# UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

## ESCUELA DE POSGRADO

## MAESTRÍA EN ESTADÍSTICA APLICADA

# "REGRESIÓN BAYESIANA CON ENLACES ASIMÉTRICOS PARA LA CLASIFICACIÓN DE CLIENTES CON PROPENSIÓN A CAER EN MORA EN UNA ENTIDAD BANCARIA"


**Presentada por:**

**RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ**


**TESIS PARA OPTAR EL GRADO DE MAESTRO**

**MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA**


**Lima - Perú**

**2018**

# UNIVERSIDAD NACIONAL AGRARIA
# LA MOLINA
## ESCUELA DE POSGRADO

## MAESTRÍA EN ESTADÍSTICA APLICADA

# "REGRESIÓN BAYESIANA CON ENLACES ASIMÉTRICOS PARA LA CLASIFICACIÓN DE CLIENTES CON PROPENSIÓN A CAER EN MORA EN UNA ENTIDAD BANCARIA"

**TESIS PARA OPTAR EL GRADO DE**
**MAESTRO MAGISTER SCIENTIAE**

**Presentada por:**

**RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ**

**Sustentada y aprobada ante el siguiente jurado:**

Mg. Jesús Salinas Flores
**PRESIDENTE**

Mg. Jorge Chue Gallardo
**PATROCINADOR**

Mg. Raphael Valencia Chacón
**MIEMBRO**

Mg.Sc. Carlos López de Castilla Vásquez
**MIEMBRO**

# DEDICATORIA

A Fernando Fernández Gómez, mi padre, por todo el apoyo brindado durante toda mi vida.

A María Vásquez Maque, mi madre, por todo el apoyo brindado.

A Lesly Basualdo, mi esposa, amiga y compañera por su amor incondicional.

A Jeferson Suarez e Israel Diestra, mis grandes amigos por su apoyo en todo momento.

A toda mi familia.

# AGRADECIMIENTO

A Jaime Porras, coordinador de la Maestría de Estadística Aplicada de la Universidad Nacional Agraria La Molina, por todo el apoyo brindado durante toda la maestría y por la gran predisposición como persona en todo momento.

A Jorge Chue, catedrático de la Maestría de Estadística Aplicada de la Universidad Nacional Agraria La Molina, por su apoyo como asesor en la elaboración de la presente investigación.

A Enver Tarazona, catedrático de la Maestría de Estadística Aplicada de la Universidad Nacional Agraria La Molina, por su apoyo en la elaboración de la presente investigación.

A Jesús Salinas, catedrático de la Maestría de Estadística Aplicada de la Universidad Nacional Agraria La Molina, por su apoyo en la elaboración de la presente investigación.

A todos los que fueron mis profesores durante mis estudios de la Maestría de Estadística Aplicada en la Universidad Nacional Agraria La Molina, en reconocimiento a sus labores como catedráticos y guías personales.

A todas la personas que de alguna manera me brindaron su apoyo en la elaboración de esta investigación.

# I. ÍNDICE GENERAL

# ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

# ÍNDICE DE ANEXOS

# II. RESUMEN

En la actualidad las entidades bancarias conviven con clientes que no cumplen con sus obligaciones crediticias y se exceden del plazo estipulado acordado con el banco, a estos clientes se les denomina clientes morosos, por tal motivo el objetivo del presente trabajo es determinar el modelo de regresión binaria bayesiano con enlace asimétrico más adecuado para clasificar a los clientes que incumplirán sus pagos de sus tarjetas de crédito según sus probabilidades de mora en la entidad bancaria UNIBANK y haciendo uso de las variables más significativas. Se realizó un análisis comparativo entre los modelos de regresión bayesiana con enlaces asimétricos cloglog, power logit y scobit, y se determinó que el modelo de regresión binaria bayesiano con enlace asimétrico cloglog fue el más adecuado para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora, pues este modelo presentó un valor mucho mayor de sensibilidad que los modelos power logit y scobit, siendo las diferencias 8.5% y 9.1%, respectivamente.

**PALABRAS CLAVE**: Mora, tarjeta de crédito, regresión bayesiana, enlace asimétrico clolog, enlace asimétrico power logit, enlace asimétrico scobit.

# ABSTRACT

At present the banking entities coexist with clients that do not fulfill their credit obligations and exceed the stipulated term agreed with the bank, these clients are called delinquent clients, for that reason the objective of the present work is to determine the regression model Bayesian binary with asymmetric link more suitable to classify customers who will default their payments on their credit cards according to their probability of default in the bank UNIBANK and making use of the most significant variables. A comparative analysis was performed between the bayesian regression models with asymmetric cloglog, power logit and scobit links, and it was determined that the bayesian binary regression model with asymmetric link cloglog was the most adequate to classify clients who breach their credit obligations with their credit cards in the bank UNIBANK according to their probability of default, since this model presented a much greater value of sensitivity than the models power logit and scobit, being the differences 8.5% and 9.1%, respectively.



**KEY WORDS**: Default, credit card, bayesian regression, asymmetric clolog link, asymmetric power logit link, asymmetric scobit link.

# ABSTRAKT

Heute arbeiten Banken häufig mit den Kunden, die ihre Kreditverpflichtungen nicht erfüllen und die mit der Bank vereinbarte Frist überschreiten; solche Kunden werden als säumige Zahler bezeichnet. Das Ziel dieser Studie ist es dementsprechend, das Bayessche binäre Regressionsmodell mit asymmetrischer Verbindung auszuwählen, welches zum Klassifizieren solcher Kunden am besten geeignet wäre, die die Zahlungsfristen für ihre Kreditkarten versäumen; dies wird durch die Wahrscheinlichkeit des Zahlungsverzugs am Beispiel der UNIBANK durchgeführt, wobei die wichtigsten Kundendaten benutzt werden. Es wurde eine komparative Analyse solcher Bayesschen Regressionsmodelle wie cloglog, logit power und scobit durchgeführt, wobei es festgestellt wurde, dass das Bayessche binäre Regressionsmodell cloglog das geeignetste ist, um Kunden zu klassifizieren, die ihre Kreditkartenverpflichtungen nicht erfüllen, was durch die Wahrscheinlichkeit des Zahlungsversugs am Beispielder UNIBANK mit Bezug auf die wichtigsten Kundendaten durchgeführt wurde. Das ausgewählte Modell wies im Vergleich zu logit und power scobit einen viel höheren Wert der Empfindlichkeit aus, wobei die Differenz in der Empfindlichkeit entsprechend bei 8,5% und 9,1% liegt.

**KEYWORDS**: Säumige Kunden, Kreditkarte, Bayessche Regression, clolog, power logit, scobit.

# III. INTRODUCCIÓN

UNIBANK es una entidad bancaria local de Perú, con una trayectoria en el mercado de más de 30 años. Sus clientes se distribuyen en banca mayorista conformada por empresas y banca minorista conformada por personas, siendo la banca minorista el segmento que más contribuye a la utilidad del negocio, el cual se encuentra por encima del 60%.

El número de clientes de la banca minorista de UNIBANK en el Perú ha aumentado de forma moderada pero constante desde su creación. Sin embargo, teniendo en cuenta el crecimiento del país en los últimos años, se ha detectado un gran aumento de la demanda de tarjetas de crédito para este segmento.

Para el otorgamiento de una tarjeta de crédito al segmento de la banca minorista la entidad bancaria usa un modelo de regresión binaria con enlace simétrico logit para predecir la probabilidad de mora de un cliente en función a un conjunto de variables explicativas o predictoras. Sin embargo, este tipo de enlace por lo general no es aplicable cuando se tienen un mayor porcentaje de una de las categorías de la variable respuesta, tal es el caso de la presente investigación, en la que se tiene un mayor porcentaje de clientes que incumplieron el pago de su tarjeta de tarjeta de crédito, los cuales representan el 70% de clientes, frente a un 30% que cumplieron con sus pagos. Para solucionar este problema, se usaron los modelos de regresión binaria bayesiana con enlaces asimétricos log-log complementario o cloglog y los planteados por Prentice (1976) y Nagler (1994) que muestran enlaces logit asimetrizados llamados power logit y scobit.

La entidad bancaria UNIBANK será el principal beneficiario y usuario, pues obtuvo un modelo adecuado para poder clasificar a los clientes que incumplirán con sus obligaciones crediticias con sus tarjetas de crédito. Por otro lado, el segundo beneficiario será la Asociación de Bancos del Perú, pues le servirá como una herramienta a imitar y proponer a

las entidades bancarias del Perú como una de las mejores prácticas para controlar la morosidad. Finalmente, es importante mencionar que la presente investigación beneficiará a otros investigadores como tema de consulta y un caso de aplicación de los modelos de regresión binaria bayesiana con enlaces asimétricos en el sector bancario peruano.

El objetivo general del presente trabajo fue determinar el modelo de regresión binaria logística bayesiano con enlace asimétrico más adecuado para clasificar a los clientes que incumplirán sus pagos de sus tarjetas de crédito según sus probabilidades de mora en la entidad bancaria UNIBANK y haciendo uso de las variables más significativas.

El primer objetivo específico fue evaluar el desempeño de los modelos de regresión binaria bayesiana con enlaces asimétricos para la clasificación de clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora mediante el uso de indicadores bayesianos, sensibilidad y área bajo la curva ROC. Finalmente, el segundo objetivo específico fue determinar las variables que son más significativas para la clasificación de clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora.

# IV. REVISIÓN DE LA LITERATURA

## 4.1 Estudios de modelos estadísticos bayesianos con enlaces asimétricos

Dávila et al. (2015) utilizaron un modelo logístico bayesiano asimétrico para analizar qué factores afectan la probabilidad de éxito de los estudiantes en superar la asignatura de Matemáticas Empresariales en el primer intento a través del análisis de diferentes variables, correspondiente al Grado en Administración y Dirección de Empresas en la Universidad de Las Palmas de Gran Canaria. Se recogió la información de 279 estudiantes matriculados en Matemáticas Empresariales en el curso académico 2011-2012, siendo las variables independientes la opinión sobre la asignatura, utilidad del material de clase y el trabajo personal del estudiante en el curso, por otro lado la variable dependiente es aprobar la asignatura, esta variable toma el valor de 1 si el estudiante ha aprobado la asignatura (representa el 34.41%, es decir 96 estudiantes), y 0 en caso contrario (representa el 65.59%, es decir 183 estudiantes). Se usó un modelo de regresión logística clásico para la estimación de los parámetros y se compararon los resultados con el modelo de regresión logística bayesiano, teniendo este último la principal ventaja de incorporar el efecto de asimetría de los estudiantes que aprobaron y no aprobaron la asignatura. Los resultados mostraron que los factores que influyen en dicha probabilidad son: la asistencia a clases de teoría y prácticas, la valoración positiva del estudiante hacia la materia, el tipo de centro en que se cursaron los estudios preuniversitarios y la asistencia a clases de apoyo.

