas del seguro de desgravamen).

Tabla 3. Parámetros aplicables en el otorgamiento de un crédito BN

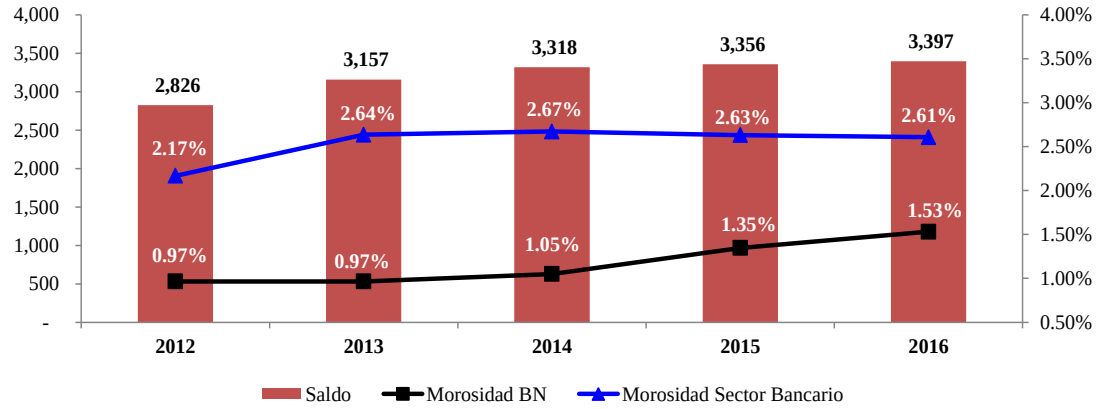
Edad del cliente	Plazo máximo	Monto máximo en soles	Tasas de interés %
Entre 83 años y 84 años	a 12 meses	2.000,00	13
Entre 78 años y 82 años	a 24 meses	2.000,00	13 – 14
Hasta 77 años	Hasta 24 meses	19.000,00	13 – 14
Hasta 76 años	Hasta 36 meses	19.000,00	13 – 15
Hasta 75 años	Hasta 48 meses	19.000,00	13 – 16
Entre 60 y 74 años	Hasta 48 meses y 60 meses	19.000,00	13 - 16, 19
Entre 18 años y 59 años	Hasta 48 meses y 60 meses	50.000,00	13 - 16, 19

Fuente: Banco de la Nación, 2017.

2. Morosidad

Al cierre del año 2016, el saldo de Préstamos Multired ascendió a S/ 3.397 millones (crecimiento anual de 1,21%) presentando una morosidad de 1,53%. Con referencia al sistema bancario, el Banco se posiciona en 5.° lugar por saldo de préstamos no revolventes y revolventes[4], y mantiene una morosidad por debajo del sistema (2,61% vs. 1,53%).

Gráfico 3. Evolución del saldo vigente de préstamos de consumo no revolvente y morosidad



Fuente: SBS -Estados Financieros del Banco de la Nación y sistema bancario, 2017.

La principal causa de la morosidad en los préstamos Multired es la pérdida del abono en el Banco de la Nación, ya sea por pasar a condición de cesante, por despido, o por migración de su cuenta de ahorros a otra entidad del sistema financiero, tomando en cuenta ello, se presenta como un riesgo que el cliente del Banco de la Nación opte por trasladar su cuenta de ahorros a otra entidad del sistema financiero[5],así como por las políticas más agresivas que tiene el sistema para ampliar su base de clientes, teniendo en consideración que ya cuentan con historial crediticio favorable construido en el Banco de la Nación y podría constituirse en un sujeto de crédito atractivo en el sistema, más aun, teniendo en cuenta que, por ejemplo, para los desembolsos del mes de noviembre 2016, el 53% mantenía deuda con el sistema financiero.

En ese sentido, en un escenario adverso, el Banco tendrá que adecuar su modelo de negocios en el que deberá evaluar el requisito del cargo automático de las cuotas de los préstamos en la

[4] Ver anexo 1.
[5] Según Decreto Supremo 003-2010-TR, sobre modificación de las normas reglamentarias relativas a la obligación de los empleadores de llevar planillas de pago, modificó el artículo 18 del Decreto Supremo 001-98-TR, indicando que: «Si el pago por terceros se efectúa a través de las empresas del sistema financiero, los trabajadores tendrán derecho de elegir aquella donde se efectuarán los depósitos, (…)».

cuenta de ahorros, posiblemente porque el trabajador podría optar por migrar su cuenta, en tanto no suceda lo anterior, resulta importante identificar las variables determinantes de la probabilidad de *default* en la cartera de consumo no revolvente del BN, a través de un modelo de regresión logística de *credit scoring* debido a que con una adecuada gestión de riesgos, además le brindaría ventajas respecto a establecer ratios de morosidad objetivo en función a su perfil de riesgos y su nivel de rentabilidad esperado. En ese mismo sentido, contaría con un primer avance para postular al uso de un modelo interno para la determinación del requerimiento de capital por riesgo crediticio.

Adicionalmente, otra ventaja es la inclusión de la herramienta en su nuevo core bancario dentro de su módulo de colocaciones.

Capítulo IV. Metodología e implementación

1. Base de datos y variables utilizadas en la estimación

1.1 Construcción de la base de datos y selección de la muestra

La información corresponde a los desembolsos de los préstamos Multired a nivel nacional en el periodo enero – diciembre 2014[6] de los clientes del Banco de la Nación, la cual fue obtenida de las fuentes que se detallan a continuación:

- Sistema de desembolsos del Banco, el cual contiene el registro diario de los desembolsos a nivel nacional, incluye información de: importe desembolsado, plazo, tasa de interés, unidad ejecutora a la que pertenece el cliente, agencia en donde se desembolsó el crédito, y número de desembolsos anteriores.
- Sistema de ahorros del Banco, del que se obtuvo la información de: saldo promedio de abono y la antigüedad de la cuenta de ahorros, la cual se utilizó como variable *proxy* de la antigüedad laboral.
- Reporte crediticio de deudores, se obtuvo la información de: edad, sexo y días de atraso del cliente.
- Reporte crediticio consolidado, se obtuvo la información del saldo adeudado y número de entidades con las que el cliente tenía exposición crediticia en el momento del desembolso del crédito.

Durante el año 2014, el total de desembolsos fue de 277 mil créditos, de los cuales los créditos definidos como insuficiente e indeterminado fueron excluidos, quedando de esta manera un total 169.267 préstamos. Contando con esta información, se obtuvo dos muestras: i. muestra de desarrollo, sobre la cual se estiman los tres modelos, cuenta con 84.635 registros, y ii. muestra de validación, cuenta con 84.632 registros. Como se observa, ambas muestras tienen similar cantidad de datos, lo que permite contar con suficiente información para que el modelo aprenda, así como que permita una verificación consistente del modelo obtenido.

[6] Según Carta EF/92.6200 004-2017, el Banco de la Nación indicó que no es posible remitir información reciente de su operativa mensual (año 2016) con el fin de salvaguardar sus intereses.

Las principales limitantes de la muestra son:

- La información de desembolsos no logra cubrir un ciclo económico.
- El producto es homogéneo, sin embargo, establece algunos requisitos especiales, dependiendo de la edad y el sector en donde labora.
- La base de datos no cuenta con los créditos que fueron solicitados y que no se desembolsaron (rechazados), por lo que no se tiene las características de estos clientes. La base de rechazados se puede analizar a través de aplicación de técnicas de inferencia, aplicando el modelo estimado a los créditos rechazados (a través de técnicas como *fuzzy augmentation*, punto de corte y *parceling*).

Tabla 4. Composición de las muestras de construcción y validación

Muestra	Buenos	Malos	Total
Total	164.856	4.412	169.268
Muestra de construcción (50%)	82.478	2.158	84.636
Muestra de validación (50%)	82.378	2.254	84.632

Fuente: Elaboración propia, 2017.

1.2 Definición de variables

1.2.1 Definición de *default*

La SBS, en su Resolución 14345-2009 artículo 56.°, define como una operación para clientes minoristas en estado de incumplimiento, aquella que presenta, al menos, una de las siguientes características:

- Un atraso mayor a los 90 días.
- Pasa a una situación de reestructurado.
- Que en los últimos años la operación haya registrado más de una refinanciación, salvo que la empresa demuestre estadísticamente, a satisfacción de la SBS, que posee otras variables o mecanismos que predigan mejor la entrada en estado de incumplimiento.
- Que la empresa considera que el deudor es incapaz de honrar sus obligaciones en la forma pactada, sea parcial o totalmente.

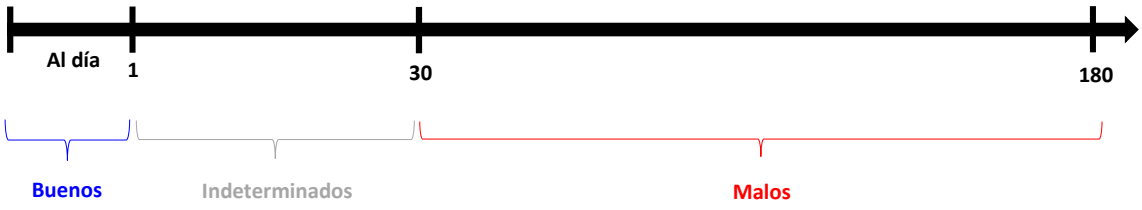
Asimismo, según Resolución SBS 11356-2008, un crédito minorista no revolvente se contabiliza como vencido de manera escalonada, es decir, después de 30 días de la fecha pactada y no haber pagado la cuota correspondiente, se considera como vencida la porción no

pagada, mientras que para considerar la totalidad del saldo insoluto como vencido, se realiza a partir de los 90 días desde la fecha pactada correspondiente.

Tomando en consideración lo anterior y el perfil crediticio del Banco de la Nación, se plantea las siguientes definiciones:

- Buen crédito: crédito que durante los siguientes 12 meses a su desembolso tuvo cero (0) días de atraso.
- Mal crédito: crédito que en alguno de los siguientes 12 meses a su desembolso acumuló más de treinta (30) días de atraso.
- Indeterminado: aquel crédito que en alguno de los siguientes 12 meses a su desembolso tuvo entre 1 y 30 días de atraso.
- Insuficiente: aquel crédito del que no se cuenta con información completa en los 12 meses siguientes a su desembolso.

Gráfico 4. Definición del cliente



Fuente: Elaboración propia, 2017.

Así, la variable dependiente binaria, toma el valor de 1 si el crédito es malo, de lo contrario toma el valor de cero.

Adicionalmente, se evidencia que la mayoría de los créditos que pasan los 30 días de atraso, continúan incrementando sus días de atraso evaluado 12 meses después, según se aprecia en la tabla 5 y evidenciaría que la definición de malos considerada para el modelo es consistente.

Tabla 5. ***Roll rate analysis***

Días de atraso (dic 14)	Días de atraso (dic 15)					Total general	Deterioro (%)
	0-8	0-30	31-60	61-120	121-más		
0-8	392.534	4.459	1.122	1.766	4.269	404.150	2,9
0-30	1.979	1.171	81	78	225	3.534	44,0
31-60	332	68	53	39	250	742	55,3
61-120	341	36	24	51	487	939	63,7
121-más	271	47	21	40	6.549	6.928	96,1
Total general	**395.457**	**5.781**	**1.301**	**1.974**	**11.780**	**416.293**	**5,0**

Fuente: Elaboración propia, 2017.

1.2.2 Variables independientes

Para la selección de las variables determinantes de la probabilidad de *default* del cliente de consumo no revolvente del BN se eligieron en total doce variables cualitativas y cuantitativas, las cuales poseen información propia del crédito solicitado así como características laborales y demográficas del potencial cliente.

Específicamente se tienen cuatro variables cualitativas y ocho variables cuantitativas.

Tabla 6. Descripción de variables

N°	Variable	Descripción	Tipo
1	Y	Variable dependiente: 1: mal crédito y 0: buen crédito	Binaria
2	DEPARTAMENTO	Departamento en donde se encuentra ubicada la agencia en la que se desembolsó el préstamo.	Cualitativa
3	IMPORTEPRÉSTAMO	Importe desembolsado. Según definición del Banco, puede ir de S/ 300 a S/ 50.000.	Cuantitativa
4	PLAZO	Plazo del préstamo. Según definición del Banco puede ir desde 01 mes hasta 60 meses.	Cuantitativa
5	TIPO_PRÉSTAMO	Sector al cual pertenece el cliente al que se le otorgó el préstamo.	Cualitativa
6	SITUACIONCLIENTE	Puede ser activo o pensionista, según su situación en su centro de labores.	Cualitativa
7	EDAD	Edad del cliente.	Cuantitativa
8	SEXO	Sexo del cliente.	Cualitativa
9	INGRESO	Importe mensual que el cliente recibe en su cuenta de ahorros abierta en el Banco de la Nación.	Cuantitativa
10	ANTIG_LAB	Años de servicio en su centro de labores, se aproxima a la antigüedad de la cuenta de ahorros.	Cuantitativa
11	NUM_ENT	Número de entidades con las cuales el cliente mantiene deuda en el sistema financiero.	Cuantitativa
12	DEUDA_SF	Importe adeudado en entidades con las cuales el cliente mantiene deuda en el sistema financiero.	Cuantitativa
13	NUM_PRÉSTAMOS	Número de préstamos de consumo no revolvente solicitados en el Banco de la Nación con anterioridad.	Cuantitativa

Fuente: Elaboración propia, 2017.