Pérez et al. (2014) usaron un modelo de regresión logístico bayesiano con enlace asimétrico para determinar la probabilidad que un tenedor de una póliza de seguros de un auto reporte un reclamo en una compañía de seguros en España. Los datos corresponden a una muestra 2,000 tenedores de una póliza de seguros de un auto y describen las características relacionadas con el asegurado, el vehículo y la póliza de seguros. La variable dependiente fue clasificada como 0, si el tenedor de la póliza no ha realizado

ningún reclamo (representa el 92.9%, es decir 1,859 casos) y 1 si realizó algún reclamo (representa el 7.1%, es decir 141 casos). Los resultados mostraron que el modelo con enlace asimétrico proporciona un mejor ajuste.

Bazán y Bayes (2010) comentaron que los modelos más conocidos con enlace simétrico son la regresión logística y la regresión probit, sin embargo este tipo de suposiciones son restrictivas y no aplicables cuando se tiene una mayor frecuencia de una de las respuestas binarias.

Bermúdez et al. (2008) desarrollaron modelos binarios con enlaces asimétricos para describir el comportamiento de los clientes hacia el fraude en el mercado de seguros de automóviles en España. Los datos corresponden a una muestra aleatoria de 10,000 reclamos de automóviles en España en el años 2000, estos fueron investigados por la compañía de seguros y fueron clasificados como legítimos al 98.99% (codificados usando ceros, 9899 casos) y fraudulentos al 1.01% (codificados usando unos, 101 casos). Se hace uso de un modelo con enlace asimétrico o skewed logit propuesto por Chen et al. (1999), los resultados mostraron que el uso de un enlace asimétrico mejora el porcentaje de clasificación de los casos fraudulentos.

Collet (2003), Czado y Santner (1992) y Chen et al. (2001) mencionan que los modelos con enlaces asimétricos pueden ser más apropiados que los modelos con enlaces simétricos en situaciones específicas.

### 4.2 Prueba Chi-Cuadrado de Pearson

La prueba Chi-Cuadrado de Pearson, mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis. También contrasta la hipótesis de que las variables categóricas son independientes, frente a la hipótesis alternativa de que una variable se distribuye de modo diferente para diversos niveles de la otra, mediante la presentación de los datos en tablas de contingencia.

$H_o$: Las dos variables son independientes

$H_a$: Las dos variables no son independientes

La fórmula del estadístico, la cual se aproxima a la distribución de la Chi-Cuadrado, es la siguiente:

$$\chi^2 = \sum_i \frac{(observadq_i - teorica_i)^2}{teorica_i} \quad (1)$$

Cuanto mayor sea el valor de $\chi^2$, menos verosímil es que la hipótesis sea correcta. De la misma forma, cuanto más se aproxima a cero el valor de Chi-Cuadrado, más ajustadas están ambas distribuciones. Los grados de libertad se calculan de la siguiente manera:

grados de libertad = (r-1)(k-1). Donde r es el número de filas y k el de columnas en una tabla de contingencia.

No se rechaza Ho cuando $\chi^2 < \chi_t^2(r-1)(k-1)$. En caso contrario sí se rechaza.

t representa el valor proporcionado por las tablas, según el nivel de significación estadística elegido.

### 4.3 Algoritmo de Boruta

La selección de variables es un aspecto importante en la construcción de modelos y ayudan mucho en la construcción de modelos de predicción libre de variables correlacionadas, prejuicios y el ruido no deseado. Muchas veces se tiene la idea de que con todas las variables se tiene el mejor modelo, sin embargo esto es una idea errónea. Las variables a menudo se encuentran correlacionadas y a menudo obstaculizan el logro de mayor precisión del modelo, para apoyar en este tipo de situaciones se usa el algoritmo de Boruta.

Liaw y Wiener (2002) mencionan que el algoritmo de Boruta se utiliza para la selección de variables previo a la etapa de realizar un modelo y funciona como un algoritmo de envoltura alrededor de bosques aleatorios.

Kursa y Rudnicki (2010) presentan los pasos para realizar el algoritmo de Boruta, los cuales son:

1- En primer lugar se añade aleatoriedad a los datos dados, mediante la creación de copias de todas las funciones (que se denominan funciones de sombra).

2- Mezclar los atributos añadidos para eliminar su correlación con la respuesta.

3- A continuación, se entrenan a un clasificador de bosques al azar en los datos extendidos establecidos y se aplica una medida característica importante (el valor predeterminado es la actual disminución media) para evaluar la importancia de cada función donde medias más altas son más importantes.

4- En cada iteración, se comprueba si una característica real tiene una importancia mayor que el de sus características de sombra (es decir si la función tiene una puntuación z más alta que la máxima puntuación z de sus características de sombra) y constantemente quita características donde se consideren altamente pocos importantes.

5- Por último, el algoritmo se detiene cuando todas las funciones son confirmadas o rechazadas o se alcanza un límite especificado de pistas forestales aleatorias.

El algoritmo de Boruta sigue un método de selección de características relevantes donde capta todas las características que se encuentran en algunas circunstancias relevantes para la variable resultado. En contraste, la mayoría de los algoritmos de selección de características tradicionales siguen un método óptimo mínimo en el que se basan en un pequeño subconjunto de características que proporciona un error mínimo en un clasificador elegido.

Mientras que en un modelo de random forest ajustado sobre un conjunto de datos, se puede recursivamente deshacerse de características en cada iteración que no se desempeñan bien en el proceso. De esta manera se obtendrá un subconjunto óptimo mínimo de características como el método mínimo del error del modelo de bosque aleatorio. Esto sucede mediante la selección de una versión sobre-podado del conjunto de datos de entrada, que a su vez, se deshace de algunas de las características relevantes.

Por otro lado, BORUTA encontrará todas las características que son fuerte o débilmente relevantes para la variable de decisión.

### 4.4 Métodos de Cadenas de Markov de Monte Carlo (MCMC)

En Bazán y Bayes (2010) se señala que la inferencia Bayesiana está basada en el análisis de la distribución a posteriori, pues esta contiene toda la información sobre el parámetro a ser estimado condicional a los datos. En inferencia bayesiana es útil resumir la información que está contenida en la distribución a posteriori y toman la forma de esperanzas de funciones particulares de los parámetros, tal como:

$$I = E[g(\theta)] = \int g(\theta) f(\theta|\mathrm{y}) d\theta \qquad (2)$$

Así, el problema general de inferencia bayesiana consiste en calcular estos valores esperados según la distribución a posteriori de $\theta$.

El problema en calcular este tipo de esperanzas, es que usualmente es muy complicado o imposible evaluar la expresión que se presenta en la ecuación 2 en forma analítica. Por este motivo se hace necesario utilizar métodos aproximados para obtener estas integrales, como por ejemplo los métodos conocidos como Cadenas de Markov de Monte Carlo (MCMC). Este método permite generar de manera iterativa una cadena de Markov para $\theta$ de tal manera que $f(\theta|\mathrm{y})$ sea una distribución ergódica estacionaria. Empezando en algún estado inicial $\theta o$ la idea es simular un número suficientemente grande M de transiciones bajo la cadena de Markov y registrar los correspondientes estados simulados $\theta j$. En Ross (1995) se muestra que la media ergódica es:

$$\hat{I} = \frac{1}{M} \sum_{j=1}^{M} g(\theta_j) \qquad (3)$$

converge a la integral I deseada, en donde $\hat{I}$ provee una buena aproximación para $I$.
Con esto, se puede precisar una cadena de Markov adecuada con la distribución a posteriori $f(\theta|\mathrm{y})$ como su distribución estacionaria.

### 4.5 Regresión Logística Bayesiana

#### 4.5.1 Regresión Logística Bayesiana con Enlaces simétricos

Novales (1993) mencionó que los modelos de elección discreta binaria tienen dos alternativas mutuamente excluyentes y son adecuados para analizar los factores determinantes de la probabilidad de éxito de un suceso.

Se considera un modelo de regresión binaria:

$$Y_i \sim Bernoulli(p_i) \tag{4}$$

$$p_i = F(x_i'\beta) \tag{5}$$

donde:

- $f(y) = p^y(1-p)^y$ es la función de probabilidad de $y$ con $y$ = {0, 1}
- $Y_i$ es una variable binaria tal que $Y_i$ = 1 ocurre con probabilidad $p_i$.
- $x'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ es el vector con "p" variables explicativas.
- $\beta = (\beta_1, \beta_2, \dots, \beta_P)$ es el vector de "p" coeficientes de regresión.
- $F(.)$ denota una función de distribución acumulada (fda). La función inversa $F(.)^{-1}$ es llamada función de enlace. Para una explicación más detallada puede consultarse Agresti (2002).

Si F es una función de distribución acumulada de una distribución simétrica, la función de enlace resultante es simétrica y tiene una forma simétrica alrededor de $p_i$ = 0.5.

En el caso de que F sea la función de distribución acumulada de una distribución logística se tiene el enlace logit:

$$F(t) = \frac{e^t}{1+e^t} \tag{6}$$

donde t = $x_i'\beta$

Entonces se tiene que la función de verosimilitud será dada por:

$$L(\beta) = \prod_{i=1}^{n}\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)^{y_i}\left(1 - \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)^{1-y_i} \quad (7)$$

Se considera que la priori $\beta$ que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ, se tendrá:

$$f(\beta) = \left(\frac{1}{2\pi}\right)^{n/2} e^{\beta'\Sigma^{-1}\beta} \quad (8)$$

Luego la distribución a posteriori estará dada por:

$$f(\beta|\mathrm{y}) \ \alpha \ L(\beta) f(\beta) \quad (9)$$

$$f(\beta|\mathrm{y}) \ \alpha \ \frac{e^{\sum_{i=1}^{n} y_i x_i'\beta}}{\prod_{i=1}^{n}(1+e^{x_i'\beta})} \ e^{\beta'\Sigma^{-1}\beta} \quad (10)$$

### 4.5.2 Regresión Logística Bayesiana con Enlaces Asimétricos

Chen et al. (1999) sostuvo que si la probabilidad de la respuesta binaria se aproxima a 0 en una tasa diferente que cuando se aproxima a 1, los enlaces simétricos para el ajuste de datos pueden ser inadecuados. En este caso, hay que considerar los enlaces asimétricos.