Se realizará la exploración inicial (estadística descriptiva) de las variables independientes, a fin de conocer sus principales características como se observa en la tabla 7:

Tabla 7. Estadística descriptiva de las variables cuantitativas

Estadístico	Variables independientes cuantitativas							
	IMPORTE_PRESTAMO	PLAZO	EDAD	INGRESO	ANTIG_LAB	NUM_ENT	DEUDA_SF	NUM_PRESTAMOS
Muestra de construcción								
Media	7,411	40	51	1,366	7	2	7,243	6
Mediana	5,164	48	51	1,218	8	2	111	5
Moda	2,000	48	46	398	12	1	-	1
Desviación estándar	6,497	13	16	887	5	1	12,506	4
Curtosis	4.1	-0.1	-0.8	9.5	-1.6	0.4	7.1	0.2
Coeficiente de asimetría	1.7	-1.2	-0.0	2.1	-0.3	1.0	2.4	0.7
Rango	49,700	59	66	9,878	13	6	100,000	31
Mínimo	300	1	18	102	-	1	-	1
Máximo	50,000	60	84	9,980	13	7	100,000	32
Cuenta	84,636	84,636	84,636	84,636	84,636	84,636	84,636	84,636
Nivel de confianza(95.0%)	43.8	0.1	0.1	6.0	0.0	0.0	84.3	0.0
Muestra de validación								
Media	7,429	40	51	1,368	7	2	7,241	6
Mediana	5,200	48	51	1,215	8	2	111	5
Moda	2,000	48	49	398	12	1	-	1
Desviación estándar	6,518	13	16	893	5	1	12,496	4
Curtosis	4.2	-0.1	-0.8	9.9	-1.5	0.5	7.1	0.2
Coeficiente de asimetría	1.7	-1.2	-0.0	2.2	-0.3	1.0	2.4	0.7
Rango	49,700	59	66	9,893	13	6	100,000	30
Mínimo	300	1	18	101	-	1	-	1
Máximo	50,000	60	84	9,994	13	7	100,000	31
Cuenta	84,632	84,632	84,632	84,632	84,632	84,632	84,632	84,632
Nivel de confianza(95.0%)	43.9	0.1	0.1	6.0	0.0	0.0	84.2	0.0

Fuente: Elaboración propia, 2017.

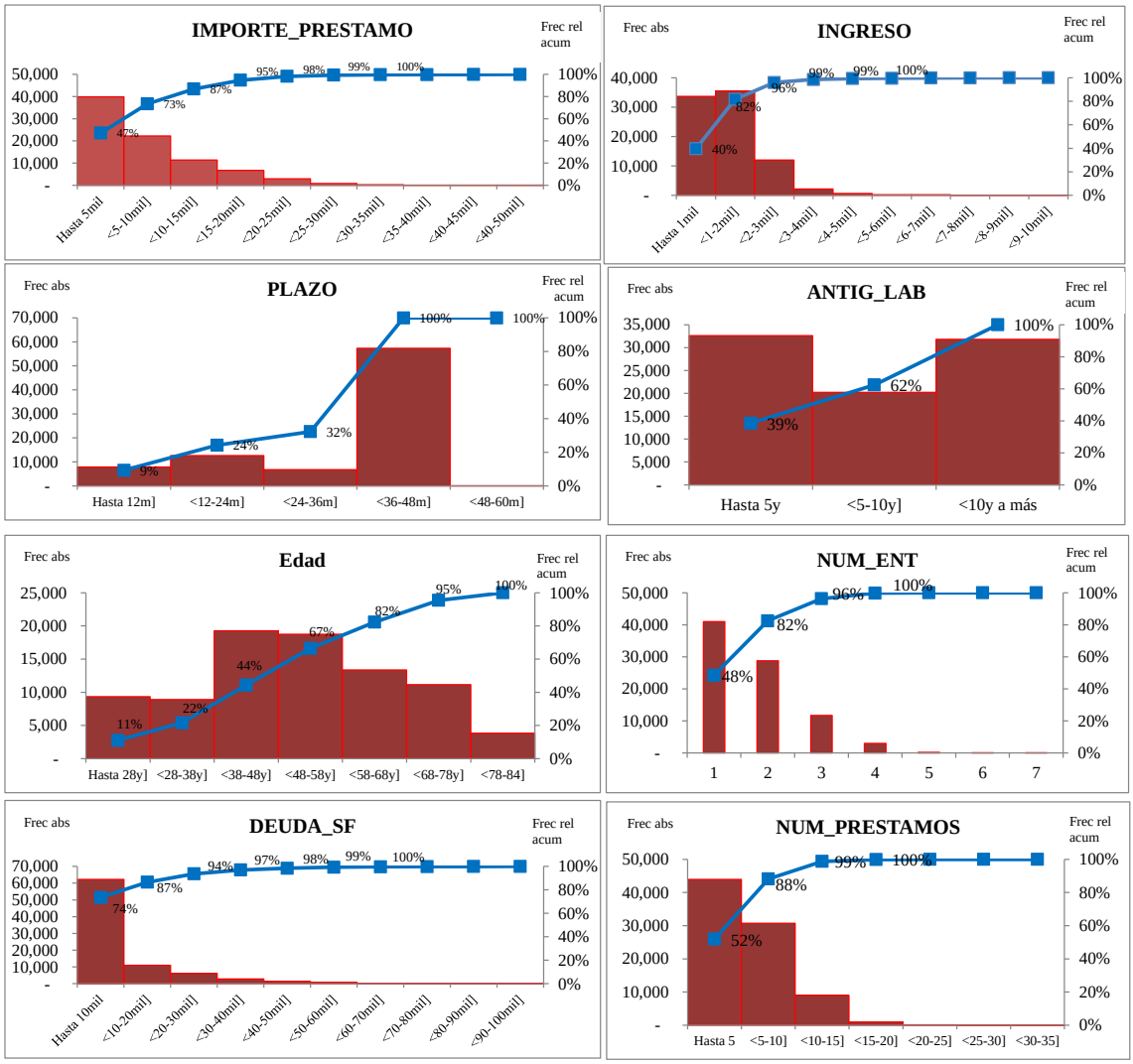
De la tabla 7 y gráfico 5 que resumen las características tanto de la muestra de construcción como de validación (compuesta por los 169.267 préstamos desembolsados entre enero y diciembre 2014), se procederá a mencionar las principales características de la muestra de construcción:

- Se tiene un promedio de importe de préstamo de S/ 7.411. El 73 % de los créditos han sido desembolsados hasta por S/ 10.000.
- El plazo promedio solicitado es de 40 meses. Solo el 32 % de los datos han sido desembolsados a plazos menores o iguales de 36 meses.
- La edad promedio del cliente al que se le desembolsó un préstamo Multired fue de 51 años. El 33 % de los créditos fue otorgado a clientes con más de 58 años de edad.
- El ingreso promedio mensual de los clientes al momento del desembolso fue S/ 1.366. El 82 % del total de desembolsos tuvo un ingreso menor o igual a S/ 2.000.

- La antigüedad laboral promedio fue 7 años, cabe indicar que la variable fue aproximada a la antigüedad de la cuenta de ahorros en el Banco de la Nación. El 62 % de los datos está situado en el periodo de hasta 10 años.
- El número de entidades del sistema financiero con la que el deudor del BN tiene deuda directa es de dos préstamos, siendo que el 82 % de los datos están situados hasta este número de entidades.
- La deuda con entidades del sistema financiero con la que el deudor del BN tiene posición es de S/ 7.243 en promedio mensual, siendo que el 74 % de los datos están situados hasta S/ 10.000.
- El número de préstamos Multired anteriores al desembolso actual que el cliente mantuvo con el BN fue seis en promedio, con una mediana de cinco préstamos. El 52 % de los datos han tenido como máximo hasta cinco préstamos con anterioridad.

Adicionalmente, se observa que tanto para la muestra de construcción y validación, los estadísticos son similares.

Gráfico 5. Histogramas y diagrama de frecuencias de las variables independientes– variables cuantitativas

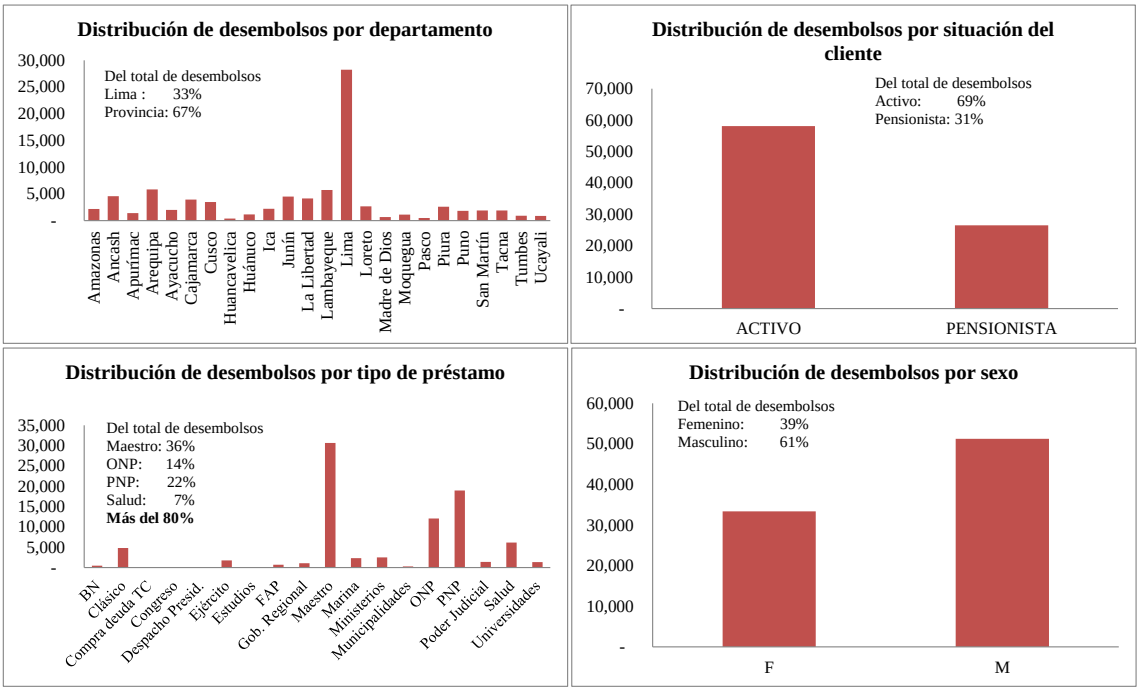


Fuente: Elaboración propia, 2017.

Del gráfico 6 se observan las siguientes características:

- En la variable Departamento, de los créditos desembolsados en provincia Arequipa, Lambayeque y Ancash cuentan con mayor participación, siendo 10 % para los dos primeros y 8 % para el último.
- Para la variable tipo de préstamo, el 81 % de desembolsos en el periodo 2014, se realizaron a los sectores Maestro, ONP, PNP y Salud.
- Las variables situación del cliente tiene una proporción de69 % y 31 % para activo y pensionista, respectivamente. Mientras que la variable sexo, muestra que el 39 % son créditos a trabajadoras del sector público y 61 % para trabajadores del sector público.

Gráfico 6. Diagrama de barras de las variables independientes cualitativas



Fuente: Elaboración propia, 2017.

Se precisa que en este estudio las variables cualitativas se trabajarán con el WOE respectivo por atributos comunes encontrados dentro de las variables a través de los árboles de decisión, por lo que no será necesario construir categorías de referencia, según el planteamiento de Siddiqi (2006). Este proceso de trasformación de variables originales en variables WOEs comprende un proceso de categorización de la variable en formato numérico.

1.2.3 Selección de variables

Dentro del análisis realizado a las variables es importante determinar las variables que mejor explican la probabilidad de *default* del cliente de consumo no revolvente del BN, para lo cual se utilizó árboles de decisión con el fin de analizar cada variable independiente y visualizar su utilidad para la discriminación en la variable dependiente (entre créditos malos y buenos); asimismo, para identificar las categorías de agrupación de las variables independientes dentro de sí mismas.