Uno de los enlaces usados es el enlace log-log complementario o cloglog, donde la función de distribución acumulada usada en el enlace corresponde a la distribución de Gumbel, cuya función de distribución no depende de ningún parámetro adicional desconocido. La distribución se muestra a continuación:

$$F(t) = 1 - (e)^{-e^t} \quad (11)$$

donde t = $x_i'\beta$

Entonces se tiene que la función de verosimilitud será dada por:

$$L(\beta) = \prod_{i=1}^{n}\left(1 - (e)^{-e^{x_i'\beta}}\right)^{y_i}\left((e)^{-e^{x_i'\beta}}\right)^{1-y_i} \quad (12)$$

Se considera que la priori $\beta$ que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ, se tendrá:

$$f(\beta) = (\frac{1}{2\pi})^{n/2} e^{\beta' \Sigma^{-1} \beta} \tag{13}$$

Luego la distribución a posteriori estará dada por:

$$f(\beta|\mathrm{y}) \ \alpha \ L(\beta) f(\beta) \tag{14}$$

$$f(\beta|\mathrm{y}) \ \alpha \ \frac{\prod_{i=1}^{n} (1-(e)^{-y_i e^{x_i' \beta}})}{e^{\sum_{i=1}^{n} (1-y_i) e^{x_i' \beta}}} \ e^{\beta' \Sigma^{-1} \beta} \tag{15}$$

La distribución a posteriori haciendo uso del enlace cloglog no presenta una forma conocida por lo que es necesario usar métodos estadísticos computacionales como los Métodos de Cadenas de Markov de Monte Carlo.

Prentice (1976) y Nagler (1994), plantean enlaces logit asimetrizados que a diferencia del enlace cloglog incorporan un parámetro $\lambda$, de este modo incluye al enlace logit como caso especial cuando el parámetro incluido es $\lambda$ igual a 1.

El primer enlace logit asimetrizado es denominado power logit, siendo su función de distribución acumulada la siguiente:

$$F(t) = (1 + e^{t})^{-\lambda} \ \ \lambda > 0 \tag{16}$$

donde t = $x_i' \beta$

Entonces se tiene que la función de verosimilitud será dada por:

$$L(\beta, \lambda) = \prod_{i=1}^{n} (1 + \mathrm{e}^{x_i' \beta})^{-\lambda y_i} (1 - (1 + \mathrm{e}^{x_i' \beta})^{-\lambda})^{(1-y_i)} \tag{17}$$

Se considera que la priori $\beta$ que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ y la priori para $\lambda$ que sigue una distribución gamma con parámetros $\alpha, \delta$; se tendrá:

$$f(\beta) = (\frac{1}{2\pi})^{n/2} e^{\beta' \Sigma^{-1} \beta} \quad (18)$$

$$f(\lambda) = \frac{\delta}{\Gamma(\alpha)} (\delta\lambda)^{\alpha-1} e^{-\delta\lambda} \quad (19)$$

$$\lambda > 0, \alpha > 0, \delta > 0$$

Luego la distribución a posteriori estará dada por:

$$f(\beta, \lambda|\text{y}) \text{ α } L(\beta, \lambda) f(\beta) f(\lambda) \quad (20)$$

$$f(\beta|\text{y}) \text{ α } (\prod_{i=1}^{n} (1 + \text{e}^{x_i'\beta})^{-\lambda y_i} (1 - \left(1 + \text{e}^{x_i'\beta}\right)^{-\lambda})^{(1-y_i)}) e^{\beta' \Sigma^{-1} \beta} (\delta\lambda)^{\alpha-1} e^{-\delta\lambda} \quad (21)$$

Por otro lado, un segundo enlace logit asimetrizado es denominado enlace scobit, siendo su función de distribución acumulada la siguiente:

$$F(t) = 1 - (1 + e^t)^{-\lambda} \;\; \lambda > 0 \quad (22)$$

donde t = $x_i'\beta$

Entonces se tiene que la función de verosimilitud será dada por:

$$L(\beta, \lambda) = \prod_{i=1}^{n} (1 - (1 + \text{e}^{x_i'\beta})^{-\lambda})^{y_i} ((1 + \text{e}^{x_i'\beta})^{-\lambda})^{(1-y_i)} \quad (23)$$

Se considera que la priori $\beta$ que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ y la priori para $\lambda$ que sigue una distribución gamma con parámetros $\alpha, \delta$; se tendrá:

$$f(\beta) = (\frac{1}{2\pi})^{n/2} e^{\beta' \Sigma^{-1} \beta} \quad (24)$$

$$f(\lambda) = \frac{\delta}{\Gamma(\alpha)} (\delta\lambda)^{\alpha-1} e^{-\delta\lambda} \quad (25)$$

$$\lambda > 0, \alpha > 0, \delta > 0$$

Luego la distribución a posteriori estará dada por:

$$f(\beta,\lambda|\mathrm{y})\ \alpha\ L(\beta,\lambda)f(\beta)f(\lambda) \tag{26}$$

$$f(\beta|\mathrm{y})\ \alpha\ (\prod_{i=1}^{n}(\ 1-\left(1+\mathrm{e}^{x_i'\beta}\right)^{-\lambda})^{y_i}(\left(1+\mathrm{e}^{x_i'\beta}\right)^{-\lambda})^{(1-y_i)})e^{\beta'\Sigma^{-1}\beta}(\delta\lambda)^{\alpha-1}e^{-\delta\lambda} \tag{27}$$

Como se puede apreciar en (21) y (27), las distribuciones a posteriori haciendo uso de los enlaces power logit y scobit tampoco presentan una forma conocida por lo que es necesario usar también métodos estadísticos computacionales como los Métodos de Cadenas de Markov de Monte Carlo.

### 4.6 Comparación de modelos

#### 4.6.1 Indicadores estadísticos bayesianos

En esta sección se revisarán diferentes criterios para la comparación de modelos que ayudará a decidir qué modelo es el más apropiado.

Los indicadores estadísticos bayesianos se basan en la media a posteriori del desvío $E[D(a,b,\lambda,\theta)]$, donde $D(a,b\ ,\lambda,\theta)$ es una medida de ajuste que se aproxima utilizando la salida de la simulación MCMC de la distribución a posteriori. Para una explicación más detallada puede consultarse Brooks (2002), Carlin y Luis (2001) y Spiegelhalter et al. (2002).

$$E[D(a,b,\lambda,\theta)] = -2\ln(p(y|a,b,\lambda,\theta) = -2\sum_{i=1}^{n}\ln P(Y_{ij}=y_{ij}|a,b,\lambda,\theta) \tag{27}$$

La aproximación de $E[D(a,b,\lambda,\theta)]$ queda expresada por:

$$D_{bar} = \frac{1}{G}\sum_{i=1}^{G} D(a^g,b^g,\lambda^g,\theta^g) \tag{28}$$

Donde el índice $g$ indica el $g$-ésimo valor simulado de un total de $G$ simulaciones.

Sea p es el número de parámetros en el modelo, N es el total de observaciones y $\rho_D$ el número efectivo de parámetros, se define $\rho_D$ como:

$$\rho_D = E[D(a,b,\lambda,\theta)] - D[E(a),E(b),E(\lambda),E(\theta)] \tag{29}$$

donde $D[E(a), E(b), E(\lambda), E(\theta)]$ es el desvío de la media a posteriori obtenido cuando se evalúa la función desvío en la media a posteriori de los parámetros, el cual se estima por:

$$D_{hat} = D(\frac{1}{G}\sum_{i=1}^{G} a^g, \frac{1}{G}\sum_{i=1}^{G} b^g, \frac{1}{G}\sum_{i=1}^{G} \lambda^g, \frac{1}{G}\sum_{i=1}^{G} \theta^g) \quad (30)$$

Haciendo uso de (28), (29) y (30), se presentan los indicadores bayesianos para la comparación de modelos en la inferencia bayesiana:

- Criterio de información de desviación (DIC), fue propuesto por Spiegelhalter et al. (2002):

$$\widehat{DIC} = Dbar + \widehat{\rho_D} = 2\,Dbar - Dhat \quad (31)$$

- El esperado del criterio de información de Akaike (EAIC), propuesto por Carlin y Luis (2001) y Brooks (2002):

$$\widehat{EAIC} = Dbar + 2p \quad (32)$$

- El esperado del criterio de información de Schwarz o Bayesiano (EBIC), propuesto por Carlin y Luis (2001) y Brooks (2002):

$$\widehat{EBIC} = Dbar + plogN \quad (33)$$

Para comparar dos o más modelos alternativos, el modelo que presente mejor ajuste al conjunto de datos será el modelo que presente el menor valor de los indicadores DIC, EAIC y EBIC.

### 4.6.2 Tabla de clasificación

La tabla de clasificación muestra la distribución de valores observados y estimados. Los valores observados son los valores reales y los valores estimados se obtienen a partir del modelo estadístico bayesiano.

López I. y Píta S. (2001) mencionan que la capacidad de que el modelo estime el suceso de interés de cuyo valor es 1, se denomina sensibilidad. Por el contrario, la capacidad de que nuestro modelo no estime el suceso de interés cuyo valor es 0, se denomina especificidad. Para la presente investigación se usará la sensibilidad como medida de precisión del modelo estadístico bayesiano.

**Cuadro Nº 1:** Tabla de Clasificación

| **Observado** | | **Pronosticado** | | |
|---|---|---|---|---|
| | | A | | Porcentaje |
| | | 1 | 0 | correcto |
| A | 1 | a | b | a/(a+b) |
| | 0 | c | d | d/(c+d) |
| Porcentaje Global | | | | (a+d)/(a+b+c+d) |

ELABORACIÓN: Propia

La sensibilidad y especificidad se calculan se calcula de la siguiente manera:

- Sensibilidad = a/(a+b), indica la capacidad o probabilidad que tiene un modelo para clasificar correctamente la categoría de interés de la variable dependiente. También llamada precisión positiva.
- Especificidad = d/(c+d), indica la capacidad o probabilidad que tiene un modelo para clasificar correctamente la categoría que no es de interés de la variable dependiente. También llamada precisión negativa.
- Precisión = (a+d)/(a+b+c+d), indica la capacidad o probabilidad que tiene un modelo para clasificar correctamente de manera global la variable dependiente.
- Falsos positivos = c/(c+d), indica la proporción de casos negativos que fueron clasificados incorrectamente como positivos.
- Falsos negativos = b/(a+b), indica la proporción de casos positivos que fueron clasificados incorrectamente como negativos.

El modelo que presente mayor sensibilidad es el modelo más adecuado.

### 4.6.3 Curva ROC

La curva ROC (Receiver Operating Characteristic), indica que cuanto más alejada esté de la diagonal principal cuya área es 0.5, mejor es el método de diagnóstico, ya que la curva ROC ideal sería con un área igual a 1. En cambio, cuanto más cercana esté a dicha diagonal principal peor será el método de diagnóstico. Pérez (2015) menciona que la diagonal principal es la que corresponde al peor test de diagnóstico y tiene un área bajo la curva de 0.5. Adicionalmente se han establecido los siguientes los intervalos para los valores de la curva ROC:

- [0.5 - 0.6>: Test malo
- [0.6 - 0.75>: Test regular
- [0.75 - 0.9>: Test bueno
- [0.9 - 0.97>: Test muy bueno
- [0.97 - 1>: Test excelente

Se plantean las siguientes hipótesis:

$H_0$: el área bajo la curva ROC es igual a 0.5
$H_1$: el área bajo la curva ROC no es igual a 0.5

Si se rechaza Ho asociado a un p-valor implica que el modelo ajustado es el adecuado, pues se rechaza que el área bajo la curva ROC es igual a 0.5, lo cual ya se explicó a inicios del punto 4.6.3.

# V. MATERIALES Y MÉTODOS

## 5.1 Hipótesis

### a- Hipótesis General

- El modelo de regresión binaria bayesiano con enlace asimétrico cloglog es el más adecuado para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según sus probabilidades de mora, pues presenta una precisión superior a los modelos con enlaces asimétricos power logit y scobit.

### b- Hipótesis Específicas

- Los modelos de regresión binaria bayesiana con enlaces asimétricos presenta un buen desempeño para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK alcanzando indicadores bayesianos, sensibilidad y área bajo la curva ROC significativos.

- Las variables situación de la casa, edad, máxima antigüedad con tarjeta de crédito, tenencia de producto pasivo, ingreso mensual, línea de tarjeta de crédito en el sistema financiero, saldo deudor en el sistema financiero, score de aprobación, abono de pago de haberes, máxima clasificación SBS, apalancamiento y número de veces sueldo ayudan a la clasificación de clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora.

## 5.2 Tipo de investigación

La presente investigación fue de corte transversal porque se tomó la muestra de clientes en un momento determinado, es decir a diciembre de 2015; fue descriptiva o correlacional porque ayudó a describir los datos de los clientes permitiendo identificar las variables que caracterizan de manera significativa los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito; y es inferencial porque a partir del estudio de la muestra de clientes que se tiene ayudó a aplicarlo a toda la cartera de clientes de la entidad bancaria.

## 5.3 Población en estudio

### 5.3.1 Universo

El universo fue la población de clientes de la entidad bancaria que posee el producto tarjeta de crédito.

### 5.3.2 Unidad de análisis

La unidad de análisis fue el cliente de la entidad bancaria que posee el producto tarjeta de crédito.

### 5.3.3 Población objetivo

La población a estudiar fueron los clientes de la entidad bancaria que posee el producto tarjeta de crédito con una antigüedad de 12 meses.

## 5.4 Fuentes de información

### 5.4.1 Datawarehouse de la entidad bancaria

El datawarehouse de la entidad bancaria es una base de datos donde se encuentra la información sociodemográfica, socioeconómica y financiera de los clientes.

### 5.4.2 Reporte Crediticio Consolidado (R.C.C.)

El Reporte Crediticio Consolidado (R.C.C.), es el reporte que emite la Superintendencia de Banca, Seguros y AFP (S.B.S.) de manera mensual de todos los deudores de las entidades financieras que reportan información.

### 5.4.3 Definición de variables

Las variable dependiente es la Mora60 e indica si los clientes de la entidad bancaria cayeron en mora en sus tarjetas de crédito con más de 60 días en el transcurso de un año luego de haberle otorgado el crédito (0: No Mora60, 1: Si Mora60).

$$Y_i = f(x) = \begin{cases} 1, & si\ P(Y_i = 1) = P_i, si\ tiene\ mora\ 60\ dias \\ 0, & si\ P(Y_i = 0) = 1 - P_i, si\ no\ tiene\ mora\ de\ 60\ dias \end{cases}$$

Las variables independientes fueron:

- VAR01: Situación de la casa. Los valores que toman son “1” para indicar que la casa donde vive el cliente es propia y “0” para indicar que no es propia.

- VAR02: Edad. Los valores que toman son “1” para indicar que la edad del clientes es menor a 29 años denominado como jóvenes y “0” para indicar que la edad del cliente es mayor o igual a 29 años o denominado como adultos.

- VAR03: Máxima antigüedad con tarjeta de crédito. Los valores que toman son “1” para indicar que la máxima antigüedad del cliente con tarjeta de crédito en el sistema financiero es mayor o igual a 12 meses y “0” para indicar que la máxima antigüedad del cliente con tarjeta de crédito en el sistema financiero es menor a 12 meses.

- VAR04: Tenencia de producto pasivo. Los valores que toman son “1” para indicar el número de meses que el cliente ha tenido algún producto pasivo es mayor o igual a 12 meses durante los 12 meses antes de la aprobación del crédito y “0” para indicar el número de meses que el cliente ha tenido algún producto pasivo es menor a 12 meses durante los 12 meses antes de la aprobación del crédito.

- VAR05: Ingreso mensual. Los valores que toman son “1” para indicar que el ingreso mensual del cliente es mayor o igual a 2,086 Soles o llamado ingreso alto y “0” para indicar que el ingreso mensual del cliente es menor a 2,086 Soles o llamado ingreso bajo.

- VAR06: Línea de tarjeta de crédito en el sistema financiero. Los valores que toman son “1” para indicar que la línea de tarjeta de crédito en el sistema financiero del cliente es mayor o igual a 7,691.12 Soles o llamada línea alta y “0” para indicar que la línea de tarjeta de crédito en el sistema financiero del cliente es menor a 7,691.12 Soles o llamada línea baja.

- VAR07: Saldo deudor en el sistema financiero. Los valores que toman son “1” para indicar que el saldo deudor promedio total en el sistema financiero es mayor o igual a 334.23 Soles o saldo alto y “0” para indicar que el saldo deudor promedio total en el sistema financiero es menor a 334.23 Soles o saldo bajo.

- VAR08: Score de aprobación. Los valores que toman son “1” para indicar que el score con el que fue aprobada la tarjeta de crédito en el banco es mayor o igual a 171 o score alto y “0” para indicar que el score con el que fue aprobada la tarjeta de crédito en el banco es menor a 171 o score bajo.

- VAR09: Abono de pago de haberes. Los valores que toman son “1” para indicar si tuvo abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito y “0” para indicar que no tuvo abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito.

- VAR010: Máxima clasificación SBS. Los valores que toman son “1” para indicar que la máxima clasificación de riesgos dada por la Superintendencia de Banca y Seguros durante los 12 meses antes de la aprobación del crédito es Normal y “0” para indicar que la máxima clasificación de riesgos dada por la Superintendencia de Banca y Seguros durante los 12 meses antes de la aprobación del crédito son CPP, Deficiente, Dudoso o Pérdida.

- VAR011: Apalancamiento. Se calcula como el monto de saldo deudor promedio total en el Sistema Financiero entre el Ingreso mensual, esto quiere decir transformándolo a nivel de empresa cuánto de patrimonio tienes para afrontar tus pasivos (deudas), a nivel de los

datos sería cuánto de poder adquisitivo tienes para afrontar tus deudas. Los valores que toman son “1” para indicar que el apalancamiento es mayor o igual a 1.19 y “0” para indicar que indicar que es menor a 1.19.

- VAR012: Número de veces sueldo. Se calcula como la línea de crédito del cliente entre el ingreso mensual del cliente. Es la cantidad de veces el su sueldo del cliente reflejándose en la línea de crédito. Los valores que toman son “1” para indicar que el número de veces sueldo es mayor o igual a 3.35 y “0” para indicar que indicar que el número de veces sueldo es menor a 3.35.

### 5.4.4 Diseño de muestreo y preparación de los datos

Mediante un muestreo aleatorio simple se obtuvo una muestra aleatoria de 5,000 clientes de la entidad bancaria. El 70% (3,500 clientes) de la muestra se usó como la muestra de construcción del modelo estadístico, mientras que el 30% (1,500 clientes) se usó como muestra para la validación del modelo estadístico.

## 5.5 Procedimiento estadístico

El procedimiento estadístico fue el siguiente:

1- Análisis exploratorio univariado de la variable dependiente y las variables independientes.

2- Análisis exploratorio bivariado de la variable dependiente Mora60 con respectos a las variables independientes.

3- Análisis de significancia bivariado de la variable dependiente Mora60 con respectos a las variables independientes. Para ello se utilizó la prueba Chi-Cuadrado para determinar si existe dependencia entre la variable Mora60 y las variables independientes.

4- Selección de las variables independientes más importantes mediante el algoritmo de Boruta.

5- División la muestra en construcción, la cual sirve para la estimación del modelo y será del 70%; y muestra de validación correspondiente al 30%.

6- Aplicación de los modelos de elección discreta binaria bayesiana con enlaces asimétricos cloglog, power logit y scobit.

7- Comparación de los modelos cloglog, power logit y scobit mediante los indicadores bayesianos Deviance Information Criterion (DIC), el Esperado del Criterio de Información de Akaike (EAIC) y el Esperado del Criterio de Información de Schwarz o Bayesiano (EBIC), el modelo que presente mejor ajuste al conjunto de datos será el modelo que presente el menor valor en los indicadores.

8- Comparación de los modelos cloglog, power logit y scobit usando la sensibilidad de la tabla de clasificación y el área bajo la curva ROC como medidas de precisión, siendo el modelo que presente mayor sensibilidad y área bajo la curva ROC el más adecuado.

### 5.6 Paquete estadístico

Los paquetes estadísticos que se usaron fueron:

- R versión 3.3.2
- RStudio versión 1.0.136

# VI. RESULTADOS Y DISCUSIÓN

## 6.1 Análisis descriptivo de las variables

### 6.1.1 Análisis exploratorio univariado de las variables

Se realizó el análisis exploratorio univariado de la variable dependiente Mora60 y de las variables independientes que corresponden a los 5,000 clientes de la entidad bancaria UNIBANK.

- Para la variable MORA60, los clientes de la entidad bancaria que si tienen mora 60 representan el 70% (3,500 clientes) frente a un 30% (1,500 clientes) que no tienen mora 60. (Ver Anexo 1, Cuadro N° 2).

- Para la situación de la casa, en su mayoría los clientes de la entidad bancaria no tienen casa propia, quienes representan el 71.4% (3,571 clientes) frente a un 28.6% (1,429 clientes) que si tiene casa propia. (Ver Anexo 1, Cuadro N° 3).