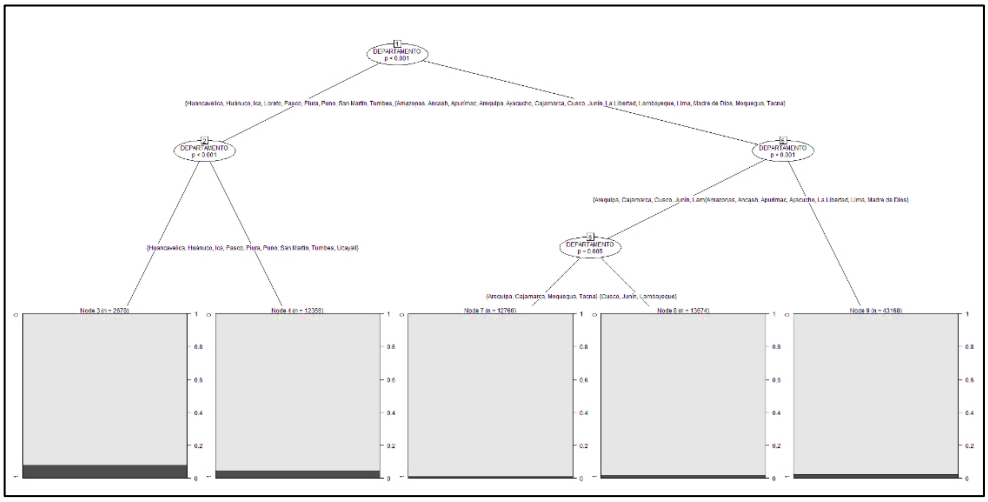
Según Siddiqi (2006) el agrupamiento de atributos dentro de las variables resulta ventajoso por lo siguiente:

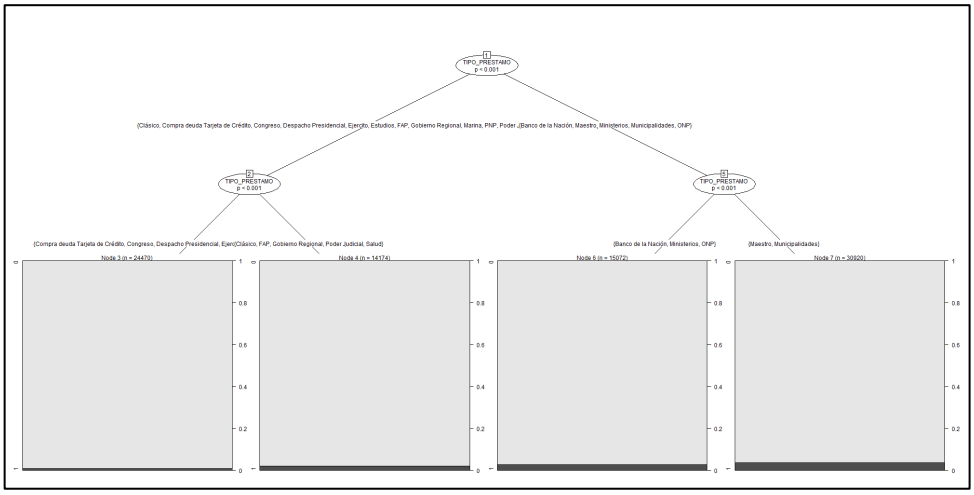
- Ofrece una alternativa fácil para tratar *outliers*, clases raras, etc.

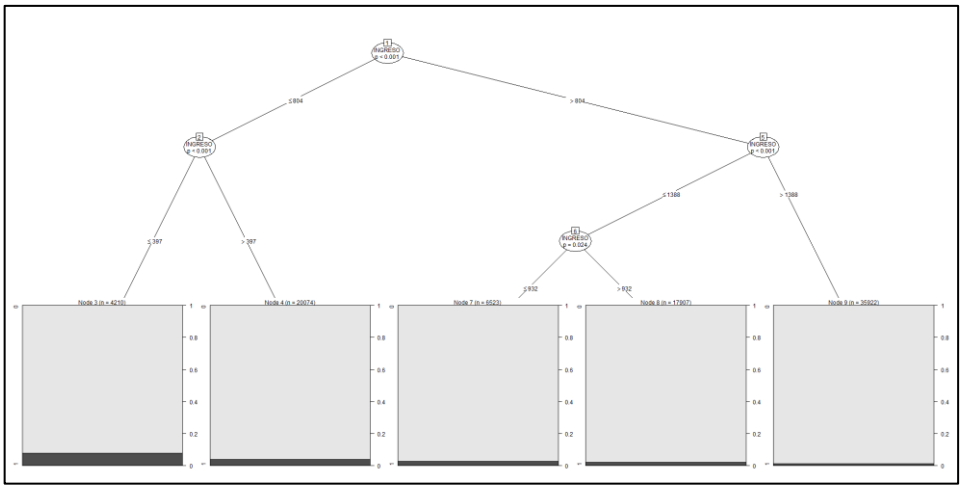
- Permite entender las relaciones y conocer el portafolio crediticio, lo que redunda sobre la gestión del portafolio.
- Las dependencias no lineales pueden ser modeladas con construcciones lineales.
- Otorga control sobre el proceso de desarrollo, la formación de los grupos finalmente impactará en el modelo.

Se muestra los árboles de tres variables. Ver anexo 2 para las 12 variables utilizadas en el estudio.

Gráfico 7. Árboles de decisión para variable. Departamento, tipo de préstamo e ingreso







Fuente: Elaboración propia, 2017

Igualmente, basado en el resultado de la agrupación en categorías de las variables usando los árboles de decisión se calculó el indicador WOE para cada variable, que permiten obtener diferencias entre los WOE de los atributos de las variables sometidas al análisis.

Seguidamente, en la selección de variables se utilizó el indicador valor de la información (IV), el cual es una medida relativa que indica que tan discriminante es cada variable predictiva, su cálculo está basado en el WOE. A mayor IV, la variable tiene más poder discriminante. Según Siddiqi (2006), la regla para considerar una variable basado en el IV es:

Tabla 8. Regla de valor de la información

IV	Conclusión
Menor a 0,02	no predictivo
Entre 0,02 y 0,10	débil
Entre 0,10 y 0,30	medio
Mayor a 0,30	fuerte

Fuente: Siddiqi, 2006.

Se utilizó el *software* R[7] para la obtención de los árboles de decisión y a partir de los cuales se construyeron el WOE y el IV, con la finalidad de determinar si la mencionada variable se incluye dentro del modelo. Se precisa que, para lo indicado, se utilizó exclusivamente la muestra de desarrollo.

Tabla 9. Resumen de IV para las variables independientes

Variable	IV	Calificación
DEPARTAMENTO	0,26	medio
IMPORTE_PRÉSTAMO	0,06	débil
PLAZO	0,10	medio
TIPO_PRÉSTAMO	0,26	medio
SITUACIÓN_CLIENTE	0,26	medio
EDAD	0,06	débil
SEXO	0,00	no predictivo
INGRESO	0,34	fuerte
ANTIG_LAB	0,06	débil
NUM_ENT	0,00	no predictivo
DEUDA_SF	0,05	débil
NUM_PRÉSTAMOS	0,05	débil

Fuente: Elaboración propia, 2017

[7] Sistema para análisis estadístico y gráficos creado por Ross Ihaka y Robert Gentleman, Ihaka R. & Gentleman R. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299–314 (https://cran.r-project.org/doc/contrib/rdebuts_es.pdf)

En detalle, con la muestra de construcción se tiene:

Tabla 10. Detalle de WOE e IV para las variables independientes

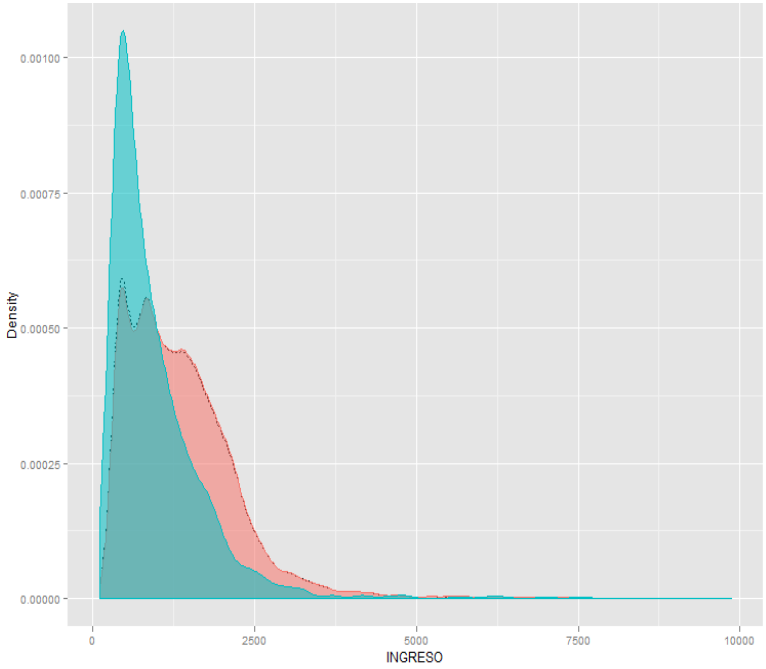
N°	Atributo	Buenos	Malos	Total	PD	WOE	IV Atributo	IV
VARIABLE: DEPARTAMENTO								
1	Loreto	2,466	212	2,678	7.9%	-1.19	0.08	0.08
2	Grupo5	6,586	358	6,944	5.2%	-0.73	0.06	0.14
3	Grupo1	5,202	210	5,412	3.9%	-0.43	0.01	0.16
4	Grupo2	42,159	1,009	43,168	2.3%	0.09	0.00	0.16
5	Grupo3	13,451	223	13,674	1.6%	0.46	0.03	0.19
6	Grupo4	12,614	146	12,760	1.1%	0.82	0.07	0.26
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: IMPORTE_PRESTAMO								
1	<5001	38,657	1,254	39,911	3.1%	-0.21	0.02	0.02
2	5001-10000	21,727	507	22,234	2.3%	0.11	0.00	0.03
3	>10001	22,094	397	22,491	1.8%	0.38	0.03	0.06
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: PLAZO								
1	Hasta 11	2,767	9	2,776	0%	2.08	0.06	0.06
2	De 12 a 40	24,358	472	24,830	2%	0.30	0.02	0.08
3	Mayor a 40	55,353	1,677	57,030	3%	-0.15	0.02	0.10
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: TIPO_PRESTAMO								
1	Grupo1	24,231	239	24,470	1.0%	0.98	0.18	0.18
2	Grupo2	13,880	294	14,174	2.1%	0.21	0.01	0.19
3	Grupo3	14,629	443	15,072	2.9%	-0.15	0.00	0.19
4	Grupo4	29,738	1,182	30,920	3.8%	-0.42	0.08	0.27
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: SITUACION_CLIENTE								
1	ACTIVO	56,489	1,612	58,101	3%	-0.09	0.01	0.01
2	PENSIONISTA	25,989	546	26,535	2.1%	0.22	0.01	0.02
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: EDAD								
1	De 18 a 21	2,470	18	2,488	0.7%	1.28	0.03	0.03
2	De 22 a 24	3,178	28	3,206	0.9%	1.09	0.03	0.06
3	De 25 a 28	3,585	71	3,656	1.9%	0.28	0.00	0.06
4	Mayor a 28	73,245	2,041	75,286	2.7%	-0.06	0.00	0.06
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			

N°	Atributo	Buenos	Malos	Total	PD	WOE	IV Atributo	IV
VARIABLE: SEXO								
1	F	32,548	823	33,371	2.5%	0.03	0.00	0.00
2	M	49,930	1,335	51,265	2.6%	-0.02	0.00	0.00
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: INGRESO								
1	De 100 a 400	6,151	360	6,511	5.5%	-0.81	0.07	0.07
2	De 401 a 800	16,807	793	17,600	4.5%	-0.59	0.10	0.17
3	De 800 a 930	6,421	183	6,604	2.8%	-0.09	0.00	0.17
4	De 931 a 1400	18,155	386	18,541	2.1%	0.21	0.01	0.18
5	Mayor a 1400	34,944	436	35,380	1.2%	0.74	0.16	0.34
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: ANTIG_LAB								
1	0	18,344	290	18,634	1.6%	0.50	0.04	0.04
2	1	3,947	89	4,036	2.2%	0.15	0.00	0.05
3	Mas de 1	60,187	1,779	61,966	2.9%	-0.12	0.01	0.06
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: NUM_ENT								
1	1	39,998	988	40,986	2%	0.06	0.00	0.00
2	Mas de 1	42,480	1,170	43,650	2.7%	-0.05	0.00	0.00
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: DEUDA_SF								
1	<6001	56,364	1,288	57,652	2.2%	0.14	0.01	0.01
2	6001-44000	24,229	754	24,983	3.0%	-0.17	0.01	0.02
3	>44001	1,885	116	2,001	5.8%	-0.86	0.03	0.05
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			
VARIABLE: NUM_PRESTAMOS								
1	1	13,119	201	13,320	1.5%	0.54	0.04	0.04
2	2	7,395	158	7,553	2.1%	0.20	0.00	0.04
3	Mas de 2	61,964	1,799	63,763	2.8%	-0.10	0.01	0.05
	Total	**82,478**	**2,158**	**84,636**	**2.5%**			

Fuente: Elaboración propia, 2017.

Considerando el IV de las variables independientes, para los casos en que el IV sea de 0,00, debería removerse y no ser considerados en la formulación del modelo de regresión, es decir, SEXO y NUM_ENT; sin embargo, el modelo considerará la variable SEXO debido a que en la práctica le agregan un ajuste marginal adicional y se removerán si es que no aportan al ajuste del modelo y resulta no significativa. De la revisión de los IVs de las variables, se observa que el INGRESO es la variable más predictiva. A continuación, se presenta el gráfico 8 que evidencia su poder discriminante.

Gráfico 8. Distribución del ingreso según buenos (rojo) y malos (verde)

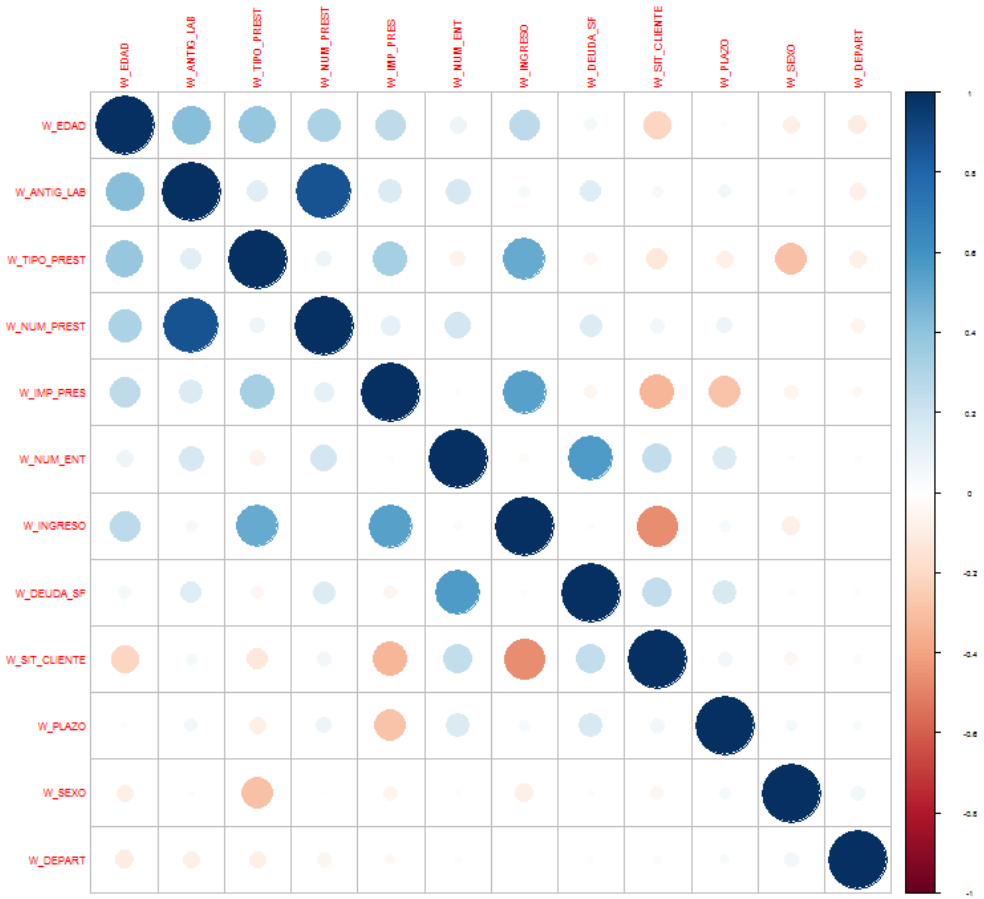


Fuente: Elaboración propia, 2017.

Es preciso indicar que las variables independientes tomadas en cuenta para regresionar el modelo, fueron convertidas previamente y asignadas con el WOE que les correspondan de acuerdo con la tabla 10, las cuales se transforman en formato Excel y son ingresadas al programa R que se utilizará para obtener el modelo de regresión; se menciona que se realiza esta transformación debido a que tiene ventajas al no considerar las diferentes unidades en las que se encuentran las mismas según lo señalado anteriormente.

Antes de obtener los modelos se analizó la correlación de las variables a través de la matriz de correlaciones y el clúster de correlaciones.

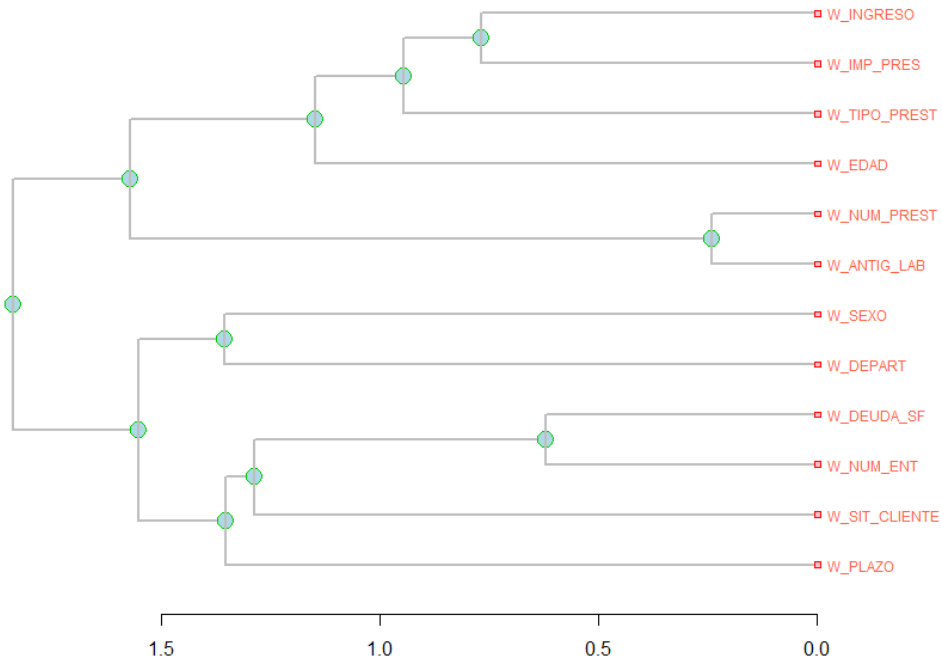
Gráfico 9. Matriz de correlaciones



Fuente: Elaboración propia, 2017.

Se observa fuerte correlación positiva entre las variables W_NUM_PREST y W_ANTIG_LAB, y en menor medida entre el W_INGRESO y (W_TIPO_PREST, W_IMP_PREST), y entre las variables W_DEUDA_SF y W_NUM_ENT. Este tipo de asociaciones fuertes se confirman en el clúster de correlaciones (jerárquico) que es representado a través de un dendograma.

Gráfico 10. Clústeres de correlación de variables



Fuente: Elaboración propia, 2017.

2. Estimación del modelo de regresión e interpretación

Se estimó el modelo, el cual muestra la probabilidad de incumplimiento, quedando definido por las siguientes variables:

Tabla 11. Modelo de regresión logístico

Variable dependiente: Y[0: Bueno, 1: Malo]				
Variable	**Coeficiente**	**Error Estándar**	**Z Value**	**Prob.**
Intercepto	-3.63	0.02	-150.56	0.00
W_DEPART_i	-0.97	0.04	-24.14	0.00
W_PLAZO_i	-0.65	0.09	-7.35	0.00
$W\_TIPO\_PREST_i$	-0.61	0.06	-10.71	0.00
$W\_SIT\_CLIENTE_i$	-2.27	0.19	-11.82	0.00
W_SEXO_i	-6.27	0.94	-6.67	0.00
$W\_INGRESO_i$	-0.98	0.05	-19.57	0.00
$W\_ANTIG\_LAB_i$	-0.61	0.11	-5.72	0.00
$W\_DEUDA\_SF_i$	-0.45	0.11	-4.13	0.00

Fuente: Elaboración propia, 2017.

El modelo de regresión logístico estima la probabilidad de incumplimiento del cliente y queda definido por variables cualitativas como el departamento en donde se desembolsó el crédito, el tipo de préstamo (sector al que pertenece el trabajador público que solicita el préstamo), la situación laboral del cliente, sexo; así como de las variables intrínsecas de la operación de préstamo como son el plazo, deuda en el sistema financiero además de la variable ingreso del trabajador (que considera lo que en neto ingresa a la cuenta de ahorros en el banco) y antigüedad laboral.

Se evidencia que estas variables independientes son significativas a por lo menos el 95 por ciento de confianza estadística, lo que indica que las variables influyen en la probabilidad de pago del prestatario. Asimismo, presentan coeficientes negativos, lo cual se explica por la relación que tiene el WOE con la probabilidad de default, siendo esta una relación negativa.

Específicamente, para obtener la interpretación de los coeficientes se reescribirá la expresión del modelo final de la siguiente manera:

Dada la especificación inicial del modelo de regresión logística

$$Prob[Y = 1] = p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}}$$

Se obtiene la expresión resultante es conocida como Logit:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

Si tomamos exponenciales, la expresión representaría el ratio *odds* y quedaría según se presenta a continuación:

$$\frac{p}{1-p} = e^{\beta_0} * e^{\beta_1 X_1} * e^{\beta_2 X_2} * \ldots$$

Es así que las estimaciones de los coeficientes del modelo muestran que las variables con las que queda definido el modelo influyen en términos del ratio *odds*, dado un incremento de una unidad en la variable señalada incrementaría el ratio según como sigue: (i) si se produce en la variable W_DEPART sería de 0,38 (resultado de $e^{(-0,97)}$, dado que $\beta_1 = -0,97$, cálculos similares se realizan para las variables independientes en adelante); (ii) si se produce en la variable W_PLAZO sería de 0,52; (iii) si se produce en la variable W_TIPO_PREST sería de 0,54; (iv) si se produce en la variable W_SIT_CLIENTE sería de 0,10; (v) si se produce en la

variable W_SEXO sería de 0,002; (vi) si se produce en la variable W_INGRESO (que equivale por ejemplo de pasar del rango de ingresos de 100-400 al 931-1400) sería de 0,38; (vii) si se produce en la variable W_ANTIG_LAB sería de 0,54; y finalmente, (viii) si se produce en la variable W_DEUDA_SF sería de 0.64.

Adicionalmente, las estimaciones muestran que medido a través de la probabilidad de incumplimiento y tomando en cuenta, por ejemplo, variaciones en la variable origen INGRESO (no considera el WOE respectivo), se obtiene que para un cliente que solicita el préstamo en el departamento de Loreto, con una solicitud de préstamo con plazo mayor a 40 meses, laborando en el sector Educación (Maestro), situación activo, de sexo femenino, con un rango de ingreso de S/ 930 a S/ 1.400, con antigüedad laboral mayor a 1 año y deuda promedio en el sistema financiero hasta S/ 6 mil, la probabilidad de *default* es 0,12. Si el ingreso estuviera en el rango de S/ 401 a S/ 800, la probabilidad de default sería de 0,23 incrementándose el riesgo crediticio en 11 puntos básicos.

2.1 Poder discriminatorio

Una vez construido el modelo de regresión logística se tiene los siguientes posibles casos:

- Se produce el *default* y el modelo lo clasifica como mal crédito (clasificación apropiada).
- Se produce el *default* y el modelo lo clasifica como buen crédito (error Tipo I).
- No se produce el *default* y el modelo lo clasifica como mal crédito (error Tipo II).
- No se produce el *default* y el modelo lo clasifica como buen crédito (clasificación apropiada).