- Para la variable edad, en su mayoría los clientes de la entidad bancaria son adultos o con edad mayor o igual a 29 años, quienes representan el 77.7% (3,883 clientes) frente a un 22.3% (1,117 clientes) que son jóvenes o tienen menos de 29 años de edad. (Ver Anexo 1, Cuadro N° 4).

- Para la variable máxima antigüedad con tarjeta de crédito, los clientes de la entidad bancaria que tienen menos de 12 meses de antigüedad con tarjeta de crédito en el sistema financiero representan el 49.2% (2,459 clientes) frente a un 50.8% (2,541 clientes) que tienen 12 meses o más de antigüedad con tarjeta de crédito en el sistema financiero. No existe mucha diferencia entre los clientes que tienen menos 12 de meses y los que tienen

12 meses a más de antigüedad con tarjeta de crédito en el sistema financiero. (Ver Anexo 1, Cuadro N° 5).

- Para la variable tenencia de producto pasivo, en su mayoría los clientes de la entidad bancaria que tienen 12 meses o más con algún producto pasivo durante los 12 meses antes de la aprobación del crédito representan el 78.6% (3,928 clientes) frente a un 21.4% (1,072 clientes) que tienen menos de 12 meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito. (Ver Anexo 1, Cuadro N° 6).

- Para la variable ingreso mensual, los clientes de la entidad bancaria de ingresos bajos o ganan menos de 2,086 soles al mes representan el 40.6% (2,030 clientes) frente a un 59.4% (2,970 clientes) que tienen ingresos altos o ganan más de 2,086 soles al mes, siendo la diferencia casi de 20 puntos porcentuales. (Ver Anexo 1, Cuadro N° 7).

- Para la variable línea de tarjeta de crédito en el sistema financiero, en su mayoría los clientes de la entidad bancaria que tienen línea baja de tarjeta de crédito en el sistema financiero o menor a 7,691.12 soles representan el 70.2% (3,510 clientes) frente a un 29.8% (1,490 clientes) que tienen línea alta de tarjeta de crédito en el sistema financiero o de 7,691.12 soles a más. (Ver Anexo 1, Cuadro N° 8).

- Para la variable saldo deudor en el sistema financiero, en su mayoría los clientes de la entidad bancaria que tienen un monto de saldo deudor promedio total alto en el sistema financiero o de 334.23 soles a más representan el 74.5% (3,725 clientes) frente a un 25.5% (1,275 clientes) que tienen un monto de saldo deudor promedio total bajo en el sistema financiero o menor a 334.23 soles. (Ver Anexo 1, Cuadro N° 9).

- Para la variable score de aprobación, en su mayoría los clientes de la entidad bancaria que tienen score alto con el que fue aprobada la tarjeta de crédito en el banco o 171 a más de score representan el 91.2% (4,561 clientes) frente a un 8.8% (439 clientes) que tienen score bajo con el que fue aprobada la tarjeta de crédito en el banco o menor a 171 de score. (Ver Anexo 1, Cuadro N° 10).

- Para la variable abono de pago de haberes, en su mayoría los clientes que tuvieron abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito

representan el 72.7% (3,634 clientes) frente a un 27.3% (1,366 clientes) que no tuvieron abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito. (Ver Anexo 1, Cuadro N° 11).

- Para la variable máxima clasificación SBS, en su mayoría los clientes de la entidad bancaria que tienen máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito igual a normal o están al día en sus pagos representan el 62.8% (3,142 clientes) frente a un 37.2% (1,858 clientes) que tienen máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito igual a CPP, deficiente, dudoso y pérdida o no están atrasados en sus pagos. (Ver Anexo 1, Cuadro N° 12).

- Para la variable apalancamiento, los clientes de la entidad bancaria que tienen menos de 1.19 de apalancamiento representan el 56% (2,800 clientes) frente a un 44% (2,200 clientes) que 1.19 o más de apalancamiento. (Ver Anexo 1, Cuadro N° 13).

- Para la variable número de veces sueldo, en su mayoría los clientes que tuvieron número de veces sueldo menos a 3.35 representan el 76.9% (3,845 clientes) frente a un 23.1% (1,155 clientes) que tuvieron número de veces sueldo de 3.35 a más. (Ver Anexo 1, Cuadro N° 14).

### 6.1.2 Análisis exploratorio bivariado de las variable dependiente y las variables independientes

Se realizó el análisis exploratorio bivariado entre la variable dependiente Mora60 y las variables independientes que corresponden a los 5,000 clientes de la entidad bancaria UNIBANK.

- Como se puede apreciar en la figura Nº 1 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que no tienen casa propia (72.2%), que aquellos que si tienen casa propia (64.5%). Es decir, que existen diferencias porcentuales entre la situación de la casa que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable situación de la casa es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 2 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que son jóvenes (78.8%), que aquellos que son adultos (67.5%). Es decir, que existen diferencias porcentuales entre la edad que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable edad del cliente es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 3 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 en la que su máxima antigüedad con tarjeta de crédito en el sistema financiero es menor a 12 meses (79.9%), que aquellos en la que su máxima antigüedad con tarjeta de crédito en el sistema financiero es de 12 meses a más (60.4%). Es decir, que existen diferencias porcentuales entre la máxima antigüedad con tarjeta de crédito en el sistema financiero que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable máxima antigüedad con tarjeta de crédito en el sistema financiero es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 4 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 en la que su número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito es menor a 12 meses (82.8%), que aquellos en la que su número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito es de 12 meses a más (66.5%). Es decir, que existen diferencias porcentuales entre el número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 5 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que tienen ingresos bajos (81.6%), que aquellos que tienen ingresos altos (62.1%). Es decir, que existen diferencias porcentuales entre el ingreso mensual que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable ingreso mensual del cliente es influyente con respecto a la variable Mora60.
- Como se puede apreciar en la figura Nº 6 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que tienen líneas bajas (75.4%), que aquellos que tienen líneas altas (57.4%). Es decir, que existen diferencias porcentuales el entre monto de línea de tarjeta de

crédito en el sistema financiero que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable monto de línea de tarjeta de crédito en el sistema financiero es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 7 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que tienen saldos altos (72.4%), que aquellos que tienen saldos bajos (63%). Es decir, que existen diferencias porcentuales entre el monto de saldo deudor promedio total en el sistema financiero que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable monto de saldo deudor promedio total en el sistema financiero es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 8 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que tienen score bajos (88.8%), que aquellos que tienen score altos (68.2%). Es decir, que existen diferencias porcentuales entre el score con el que fue aprobada la tarjeta de crédito en el banco que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable score con el que fue aprobada la tarjeta de crédito en el banco es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 9 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que no tienen abono de pago de haberes en el banco (73.3%), que aquellos que tienen si tienen abono de pago de haberes en el banco (68.8%). Es decir, que existen diferencias porcentuales entre los que no se les abonó y si se les abonó el pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 10 (ver Anexo 2), existe un mayor porcentaje de clientes con mora 60 que tienen clasificación SBS CPP, deficiente, dudoso y pérdida (75.6%), que aquellos que tienen clasificación SBS normal (66.7%). Es decir, que existen diferencias porcentuales entre la variable máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable máxima clasificación de riesgos SBS durante

los 12 meses antes de la aprobación del crédito es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 11 (ver Anexo 2), existe un mayor porcentaje de clientes con apalancamiento igual o más a 1.19 (77.6%), que aquellos clientes que tienen apalancamiento menor a 1.19 (64%). Es decir, que existen diferencias porcentuales entre la variable apalancamiento que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable apalancamiento es influyente con respecto a la variable Mora60.

- Como se puede apreciar en la figura Nº 12 (ver Anexo 2), existe un mayor porcentaje de clientes con número de veces sueldo menor a 3.35 (73.2%), que aquellos clientes que número de veces sueldo igual o más de 3.35 (59.3%). Es decir, que existen diferencias porcentuales entre la variable número de veces sueldo que tienen los clientes con mora 60. Por tal motivo, puede hacer pensar que la variable número de veces sueldo es influyente con respecto a la variable Mora60.

### 6.1.3 Análisis de significancia entre la variable dependiente y las variables independientes

Se realizó el análisis de significancia mediante la prueba Chi-Cuadrado de Pearson entre la variable Mora60 y las variables independientes situación de la casa, edad del cliente, máxima antigüedad con tarjeta de crédito en el sistema financiero, número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito, ingreso mensual del cliente, monto de línea de tarjeta de crédito en el sistema financiero, monto de saldo deudor promedio total en el sistema financiero, score con el que fue aprobada la tarjeta de crédito en el banco, si tuvo abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito, máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito, apalancamiento y número de veces sueldo.

El cuadro Nº 15 presenta la significancia entre la variable Mora60 y las variables independientes haciendo uso de la prueba Chi-Cuadrado de Pearson. En este cuadro se puede apreciar que todas las variables independientes tienen un nivel de significancia menor a 0.05, por lo que se rechaza la hipótesis nula de independencia con respecto a la variable Mora60 y se puede concluir que con un nivel de confianza del 95% la variable

Mora60 presenta dependencia con respecto a las variables independientes situación de la casa, edad, máxima antigüedad con TC, tenencia de producto pasivo, ingreso mensual, línea de TC en el SF, saldo deudor en el SF, score de aprobación, abono de pago de haberes, máxima clasificación SBS, apalancamiento y número de veces sueldo.

**Cuadro N° 15**: Prueba Chi-Cuadrado de Pearson

| Variables vs Morosidad | Valor | gl | Sig. Asintótica (bilateral) |
|---|---|---|---|
| Situación de la casa | 28.61 | 1 | 0.00 |
| Edad | 52.83 | 1 | 0.00 |
| Máxima antigüedad con TC | 226.31 | 1 | 0.00 |
| Tenencia de producto pasivo | 107.06 | 1 | 0.00 |
| Ingreso mensual | 219.95 | 1 | 0.00 |
| Línea de TC en el SF | 160.91 | 1 | 0.00 |
| Saldo deudor en el SF | 40.16 | 1 | 0.00 |
| Score de aprobación | 81.33 | 1 | 0.00 |
| Abono de pago de haberes | 9.63 | 1 | 0.00 |
| Máxima clasificación SBS | 43.61 | 1 | 0.00 |
| Apalancamiento | 107.80 | 1 | 0.00 |
| Número de veces sueldo | 81.77 | 1 | 0.00 |

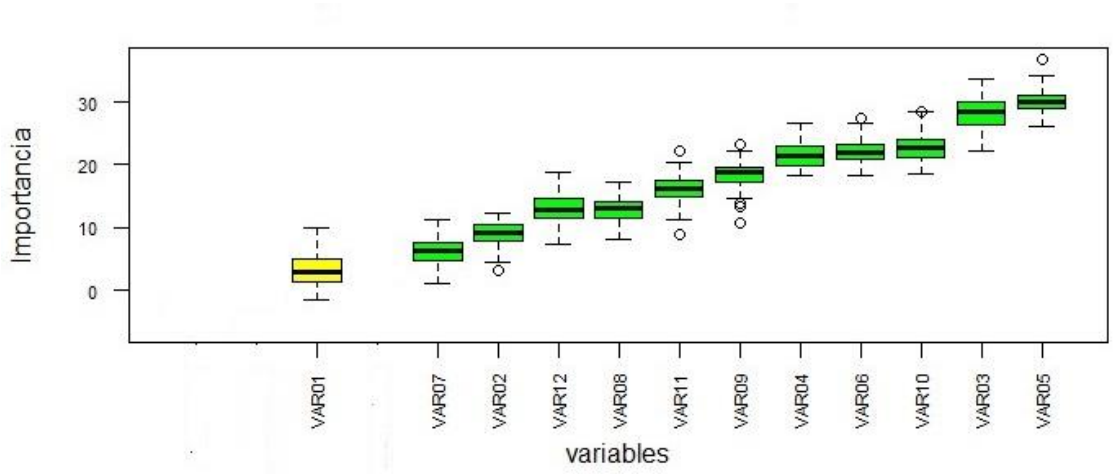
FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

### 6.2 Selección de variables independientes

Se usó el algoritmo de boruta para la selección de variables y se encontró según la figura Nº 13 que las variables independientes situación de la casa, edad, máxima antigüedad con TC, tenencia de producto pasivo, ingreso mensual, línea de TC en el SF, saldo deudor en el SF, score de aprobación, abono de pago de haberes, máxima clasificación SBS, apalancamiento y número de veces sueldo, son importantes y serán ingresadas al modelo de regresión con enlaces asimétricos.

**Figura N° 13:** Importancia de las variables independientes




### 6.3 Modelos de regresión logístico bayesiano con enlaces asimétricos

Antes de usar los modelos de regresión logística bayesiana con enlaces asimétricos se procedió a dividir a muestra de clientes en muestra de construcción y muestra de validación (revisar punto 5.4.4).

#### 6.3.1 Modelos de regresión logístico bayesiano con enlace asimétrico cloglog

En el cuadro Nº 16 se muestra el resultado del modelo de regresión logístico bayesiano con enlace asimétrico cloglog. La columna Mean B representa los coeficientes del modelo, los cuales tienen una relación no lineal con la probabilidad de que un cliente presente mora 60. La columna 2.50% es el límite inferior del intervalo de confianza al 95% de confianza para la columna Mean B y la columna 97.50% es el límite superior del intervalo de confianza al 95% de confianza para la columna Mean B.

**Cuadro N° 16**: Modelo de regresión logístico bayesiano con enlace asimétrico cloglog

| Variables | Mean B | 2.50% | 97.50% |
|---|---|---|---|
| Constante | 1.3 | 1.2 | 1.4 |
| Edad (X2) | 0.1 | 0 | 0.2 |
| Máxima antigüedad con TC (X3) | -0.3 | -0.4 | -0.3 |
| Tenencia de producto pasivo (X4) | -0.4 | -0.5 | -0.3 |
| Ingreso mensual (X5) | -0.4 | -0.5 | -0.3 |
| Línea de TC en el SF (X6) | -0.1 | -0.3 | 0 |
| Saldo deudor en el SF (X7) | 0.2 | 0 | 0.3 |
| Score de aprobación (X8) | -0.1 | -0.2 | 0 |
| Abono de pago de haberes (X9) | -0.2 | -0.3 | -0.1 |
| Máxima clasificación SBS (X10) | -0.3 | -0.4 | -0.3 |
| Apalancamiento (X11) | 0.1 | 0 | 0.2 |
| Número de veces sueldo (X12) | -0.2 | -0.4 | -0.1 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

De este modo, el modelo de regresión logística bayesiano con enlace asimétrico cloglog se presentará con 12 variables, siendo la ecuación para determinar las probabilidades de mora de los clientes que incumplirán sus pagos de sus tarjetas de crédito en la entidad bancaria UNIBANK haciendo uso de las variables más importantes el siguiente:

$$p_i = F(x_i'\beta)$$

$$F(t) = 1 - (e)^{-e^t}$$

donde t = $x_i'\beta$

Por lo tanto, usando los coeficientes Mean B del cuadro Nº 16, el modelo se representó de la siguiente manera:

$$p_i = F(x_i'\beta) \ = F(t) = 1 - (e)^{-e^t}$$

Donde "t" queda expresado de la siguiente manera:

$$t = 1.3 + 0.1X_2 - 0.3X_3 - 0.4X_4 - 0.4X_5 - 0.1X_6 + 0.2X_7 - 0.1X_8 - 0.2X_9 - 0.3X_{10} + 0.1X_{11} - 0.2X_{12}$$

### 6.3.2 Modelos de regresión logístico bayesiano con enlace asimétrico power logit

En el cuadro Nº 17 se muestra el resultado del modelo de regresión logístico bayesiano con enlace asimétrico power logit. La columna Mean B representa los coeficientes del modelo, los cuales tienen una relación no lineal con la probabilidad de que un cliente presente mora 60. La columna 2.50% es el límite inferior del intervalo de confianza al 95% de confianza para la columna Mean B y la columna 97.50% es el límite superior del intervalo de confianza al 95% de confianza para la columna Mean B.

**Cuadro N° 17**: Modelo de regresión logístico bayesiano con enlace asimétrico power logit

| **Variables** | **Mean B** | **2.50%** | **97.50%** |
|---|---|---|---|
| Constante | 2.5 | 2.3 | 2.8 |
| Situación de la casa (X1) | -0.1 | -0.3 | 0.1 |
| Edad (X2) | 0.2 | -0.1 | 0.4 |
| Máxima antigüedad con TC (X3) | -0.7 | -0.9 | -0.4 |
| Tenencia de producto pasivo (X4) | -0.7 | -1 | -0.5 |
| Ingreso mensual (X5) | -0.8 | -1 | -0.6 |
| Línea de TC en el SF (X6) | -0.3 | -0.6 | -0.1 |
| Saldo deudor en el SF (X7) | 0.4 | 0.1 | 0.6 |
| Score de aprobación (X8) | -0.3 | -0.7 | 0.1 |
| Abono de pago de haberes (X9) | -0.4 | -0.7 | -0.2 |
| Máxima clasificación SBS (X10) | -0.7 | -0.9 | -0.6 |
| Apalancamiento (X11) | 0.3 | 0.1 | 0.5 |
| Número de veces sueldo (X12) | -0.4 | -0.7 | -0.1 |
| lambda | 0.7 | 0.4 | 1 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

De este modo, el modelo de regresión logística bayesiano con enlace asimétrico power logit se presentará con 12 variables, siendo la ecuación para determinar las probabilidades de mora de los clientes que incumplirán sus pagos de sus tarjetas de crédito en la entidad bancaria UNIBANK haciendo uso de las variables más importantes el siguiente:

$$p_i = F(x_i'\beta)$$

$$F(t) = (1 + e^t)^{-\lambda} \;\; \lambda > 0$$

donde t = $x_i'\beta$

Por lo tanto, usando los coeficientes Mean B del cuadro Nº 17, el modelo se representó de la siguiente manera:

$$p_i = F(x_i'\beta) = F(t) = (1 + e^t)^{-\lambda}$$

Donde "t" y "λ" queda expresado de la siguiente manera:

$t = 2.5 - 0.1X_1+0.2X_2-0.7X_3-0.7X_4-0.8X_5-0.3X_6+0.4X_7-0.3X_8-0.4X_9-0.7X_{10}+0.3X_{11}-0.4X_{12}$

$$\lambda = 0.7$$

### 6.3.3 Modelos de regresión logístico bayesiano con enlace asimétrico scobit

En el cuadro Nº 18 se muestra el resultado del modelo de regresión logístico bayesiano con enlace asimétrico scobit. La columna Mean B representa los coeficientes del modelo, los cuales tienen una relación no lineal con la probabilidad de que un cliente presente mora 60. La columna 2.50% es el límite inferior del intervalo de confianza al 95% de confianza para la columna Mean B y la columna 97.50% es el límite superior del intervalo de confianza al 95% de confianza para la columna Mean B.

**Cuadro N° 18**: Modelo de regresión logístico bayesiano con enlace asimétrico scobit

| Variables | Mean B | 2.50% | 97.50% |
|---|---|---|---|
| Constante | 1.2 | 1 | 1.4 |
| Situación de la casa (X1) | -0.1 | -0.2 | 0.1 |
| Edad (X2) | 0.1 | 0 | 0.3 |
| Máxima antigüedad con TC (X3) | -0.5 | -0.6 | -0.3 |
| Tenencia de producto pasivo (X4) | -0.5 | -0.7 | -0.4 |
| Ingreso mensual (X5) | -0.6 | -0.7 | -0.4 |
| Línea de TC en el SF (X6) | -0.2 | -0.3 | 0 |
| Saldo deudor en el SF (X7) | 0.2 | 0.1 | 0.4 |
| Score de aprobación (X8) | -0.2 | -0.5 | 0 |
| Abono de pago de haberes (X9) | -0.2 | -0.4 | -0.1 |
| Máxima clasificación SBS (X10) | -0.5 | -0.6 | -0.3 |
| Apalancamiento (X11) | 0.2 | 0.1 | 0.3 |
| Número de veces sueldo (X12) | -0.3 | -0.5 | -0.1 |
| lambda | 2.2 | 1.7 | 2.7 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

De este modo, el modelo de regresión logística bayesiano con enlace asimétrico scobit se presentará con 12 variables, siendo la ecuación para determinar las probabilidades de mora de los clientes que incumplirán sus pagos de sus tarjetas de crédito en la entidad bancaria UNIBANK haciendo uso de las variables más importantes el siguiente:

$$p_i = F(x_i'\beta)$$
$$F(t) = 1 - (1 + e^t)^{-\lambda} \;\; \lambda > 0$$
$$\text{donde } t = x_i'\beta$$

Por lo tanto, usando los coeficientes Mean B del cuadro Nº 18, el modelo se representó de la siguiente manera:

$$p_i = F(x_i'\beta) = F(t) = 1 - (1 + e^t)^{-\lambda}$$

Donde "t" y "λ" queda expresado de la siguiente manera:

$t = 1.2 - 0.1X_1+0.1X_2-0.5X_3-0.5X_4-0.6X_5-0.2X_6+0.2X_7-0.2X_8-0.2X_9-0.5X_{10}+0.2X_{11}-0.3X_{12}$

$$\lambda = 2.2$$

### 6.4 Comparación de modelos

#### 6.4.1 Uso de indicadores bayesianos

En el cuadro Nº 19, se presentan los indicadores bayesianos para los modelos de regresión logística bayesiano con enlaces asimétricos cloglog, power logit y scobit.