Se plantea un *cut off* de 2,7%, denotando Ct y Ft, respectivamente, el número de casos correctamente y erróneamente clasificados como malos, por lo que la sensibilidad (Ht) y la falsa alarma (F) son:

$$H_T = \frac{C_T}{D}\,,\; F_T = \frac{W_T}{N-D}$$

La curva ROC muestra la sensibilidad contra las falsas alarmas. Cuanto más pronunciada es la curva ROC, es mejor, debido a que implica que hay pocas falsas alarmas en comparación con los malos créditos detectados correctamente. La curva ROC también puede ser interpretada como un *trade off* entre el error tipo I y error tipo II (Ht=1-E1 y Ft=E2).

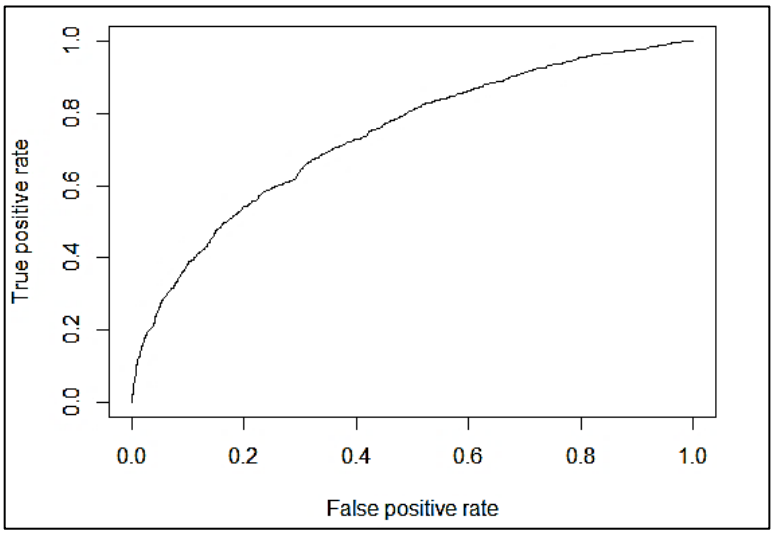
El área bajo la curva ROC es una medida de discriminación ampliamente utilizada y, para el modelo, es de 0,7379, lo que implica que el modelo tiene poder de discriminación aceptable, de acuerdo con lo señalado por Hosmer *et al.*(2013).

Tabla 12. Regla de decisión para el ROC

ROC	Calificación
Igual a 0,5	No hay discriminación
Mayor a 0,5 y menor a 0,7	Discriminación baja
Mayor o igual a 0,7 y menor a 0,8	Discriminación aceptable
Mayor o igual a 0,8 y menor a 0,9	Discriminación excelente
Mayor o igual a 0,9	Discriminación sobresaliente

Fuente: Hosmer *et al.*, 2013.

Gráfico 11. Curva ROC



Fuente: Elaboración propia, 2017.

No obstante, las ventajas de la curva ROC, según Servigny y Renault (2004), se tienen ciertas limitantes:

- La medida ROC está centrada en el ordenamiento en rango y, por lo tanto, solo se ocupa de la clasificación relativa. En términos de crédito, siempre y cuando el modelo produzca una clasificación correcta de las empresas en términos de probabilidades de incumplimiento, tendrá un buen coeficiente ROC, independientemente de si todas las empresas asignaron probabilidades mucho más bajas (o más) que sus valores reales. Por lo tanto, uno puede

tener un modelo que subestima el riesgo sustancialmente, pero todavía tiene un coeficiente ROC satisfactorio.

- ROC es una medida aceptable siempre y cuando la distribución de clase no sea sesgada. Este es el caso del crédito, donde la población que cumple sus pagos es mucho mayor que la que no cumple. Las curvas ROC pueden no ser la medida más adecuada en tales circunstancias.

Además, se realiza el análisis del KS, el cual mide el poder predictivo del modelo a través de la divergencia entre la distribución de buenos y malos

Tabla 13. Análisis K-S

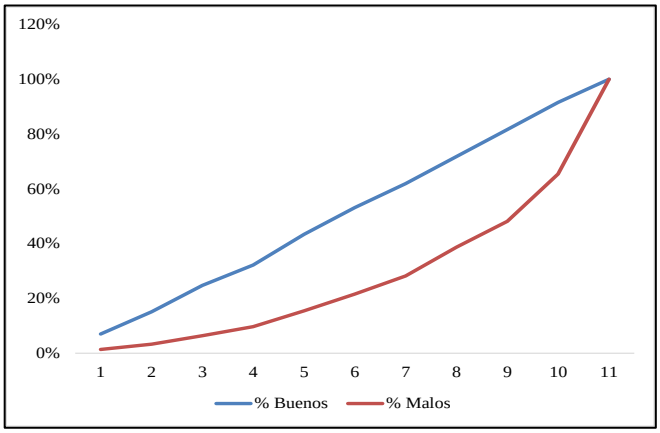
PD	Bueno	Malo	Total	% Total	PD	PD Acum.	% Buenos	% Malos	KS	Odds	Tasa aprob
<= 0.004	5,749	31	5,780	7%	0.5%	0.5%	7%	1%	6%	185.45	0.07
0.005 - 0.006	6,726	44	6,770	8%	0.6%	0.6%	15%	3%	12%	152.86	0.15
0.007 - 0.008	7,919	69	7,988	9%	0.9%	0.7%	25%	6%	18%	114.77	0.24
0.009 - 0.01	6,100	74	6,174	7%	1.2%	0.8%	32%	10%	22%	82.43	0.32
0.011 - 0.013	9,258	130	9,388	11%	1.4%	1.0%	43%	15%	28%	71.22	0.43
0.014 - 0.017	8,041	139	8,180	10%	1.7%	1.1%	53%	22%	32%	57.85	0.52
0.018 - 0.021	7,191	148	7,339	9%	2.0%	1.2%	62%	28%	34%	48.59	0.61
0.022 - 0.027	8,166	238	8,404	10%	2.8%	1.5%	72%	39%	33%	34.31	0.71
0.028 - 0.036	8,066	212	8,278	10%	2.6%	1.6%	82%	48%	33%	38.05	0.81
0.037 - 0.059	8,192	388	8,580	10%	4.5%	1.9%	92%	65%	26%	21.11	0.91
>= 0.06	6,970	781	7,751	9%	10.1%	2.7%	100%	100%	0%	8.92	1.00
Total	**82,378**	**2,254**	**84,632**	**100%**	**2.7%**						

Fuente: Elaboración propia, 2017

Se observa una divergencia entre buenos créditos y malos créditos obteniéndose un *KS* de 34%, el cual se encuentra dentro del rango satisfactorio entre 20% y 40%, según referencia de Mays (2004). De los resultados obtenidos, con el *cut off* de 2,7% para la probabilidad de incumplimiento, permite obtener una tasa de aprobación de solicitudes de crédito del 71%.

En base a este análisis es posible establecer objetivos de riesgo que está dispuesto a aceptar el BN de acuerdo a su perfil de riesgos.

Gráfico 12. Análisis K-S

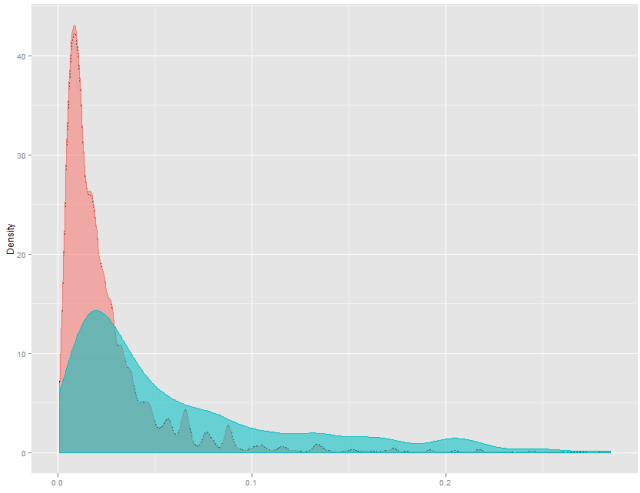


Fuente: Elaboración propia, 2017

2.2 Validación

Empleando la muestra de validación, se procederá a aplicar el modelo planteado, con la finalidad de revisar cuál hubiese sido el comportamiento de este para predecir los desembolsos realizados entre el periodo enero 2014 a diciembre 2014.

Gráfico 13. Distribución del probabilidad de *default* según buenos (rojo) y malos (verde)



Fuente: Elaboración propia, 2017.

La muestra de validación consta de 84.632 registros, de los cuales 82.378 son buenos y 2.254 han sido definidos como malos, esta muestra contiene similar número de buenos y malos créditos, respecto de la muestra de construcción, como se observó en la tabla 4.

Utilizando los coeficientes estimados por el modelo de regresión logística, se estimó la probabilidad de *default* en plantilla Excel. Realizados los cálculos, se obtuvo una sensibilidad de

61% (malos detectados por el modelo) y una especificidad de 72% (buenos detectados por el modelo).

Tabla 14. Validación del modelo (matriz de confusión)

Observación/ estimación		Estimación - modelo de regresión		
		Bueno	**Malo**	**Total**
Observación	**Bueno**	59.150	23.228	82.378
	Malo	873	1.381	2.254
	Total	60.023	24.609	84.632

Fuente: Elaboración propia, 2017.

Conclusiones y recomendaciones

1. Conclusiones

- Con el modelo de regresión logística se demuestra que la probabilidad de *default* para la cartera de créditos de consumo no revolvente del BN viene determinado por variables cualitativas como departamento donde se desembolsa el crédito, sector en el que labora el trabajador público, su situación laboral: activo o pensionista y sexo en combinación con variables cuantitativas como plazo del préstamo, antigüedad laboral, ingreso del trabajador y la deuda en el sistema financiero al momento de solicitar el crédito.

- El modelo de regresión logístico obtuvo un poder de discriminatorio aceptable medido a través del estadístico K-S y la curva ROC y área bajo la curva ROC siendo de 0,7379. Con la validación del modelo se obtuvo que el 61% de los créditos malos observados son estimados por el modelo.

- La identificación de las variables determinantes de la probabilidad de *default* permitirá una mejora en la gestión del riesgo crediticio del BN en el entendido que se desenvuelve en un entorno externo e interno que requerirían modificar su modelo de negocios actual (recupero de los préstamos a través de débito automático sobre la cuenta de haberes que posee el cliente en el BN).

2. Recomendaciones

- Implementar la técnica del *credit scoring* en el proceso de evaluación crediticia para los préstamos no revolventes del Banco de la Nación a través de un modelo de regresión logística, tomando en cuenta las ventajas que esta herramienta le ofrece.

- Dado que el presente trabajo de investigación es un primer acercamiento a una propuesta la aplicación del *credit scoring* en el Banco de la Nación, resulta necesario analizar el impacto que ocasionaría una eventual implementación que, además del entendimiento de las ventajas y desventajas que esta técnica ofrece, requerirá cambios en la mentalidad de los funcionarios de negocios.

- Enriquecer la estimación del modelo de regresión con información de desembolsos que abarque un ciclo económico.

- Explorar la incorporación en el modelo de variables macroeconómicas como por ejemplo las relacionadas a demanda interna, desempleo, producto bruto interno, tasas de interés de la política monetaria. Siguiendo a Carrera (2011), estimaciones de corte deberían extenderse a paneles que capturen hechos estilizados a nivel macroeconómico.

- Teniendo en consideración las limitaciones del *scoring* que utiliza el pasado para explicar el futuro, y la dinámica en las características o principales variables determinantes, se recomienda la revisión del modelo por lo menos con una periodicidad anual, con el objetivo de evitar el deterioro en el tiempo.

- Es preciso indicar que el presente trabajo de investigación no ha tenido por objetivo desarrollar la estimación del componente probabilidad de incumplimiento, satisfaciendo los requisitos mínimos que el regulador exige. No obstante, es un primer avance que, aunado al inicio de los esfuerzos para el desarrollo de los componentes Pérdida dado el incumplimiento y Exposición ante incumplimiento, podrían, en un futuro, permitirle optar por la solicitud de la aprobación por parte de la SBS para el cálculo de requerimiento de capital por riesgo crediticio a través de modelo interno.

Bibliografía

Anderson, Raymond (2007). *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation*. Estados Unidos de América: Oxford University Press Inc.

Banco de la Nación. (1994). Estatuto del Banco de la Nación – Texto actualizado. Fecha de consulta: 31.03.2017. Disponible en: http://www.bn.com.pe/nosotros/estatuto.asp

Carrera, César. (2011). "El canal del crédito bancario en el Perú: Evidencia y mecanismo de transmisión". En: *Banco Central de Reserva del Perú*. Diciembre 2011. Fecha de consulta: 14/06/2017. <http://www.bcrp.gob.pe/docs/Publicaciones/Revista-Estudios-Economicos/22/ree-22-carrera.pdf>.

De Servigny, Arnaud y Renault, Olivier (2004). *Measuring and Managing Credit Risk*. México: McGraw-Hill Companies, Inc.

Escalona, Arturo (2011). *Uso de Modelos Credit Scoring en Microfinanzas.* Tesis presentada como requisito parcial para obtener el grado de Maestro en Ciencias. Institución de enseñanza e investigación en ciencias agrícolas Campus Montecillo - México.

Hosmer, David; Lemeshow, Stanley; Sturdivant, Rodney (2013). *Applied Logistic Regression*. 3ra edición. Estados Unidos de América: John Wiley&Sons, Inc.

Lawrence, David y Arlene Solomon (2002). *Managing a Consumer Lending Business*. Estados Unidos de América: Solomon Lawrence Partners.

Maddala, G.S. (1996). *Introducción a la Econometría*. 2da. Edición. México: Prentice-Hall Hispanoamérica, S.A.

Mays E; Nuetzel P. (2004) *Scorecard Monitoring Reports*. In Mays, E. (ed.) *Credit Scoring for Risk Managers: The Handbook for Lenders*. Estados Unidos de América: South-Western Publishing: Mason, OH.

Samaniego, Reyes (2008). *El Riesgo de Crédito en el Marco del Acuerdo de Basilea II*. 1ra edición. España: Delta Publicaciones Universitarias.

Siddiqi, Naeem (2006). *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring*. Estados Unidos de América: John Wiley&Sons, Inc.

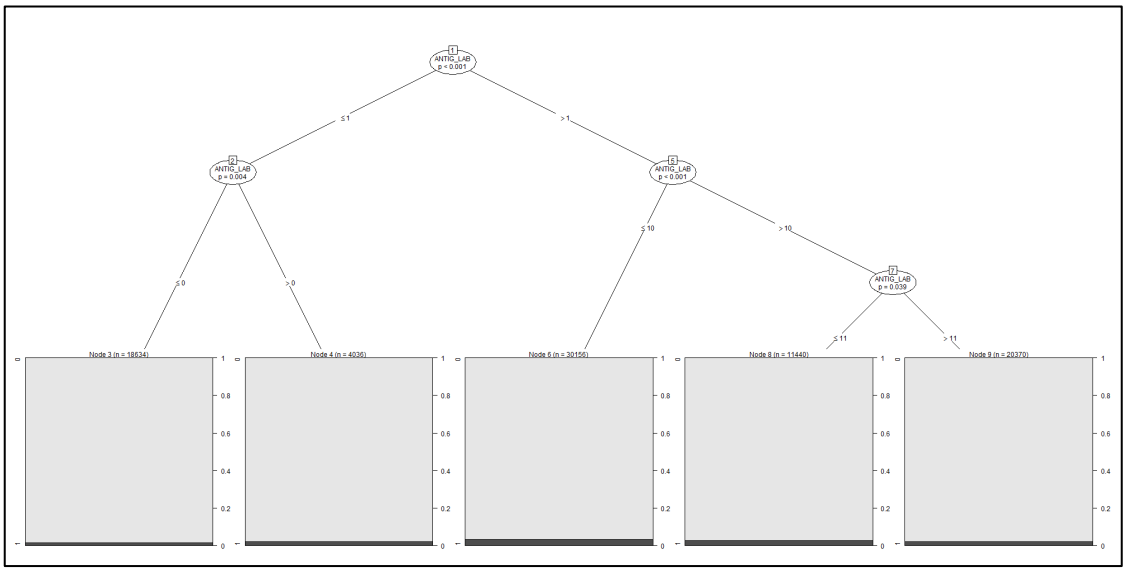
Superintendencia de Banca, Seguros y AFP (SBS). (2017). Sistema Financiero. Fecha de consulta: 15.03.2017. Disponible en: http://www.sbs.gob.pe/principal/categoria/sistema financiero/148/c-148

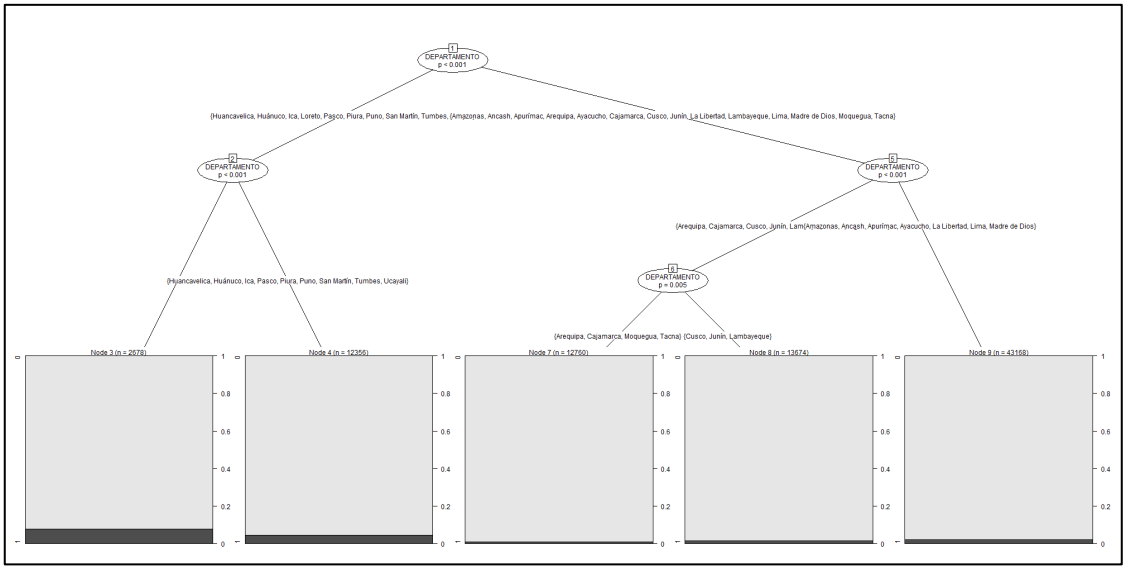
Anexos

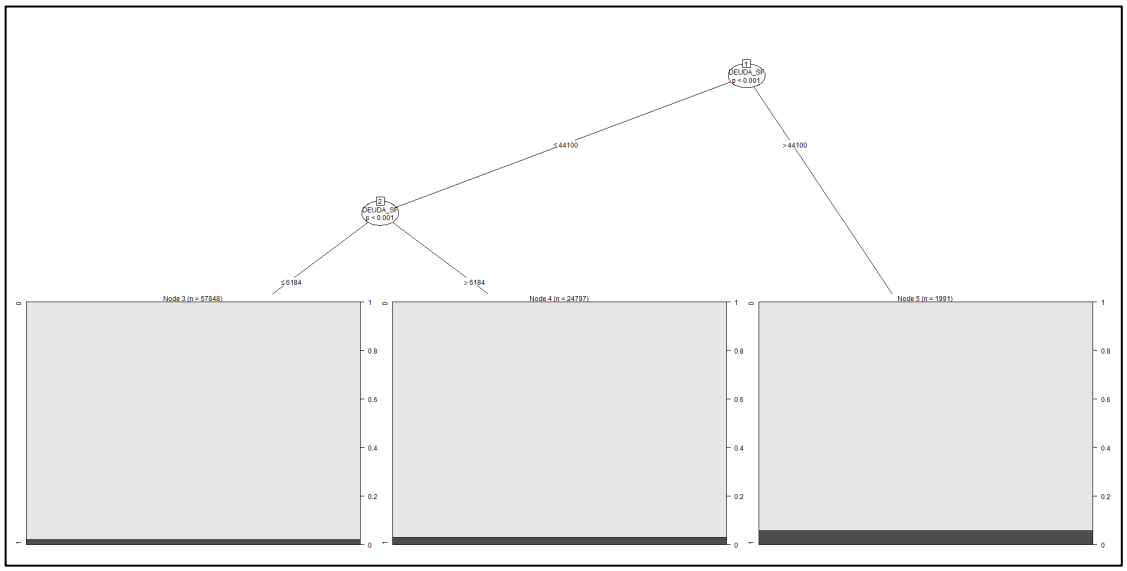
Anexo 1. Ranking del sistema bancario y el BN por créditos de consumo no revolvente y revolvente

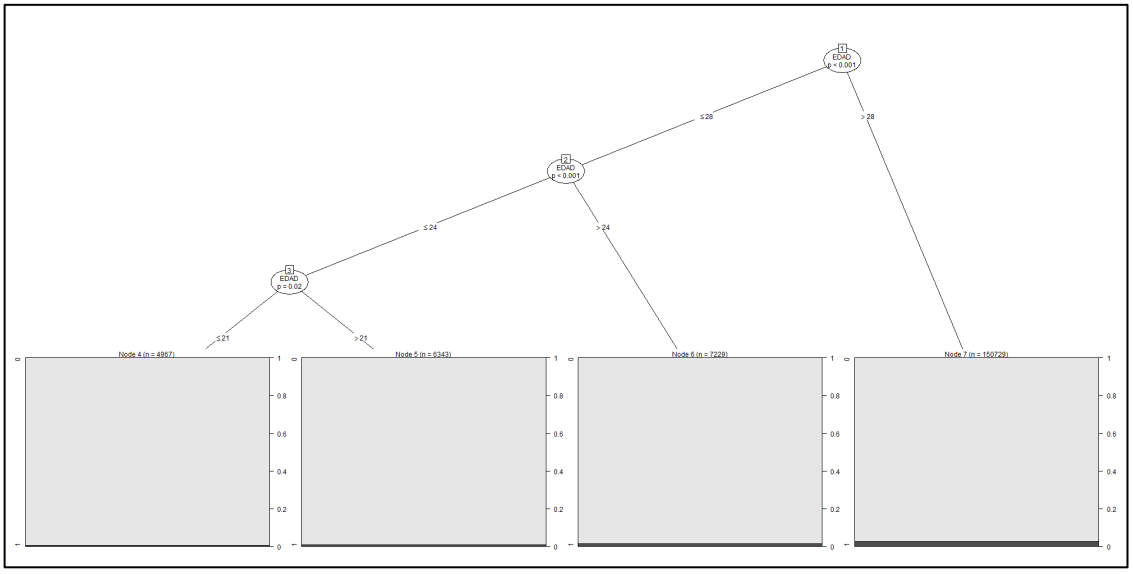
	BANCOS	Nov-15		Dic-15		Nov-16		Dic-16	
		Millones S/.	Part. %	Millones S/.	Part. %	Millones S/.	Part. %	Millones S/.	Part. %
1	CREDITO	10,310	24.2%	10,276	23.9%	10,698	23.2%	10,571	22.9%
2	INTERBANK	8,366	19.6%	8,416	19.6%	8,999	19.5%	8,889	19.3%
3	SCOTIABANK	5,634	13.2%	5,649	13.1%	6,419	13.9%	6,480	14.0%
4	CONTINENTAL	4,091	9.6%	4,073	9.5%	4,488	9.7%	4,422	9.6%
5	BN	**3,867**	9.1%	**3,867**	9.0%	**4,142**	9.0%	**4,175**	9.0%
6	FALLABELLA	3,689	8.6%	3,848	8.9%	3,831	8.3%	3,925	8.5%
7	RIPLEY	1,444	3.4%	1,541	3.6%	1,695	3.7%	1,776	3.8%
8	FINANCIERO	1,302	3.1%	1,315	3.1%	1,526	3.3%	1,555	3.4%
9	COMERCIO	858	2.0%	853	2.0%	939	2.0%	948	2.1%
10	BIF	780	1.8%	794	1.8%	937	2.0%	940	2.0%
11	GNB	899	2.1%	894	2.1%	902	2.0%	902	2.0%
12	MIBANCO	593	1.4%	604	1.4%	679	1.5%	686	1.5%
13	CENCOSUD	410	1.0%	444	1.0%	499	1.1%	517	1.1%
14	AZTECA	433	1.0%	430	1.0%	339	0.7%	360	0.8%
15	SANTANDER	1.2	0.0%	1.2	0.0%	0.8	0.0%	1.2	0.0%
16	CITIBANK	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	TOTAL	**42,677**	**100.0%**	**43,005**	**100.0%**	**46,094**	**100.0%**	**46,147**	**100.0%**