**Cuadro N° 19:** Indicadores bayesianos

| Modelo | DIC | EAIC | EBIC |
|---|---|---|---|
| Cloglog | 3,875.0 | 3,887.7 | 3,906.2 |
| Power logit | 3,862.0 | 3,873.0 | 3,891.5 |
| Scobit | 3,864.0 | 3,875.2 | 3,893.7 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

Del cuadro Nº 19 se puede apreciar que el modelo power logit es el que presenta mejor ajuste, pues tiene los menores valores de los indicadores de DIC, EAIC y EBIC, siendo estos 3,862, 3,873 y 3,891.5, respectivamente. Sin embargo esta diferencia no es muy grande en comparación con los modelos cloglog y scobit.

### 6.4.2 Uso de la sensibilidad de la tabla de clasificación

Los cuadros Nº 20, Nº 21 y Nº 22 presentan las tablas de clasificación de los modelos de regresión logística bayesiana con enlaces asimétricos cloglog, power logit y scobit, respectivamente.

**Cuadro Nº 20:** Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico cloglog

| **Observado** | | **Pronosticado** | | |
|---|---|---|---|---|
| | | Mora60 | | Porcentaje |
| | | NO | SI | correcto |
| Mora60 | NO | 294 | 130 | 69.3% |
| | SI | 404 | 672 | 62.5% |
| Porcentaje global | | | | 64.4% |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro Nº 21:** Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico power logit

| **Observado** | | **Pronosticado** | | |
|---|---|---|---|---|
| | | Mora60 | | Porcentaje |
| | | NO | SI | correcto |
| Mora60 | NO | 329 | 95 | 77.6% |
| | SI | 495 | 581 | 54.0% |
| Porcentaje global | | | | 60.7% |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro Nº 22:** Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico scobit

| Observado | | Pronosticado | | |
|---|---|---|---|---|
| | | Mora60 | | Porcentaje |
| | | NO | SI | correcto |
| Mora60 | NO | 326 | 98 | 76.9% |
| | SI | 502 | 574 | 53.3% |
| Porcentaje global | | | | 60.0% |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

De los cuadros Nº 20, Nº 21 y Nº 22, la sensibilidad para el modelo de regresión logística bayesiana con enlace asimétrico cloglog es igual a 62.5%, para el modelo de regresión logística bayesiana con enlace asimétrico power logit es 54.0% y para el modelo de regresión logística bayesiana con enlace asimétrico scobit es 53.3%. Los valores de sensibilidad tanto para los modelos power logit y scobit son casi similares, sin embargo si se compara con el modelo cloglog, este los supera en 8.5% y 9.1%, respectivamente. Asimismo su variación porcentual es de 15.7% y 17.1%, respectivamente.

Por lo tanto el modelo más adecuado haciendo uso de la sensibilidad de la tabla de clasificación es el modelo de regresión logística bayesiano con enlace asimétrico cloglog.

### 6.4.3 Uso de la curva ROC

El cuadro Nº 23 presenta el área bajo las curva entre las probabilidades obtenidas mediante los modelos de regresión logística bayesiana con enlaces asimétricos cloglog, power logit y scobit, y la variable Mora60, la cual se encuentra en la que la columna área.

La columna significancia asintótica para los modelos de regresión logística bayesiana con enlaces asimétricos cloglog, power logit y scobit son menores a un nivel de significancia del 5%, por lo que se rechaza la hipótesis nula que el área bajo la curva ROC es igual a 0.5. En conclusión los modelos cloglog, power logit y scobit son adecuados.

La columna límite inferior y límite superior, representan el intervalo de confianza al 95% de confianza para la columna área.

**Cuadro Nº 23:** Área bajo la curva para los modelos de regresión logística bayesiana con enlaces asimétricos

| Modelo | Área | Sig. asintóticab | Límite inferior | Límite superior |
|---|---|---|---|---|
| Cloglog | 0.727 | 0.000 | 0.699 | 0.755 |
| Power logit | 0.726 | 0.000 | 0.698 | 0.754 |
| Scobit | 0.726 | 0.000 | 0.697 | 0.754 |

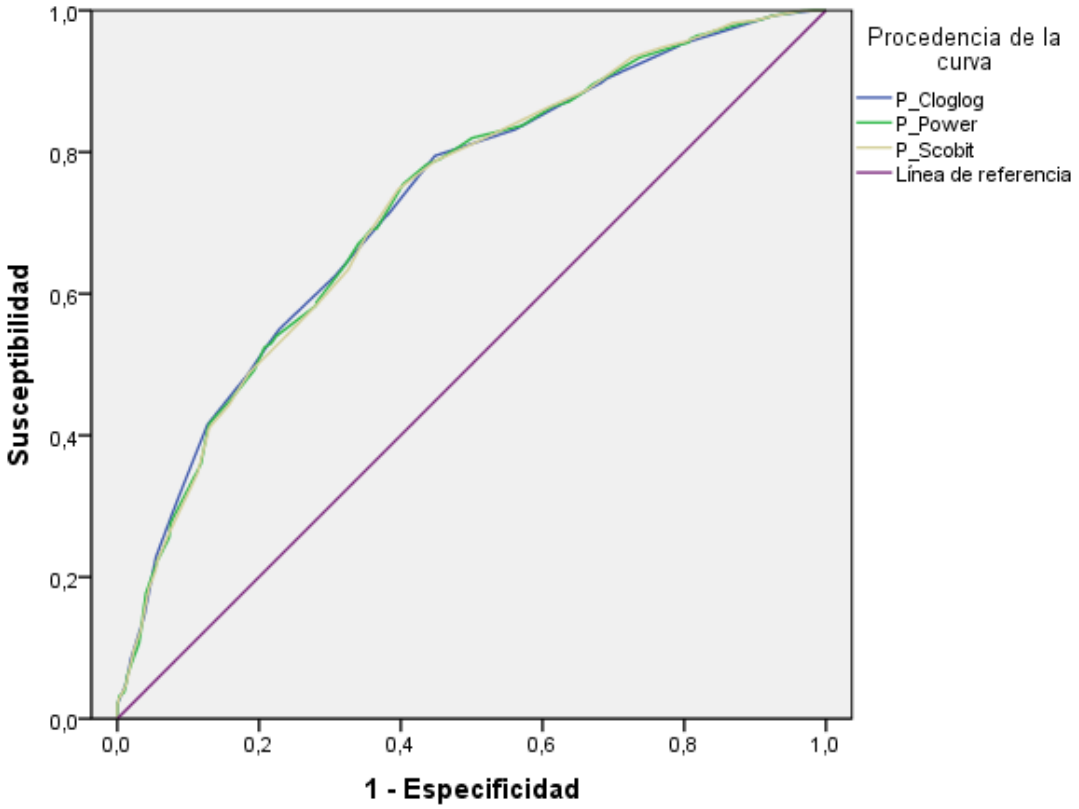
FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

Finalmente, se puede apreciar que el área bajo la curva ROC para los modelos cloglog, power logit y scobit son similares.

A continuación se presenta en la figura Nº 14 las curvas ROC para los modelos cloglog, power logit y scobit, en la que se aprecia que estas curvas se superponen unas a otras.

**Figura N° 14:** Curva ROC para los modelos de regresión logística bayesiana con enlaces asimétricos



# VII. CONCLUSIONES

En esta parte se expusieron las conclusiones producto del desarrollo de la presente investigación. Por lo tanto, como consecuencia de la investigación realizada se presentaron las siguientes conclusiones:

- El modelo de regresión binaria bayesiano con enlace asimétrico cloglog fue el más adecuado para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora pues presentó un valor mucho mayor de sensibilidad que los modelos power logit y scobit, siendo las diferencias 8.5% y 9.1%, respectivamente y variación porcentual de 15.7% y 17.1%, respectivamente.

- Los modelos de regresión binaria bayesiana con enlaces asimétricos cloglog, power logit y scobit presentaron un buen desempeño para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK. Los indicadores bayesianos DIC, EAIC y EBIC muy parecidos. Las sensibilidades que se obtuvieron fueron superiores al 50% y área bajo la curva ROC significativos con valores muy parecidos.

- Mediante el uso del algoritmo de boruta se encontró que las variables situación de la casa, edad, máxima antigüedad con tarjeta de crédito, tenencia de producto pasivo, ingreso mensual, línea de tarjeta de crédito en el sistema financiero, saldo deudor en el sistema financiero, score de aprobación, abono de pago de haberes, máxima clasificación SBS, apalancamiento y número de veces sueldo ayudan a la clasificación de clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK según su probabilidad de mora.

# VIII. RECOMENDACIONES

- Se sugiere continuar el estudio, planteando modelos alternativos como por ejemplo de machine learning y determinar si mejora la precisión para clasificar a los clientes que incumplen sus obligaciones crediticias con sus tarjetas de crédito en la entidad bancaria UNIBANK.

- Se sugiere usar los modelos de regresión binaria logística bayesiano con enlaces asimétricos a otros tipos de aplicaciones, como por ejemplo en el caso de fraudes de tarjeta de crédito en la que la tasa de fraudulentos por lo general es siempre baja frente a los que no comenten fraude.

# IX. REFERENCIAS BIBLIOGRÁFICAS


Agresti, A. 2002. Categorical Data Analysis. Second Edition. John Wiley.

Albert, J. H. 2009. Bayesian Computation with R. Springer Verlag.

Bazán, J. and Bayes, C. 2010. Inferencia Bayesiana en Modelos de Regresión Bayesiana Binaria usando BRMUV. Departamento de Ciencias de la Sección de Matemática, Pontificia Universidad Católica del Perú, Febrero 2010.

Bermúdez, LL., Pérez, J.M., Ayuso, M., Gómez, E. y Vásquez, F.J. 2008. A Bayesian dichotomous model con asymmetric link for fraud in insurance. Mathematics and Economics 42 2008, p. 779-786.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002). Journal of the Royal Statistical Society Series B, 64, 3,616-618.

Carlin, B.P. y Louis, T.A. (2001).Bayes and Empirical Bayes Methods for Data Analysis Essays on Item Response Theory. Second edition. New York: Chapman & Hall.

Chen, M.H., Dey, D.K., and Shao, Q-M. 1999. A new skewed link model for dichotomous quantal response data. Journal of the American Statistical Association, 94, p. 1172-1186.