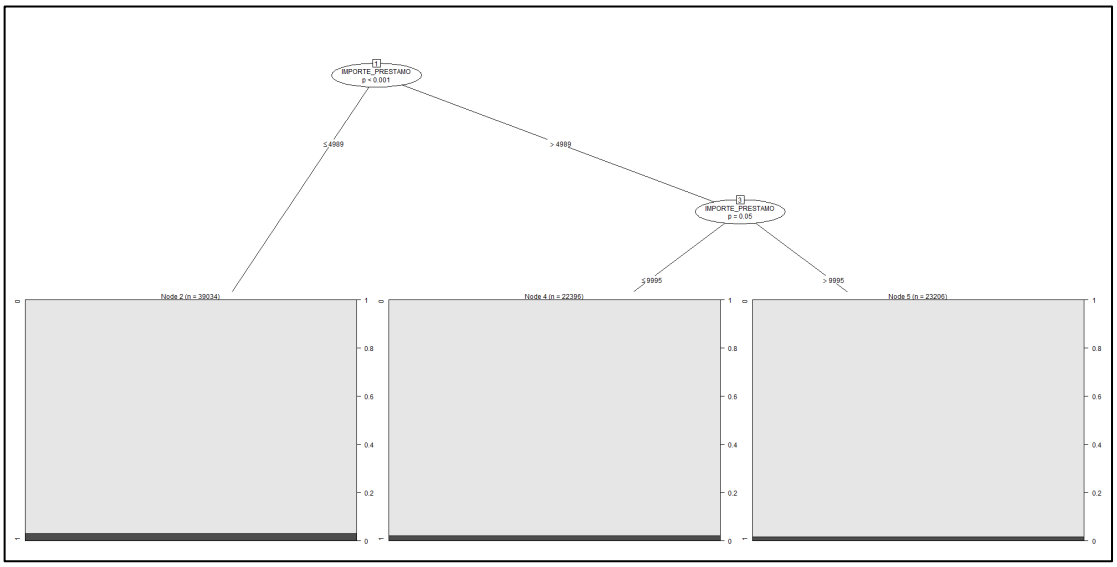
Anexo 2. Árboles de decisión para las variables independientes

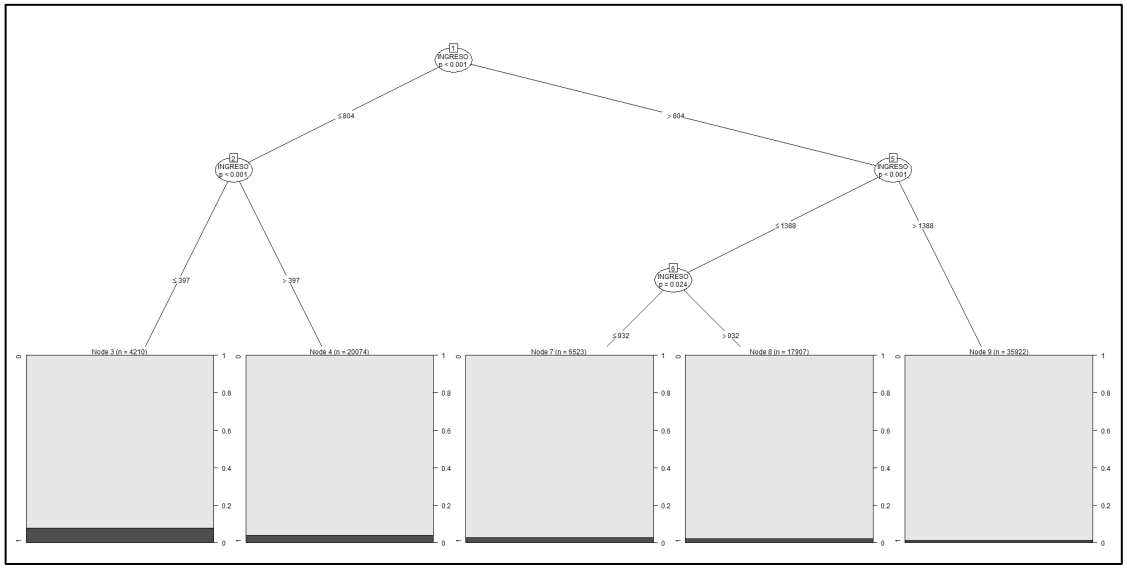


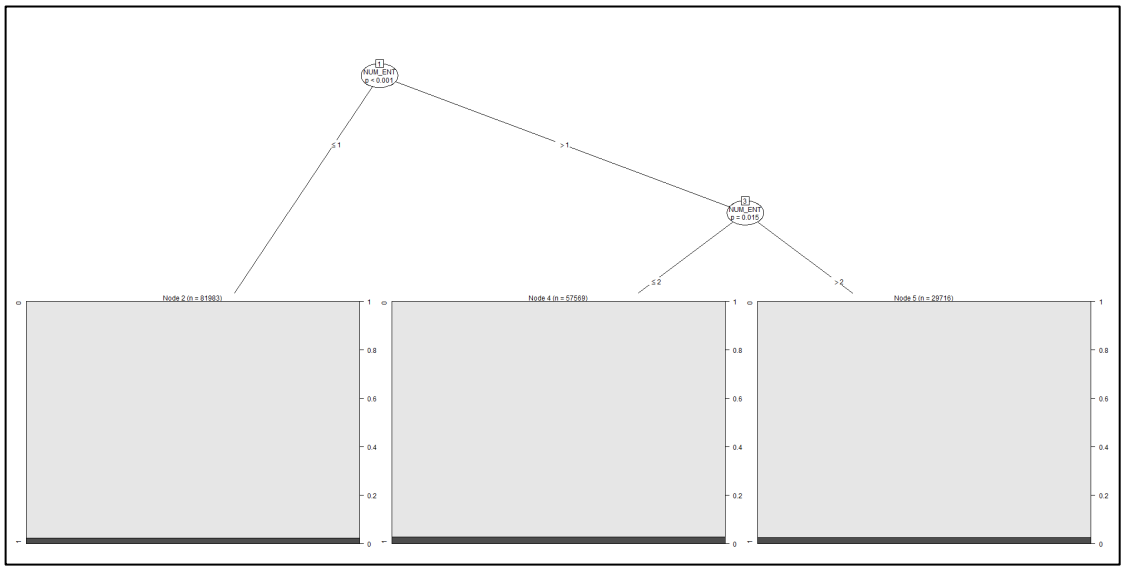


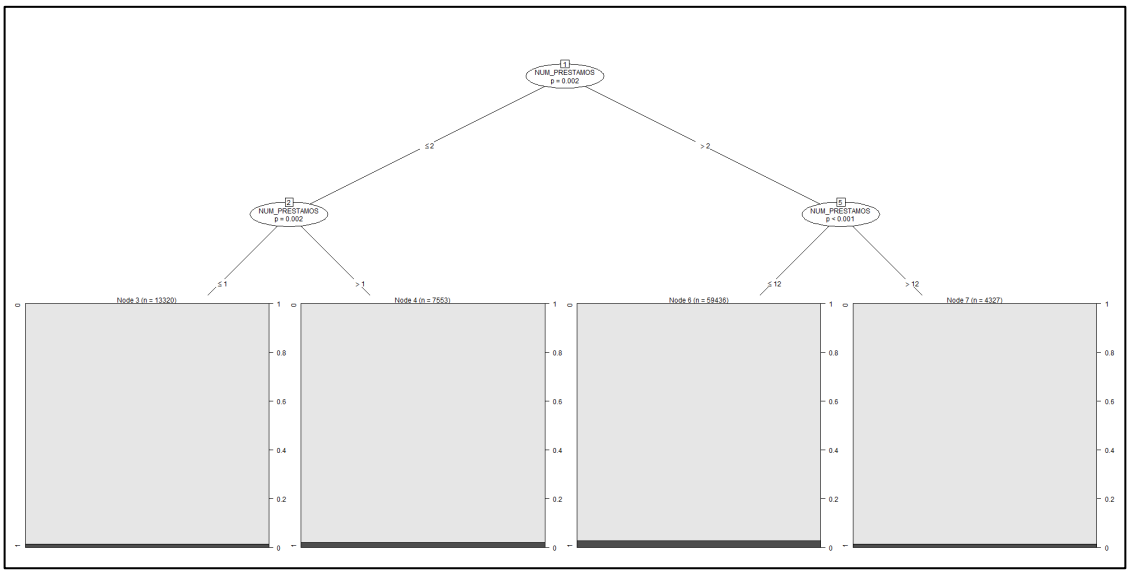

1
DEUDA_SP
p < 0.001
≤ 44100
> 44100
2
DEUDA_SP
p < 0.001
≤ 6184
> 6184
Node 3 (n = 57848)
Node 4 (n = 24797)
Node 5 (n = 1991)

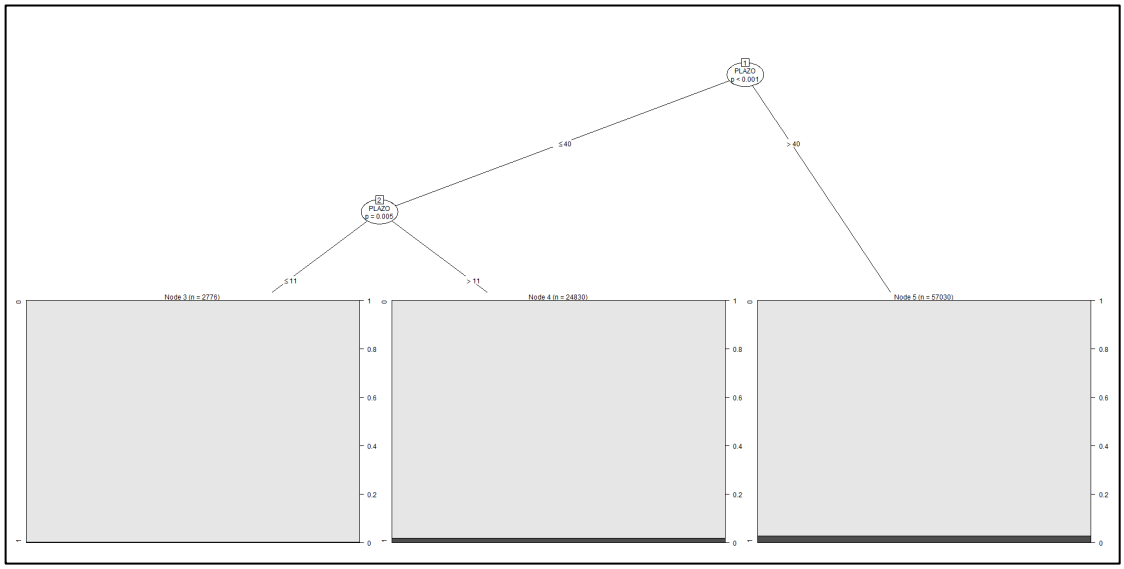

1
EDAD
p < 0.001
≤ 28
> 28
2
EDAD
p < 0.001
≤ 24
> 24
3
EDAD
p = 0.02
≤ 21
> 21
Node 4 (n = 4967)
Node 5 (n = 6343)
Node 6 (n = 7229)
Node 7 (n = 150729)

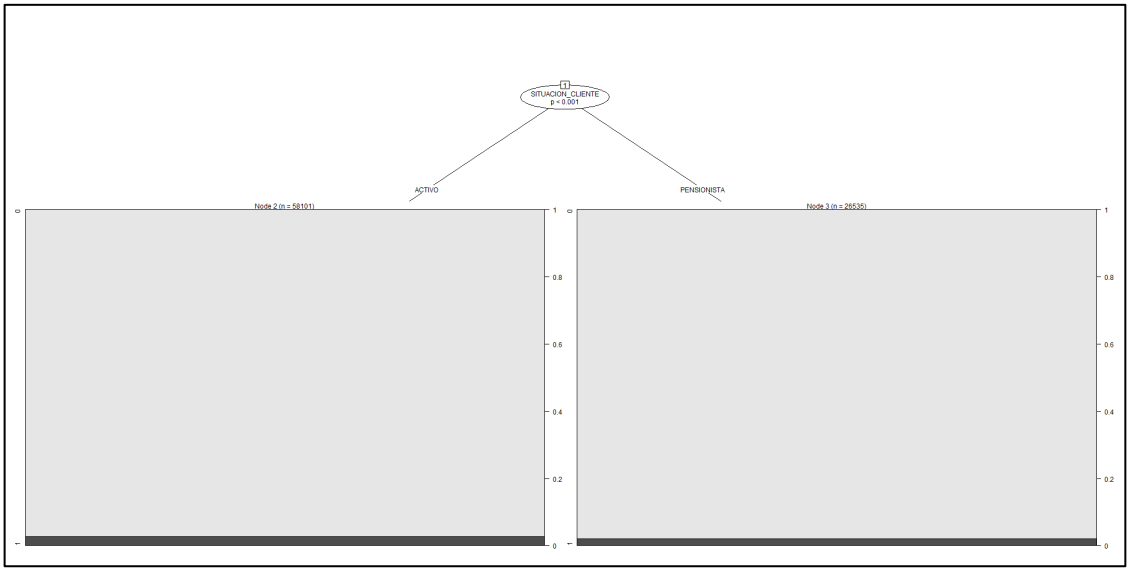

1
IMPORTE_PRESTAMO
p < 0.001
≤ 4989
> 4989
3
IMPORTE_PRESTAMO
p = 0.05
≤ 9995
> 9995
Node 2 (n = 39034)
Node 4 (n = 22395)
Node 5 (n = 23206)

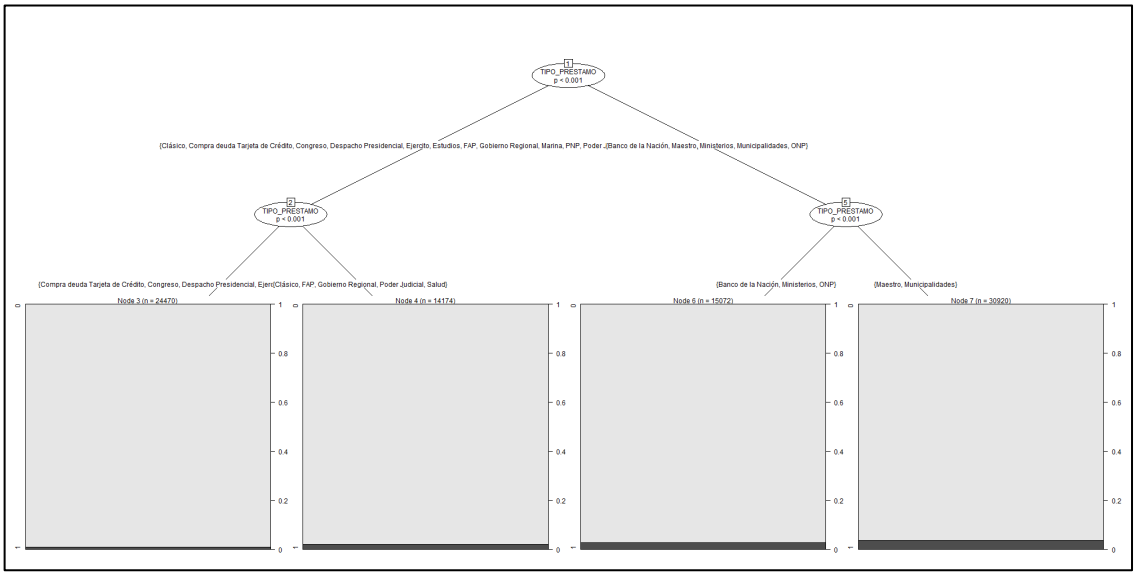

1
INGRESO
p < 0.001
≤ 804
> 804
2
INGRESO
p < 0.001
3
INGRESO
p < 0.001
≤ 397
> 397
≤ 1388
> 1388
5
INGRESO
p = 0.024
≤ 932
> 932
Node 3 (n = 4210)
Node 4 (n = 20074)
Node 7 (n = 6523)
Node 8 (n = 17907)
Node 9 (n = 35922)


1
NUM_ENT
p < 0.001
≤ 1
> 1
3
NUM_ENT
p = 0.015
≤ 2
> 2
Node 2 (n = 81983)
Node 4 (n = 57569)
Node 5 (n = 29716)
0
0.2
0.4
0.6
0.8
1


1
NUM_PRESTAMOS
p = 0.002
≤ 2
> 2
2
NUM_PRESTAMOS
p = 0.002
6
NUM_PRESTAMOS
p < 0.001
≤ 1
> 1
≤ 12
> 12
Node 3 (n = 13320)
Node 4 (n = 7553)
Node 6 (n = 59436)
Node 7 (n = 4327)
0
0.2
0.4
0.6
0.8
1


1
PLAZO
p < 0.001
≤ 40
> 60
2
PLAZO
p = 0.005
≤ 11
> 11
Node 3 (n = 2776)
Node 4 (n = 24830)
Node 5 (n = 57030)
1
0.8
0.6
0.4
0.2
0


1
SITUACION_CLIENTE
p < 0.001
ACTIVO
PENSIONISTA
Node 2 (n = 58101)
Node 3 (n = 26535)
1
0.8
0.6
0.4
0.2
0



Fuente: Elaboración propia, 2017.

Anexo 3. Código usado para el modelo de regresión

```
################################################
# Configuración espacio de trabajo
################################################
setwd("C:/Users/INTEL/Desktop/Tesis 2017/Scoring/Scoring")

################################################
# Carga de datos y exploración
################################################
datos.total<-read.table("datos_scoring.csv",header=T,sep=",")
head(datos.total)
str(datos.total) # muestra los tipos de variables
summary(datos.total) # Resumen de rango de valores
dim(datos.total)
length(unique(datos.total$ID)) # duplicado

# Transformar variable ID
datos.total$ID<-as.character(datos.total$ID)

# Exclusiones
datos.total<- subset(datos.total, DEF_MALO < 2) # Indeterminados
datos.total<- subset(datos.total, INGRESO <= 10000) # Valores extremos
datos.total<- subset(datos.total, DEUDA_SF <= 100000) # Valores extremos

# Transformando la variable dependiente de numérica a factor (categórica)
datos.total$DEF_MALO<-as.factor(datos.total$DEF_MALO)

# verificar niveles de datos categóricos
levels(datos.total$DEF_MALO)

# Quitar variables de la base
datos.total<- datos.total[,c(-14)]

# Datos categóricos
prop.table(table(datos.total$DEPARTAMENTO,datos.total$DEF_MALO)) # global
prop.table(table(datos.total$DEPARTAMENTO,datos.total$DEF_MALO),margin=1) # por
variable indep
prop.table(table(datos.total$DEPARTAMENTO,datos.total$DEF_MALO),margin=2) # por
variable dep-respuesta
barplot(table (datos.total$DEPARTAMENTO), col=c("blue","lightblue"),
main =" Diagrama de barras de las frecuencias absolutas \n de la variable \"
DEPARTAMENTO\"")
barplot(table(datos.total$DEF_MALO,datos.total$SEXO),beside=TRUE,names=c("Mujer","Ho
mbre"),col=c("red","blue"),legend.text=c("Buenos","Malos"))

# Datos continuos
hist(datos.total$IMPORTE_PRESTAMO)
hist(datos.total$EDAD)
hist(datos.total$PLAZO)
hist(datos.total$INGRESO)
hist(datos.total$ANTIG_LAB)
```

```
hist(datos.total$DEUDA_SF)
hist(datos.total$NUM_ENT)
hist(datos$NUM_ENT)

#% malos
table(datos.total$DEF_MALO)/nrow(datos.total)*100

# % malos x variable
tabla1=table(datos.total$ANTIG_LAB,datos.total$DEF_MALO)
tabla1
tabla2=prop.table(tabla1,margin=1)
tabla2

# Barras agrupadas
barplot(tabla2,col=2:6,beside = T,
xlab="Bueno = 0 y Malo = 1",
ylab="Proporción de Clientes",
main="Porcentaje por clasificación riesgo")


#Selección muestra: data de desarrollo y validación
#install.packages("caTools")
library(caTools)
dim(datos.total)
set.seed(123456)
muestra<- sample.split(datos.total, SplitRatio = 1/2)
datos<- subset(datos.total, muestra == TRUE)
datos.test<- subset(datos.total, muestra == FALSE)
dim(datos)
dim(datos.test)

#% malos base desarrollo
table(datos$DEF_MALO)/nrow(datos)*100

#% malos base validación
table(datos.test$DEF_MALO)/nrow(datos.test)*100

############################################################
# Árbol de clasificación
############################################################
# Build a conditional tree using the party package.
# install.packages("party")
library(party)
require(party, quietly = TRUE)
rpart<- ctree(DEF_MALO ~ INGRESO, data=datos)
rpart<- ctree(DEF_MALO ~ EDAD, data=datos)
rpart<- ctree(DEF_MALO ~ IMPORTE_PRESTAMO, data=datos)
rpart<- ctree(DEF_MALO ~ PLAZO, data=datos)
rpart<- ctree(DEF_MALO ~ ANTIG_LAB, data=datos)
rpart<- ctree(DEF_MALO ~ DEUDA_SF, data=datos)
rpart<- ctree(DEF_MALO ~ NUM_PRESTAMOS, data = datos)
rpart<- ctree(DEF_MALO ~ NUM_ENT, data = datos)
rpart<- ctree(DEF_MALO ~ SITUACION_CLIENTE, data = datos)
```

```
rpart<- ctree(DEF_MALO ~ SEXO, data = datos.total) # no sale relevante
rpart<- ctree(DEF_MALO ~ TIPO_PRESTAMO, data = datos)
rpart<- ctree(DEF_MALO ~ DEPARTAMENTO, data = datos)
print(rpart)
plot(rpart)

## Guardar gráfico con 3 o más variables (para 2 variables usar w=2000 y h=1000)
png(filename = "rpart_tree_scoring.png",
width = 2000,
height = 1000)
plot(rpart)
dev.off()

############################################################
# Cargando las bases con variables WoEs
############################################################
datos.woe.cons<-read.table("bd_cons.csv",header=T,sep=",")
datos.woe.valid<-read.table("bd_valid.csv",header=T,sep=",")
str(datos.woe.cons) # muestra los tipos de variables

############################################################
# Regresión logística binaria
############################################################
# Estimación del modelo
modelo_logit<- glm(DEF_MALO ~
W_DEPART+W_IMP_PRES+W_PLAZO+W_TIPO_PREST+W_SIT_CLIENTE+W_EDAD+
W_SEXO+W_INGRESO+W_ANTIG_LAB+W_NUM_ENT+W_DEUDA_SF+W_NUM_PRE
ST, family = binomial(), data = datos.woe.cons)
modelo_logit<- glm(DEF_MALO ~
W_DEPART+W_PLAZO+W_TIPO_PREST+W_SIT_CLIENTE+W_SEXO+W_INGRESO+
W_ANTIG_LAB+W_DEUDA_SF, family = binomial(), data = datos.woe.cons)
summary(modelo_logit)

# Probabilidades de default - base validación
pd_proy<- predict(modelo_logit, newdata=datos.woe.valid[,-1], type="response")

# Valores predichos de Y
ypred<- as.numeric((predict(modelo_logit, newdata=datos.woe.valid[,-1], type="response") >=
0.1) )

# Valores reales de Y
ytrue<- datos.woe.valid$DEF_MALO

## Matriz de confusión (error Clasificación)
modelo.mc <- table(ypred,ytrue)
modelo.mc

# AUC
install.packages("pROC")
library(pROC)
logit.analysis<- roc(response=ytrue, predictor=pd_proy)
logit.analysis$auc
```

```
# ROC
install.packages("ROCR")
library(ROCR)
pred<-prediction(pd_proy,ytrue)
perf<- performance(pred,"tpr","fpr")
plot(perf)

# Cálculodel KS
max(attr(perf,'y.values')[[1]]-attr(perf,'x.values')[[1]])

########################################
# Base validación con probproyectada #
########################################
Base_Validacion = data.frame(datos.woe.valid, prob_logit = pd_proy)
head(Base_Validacion)

# Exportando base en formato csv
write.csv(Base_Validacion, file="data_valid_completa.csv")

## Exportamos datos a csv o txt
write.csv(datos, file="data_desarrollo.csv")
write.csv(datos.test, file="data_validacion.csv")
write.table(datos.test, file="data_valid.txt")

# Exportar salidas
summary(modelo_logit)
Resumen = summary(modelo_logit)
capture.output(Resumen, file="Resumen_Logit.doc")
```

Fuente: Elaboración propia, 2017.

Nota biográfica

Paola Aracelli Roxana Tamayo Medrano

Nació en Lima, el 27 de agosto de 1986. Economista de la Universidad Nacional Mayor de San Marcos. Obtuvo la certificación internacional Chartered Risk Analyst (CRA) otorgada por la American Academy of Financial Management y cuenta con especialización en Gestión de Riesgos Financieros en el Tecnológico de Monterrey y Finanzas Corporativas en ESAN.

A la fecha, cuenta con experiencia de siete años en el sistema bancario, específicamente en evaluación financiera de proyectos, análisis de rentabilidad por productos, canales y clientes, proyección de estados financieros y gestión de riesgos crediticios y financieros. Participó en el proyecto de implementación del Sistema de Rentabilidad del Banco de la Nación (líder usuario) así como en la emisión del Primer Programa de Bonos Subordinados del Banco de la Nación (análisis de riesgo). Se desempeñó como analista y jefa de la Sección Estudios Financieros, como subgerente (e) de Riesgos Crediticios y Financieros del Banco de la Nación; y actualmente como subgerente (e) de Estudios Económicos y Financieros de la Gerencia de Finanzas y Contabilidad, funciones desempeñadas en el Banco de la Nación.