Chen, M.H., Dey, D.K., and Shao, Q-M. 2001. Bayesian analysis of binary data using skewed logit models. Calcutta Statistical Association Bulletin, 51, p. 201-202.

Collet, D. 2003. Modelling binary data. Chapman & Hall/CRC, Second Edition, Boca Raton, USA.

Czado, C., and Santner, T. J. 1992. The efect of link misspecicaction on binary regression inference. Journal of Statistical Planning and Inference, 33, p. 213-231.

Dávila N., García M.D., Pérez J.M. and Gómez E. 2015. An Asymmetric Logit Model to explain the likelihood of succes in academic results. Revista de Investigación Educativa, 33 (1), 27-45.

Kursa M. and Rudnicki W. 2010. Feature Selection with Boruta Package. Journal of Statistical Software, Volumen 36, Issue 11.

Liaw A. and Wiener M. 2002. Classification and Regression by random Forest. R News, 2(3), p. 18-22.

López I. y Píta S. (2001). Curvas ROC. Unidad de Epidemiología Clínica y Bioestadística. España.

Nagler, J. 1994. Scobit: an alternative estimator to logit and probit. American Journal Political Science, 38, p. 230-255.

Novales, A. 1993. Econometría. Segunda edición. McGraw-Hill. España.

Pérez, J.M., Negrín, M.A., García, C. y Gómez, E. 2014. Bayesian Asymmetric Logit Model for Detecting Risk Factors in Motors Ratemaking. Astin Bulletin 44 (2), p. 445-457.

Pérez, S. 2015. Estimación de la curva ROC acumulativa / dinámica. Facultad de Ciencias, Universidad de Oviedo.

Prentice, R. L. 1976. A Generalization of the probit and logit methods for dose-response curves. Biometrika, 32, p. 761-768.

Ross, S. 1995. Stochastic Processes, Wiley: New York, NY.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y Van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B, 64, 583-639.

# X. ANEXOS

## Anexo 1: Análisis de frecuencias de la variable dependiente y de las variables independientes

**Cuadro N° 2:** Tabla de frecuencias de la variable mora 60

| MORA60 | Frecuencia | Porcentaje |
|---|---|---|
| No Mora60 | 1,500 | 30.0 |
| Si Mora60 | 3,500 | 70.0 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 3:** Tabla de frecuencias de la variable situación de la casa

| VAR01 | Frecuencia | Porcentaje |
|---|---|---|
| NO PROPIA | 3,571 | 71.4 |
| PROPIA | 1,429 | 28.6 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 4:** Tabla de frecuencias de la variable edad del cliente

| VAR02 | Frecuencia | Porcentaje |
|---|---|---|
| Joven | 1,117 | 22.3 |
| Adulto | 3,883 | 77.7 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 5:** Tabla de frecuencias de la variable máxima antigüedad con tarjeta de crédito en el sistema financiero

| VAR03 | Frecuencia | Porcentaje |
|---|---|---|
| Menor a 12 meses | 2,459 | 49.2 |
| Mayor o igual a 12 meses | 2,541 | 50.8 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 6:** Tabla de frecuencias de la variable número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito

| VAR04 | Frecuencia | Porcentaje |
|---|---|---|
| Menor a 12 meses | 1,072 | 21.4 |
| Mayor o igual a 12 meses | 3,928 | 78.6 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 7:** Tabla de frecuencias de la variable ingreso mensual del cliente

| VAR05 | Frecuencia | Porcentaje |
|---|---|---|
| Ingreso bajo | 2,030 | 40.6 |
| Ingreso alto | 2,970 | 59.4 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 8:** Tabla de frecuencias de la variable monto de línea de tarjeta de crédito en el sistema financiero

| VAR06 | Frecuencia | Porcentaje |
|---|---|---|
| Línea baja | 3,510 | 70.2 |
| Línea alta | 1,490 | 29.8 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 9:** Tabla de frecuencias de la variable monto de saldo deudor promedio total en el sistema financiero

| VAR07 | Frecuencia | Porcentaje |
|---|---|---|
| Saldo bajo | 1,275 | 25.5 |
| Saldo alto | 3,725 | 74.5 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 10:** Tabla de frecuencias de la variable score con el que fue aprobada la tarjeta de crédito en el banco

| VAR08 | Frecuencia | Porcentaje |
|---|---|---|
| Score bajo | 439 | 8.8 |
| Score alto | 4,561 | 91.2 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 11:** Tabla de frecuencias de la variable si tuvo abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito

| VAR09 | Frecuencia | Porcentaje |
|---|---|---|
| No tiene abono PDH | 1,366 | 27.3 |
| Si tiene abono PDH | 3,634 | 72.7 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 12:** Tabla de frecuencias de la variable máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito

| VAR10 | Frecuencia | Porcentaje |
|---|---|---|
| CPP, Deficiente, Dudoso y Pérdida | 1,858 | 37.2 |
| Normal | 3,142 | 62.8 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 13:** Tabla de frecuencias de la variable apalancamiento

| VAR11 | Frecuencia | Porcentaje |
|---|---|---|
| Apalancamiento < 1.19 | 2,800 | 56.0 |
| Apalancamiento >= 1.19 | 2,200 | 44.0 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Cuadro N° 14:** Tabla de frecuencias de la variable número de veces sueldo

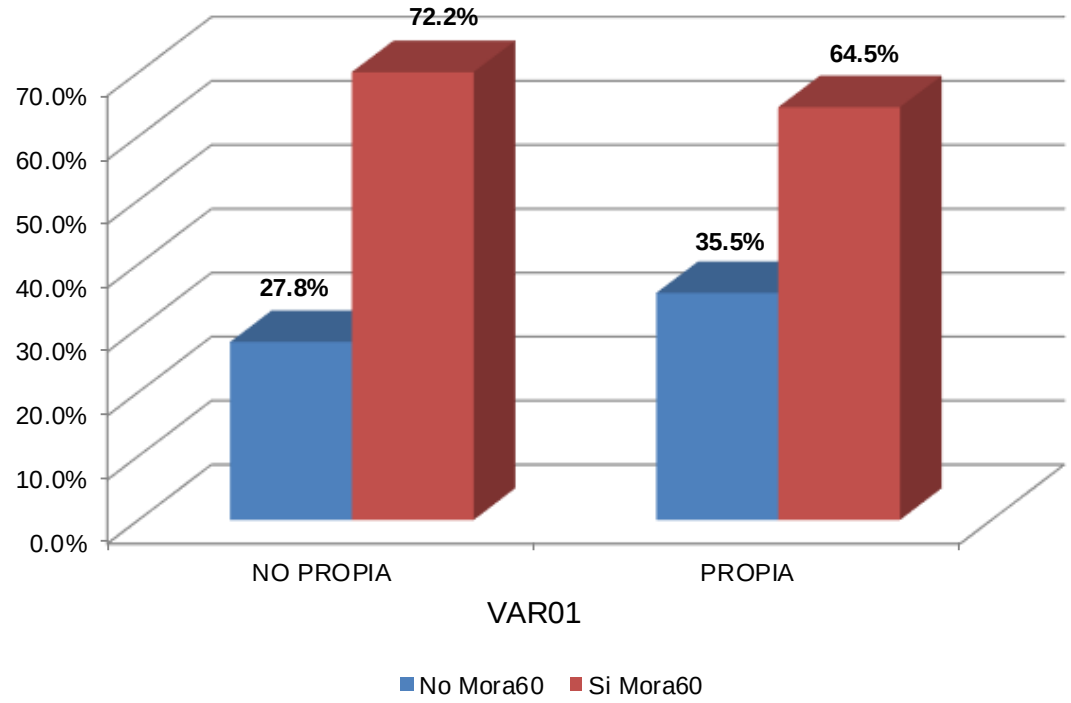
| VAR12 | Frecuencia | Porcentaje |
|---|---|---|
| Veces_Sueldo < 3.35 | 3,845 | 76.9 |
| Veces_Sueldo >= 3.35 | 1,155 | 23.1 |
| Total | 5,000 | 100.0 |

FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

## Anexo 2: Gráficos descriptivos bivariados entre la variable dependiente y las variables independientes

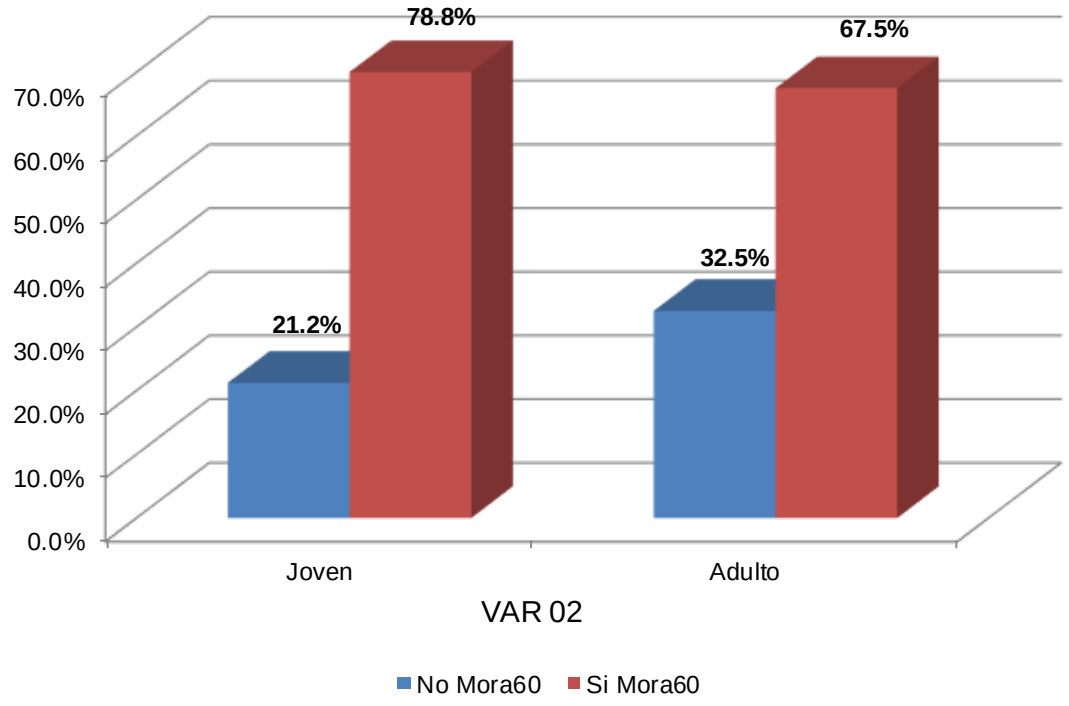
**Figura N° 1:** Gráfico porcentual entre la variable mora 60 y la variable situación de la casa




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

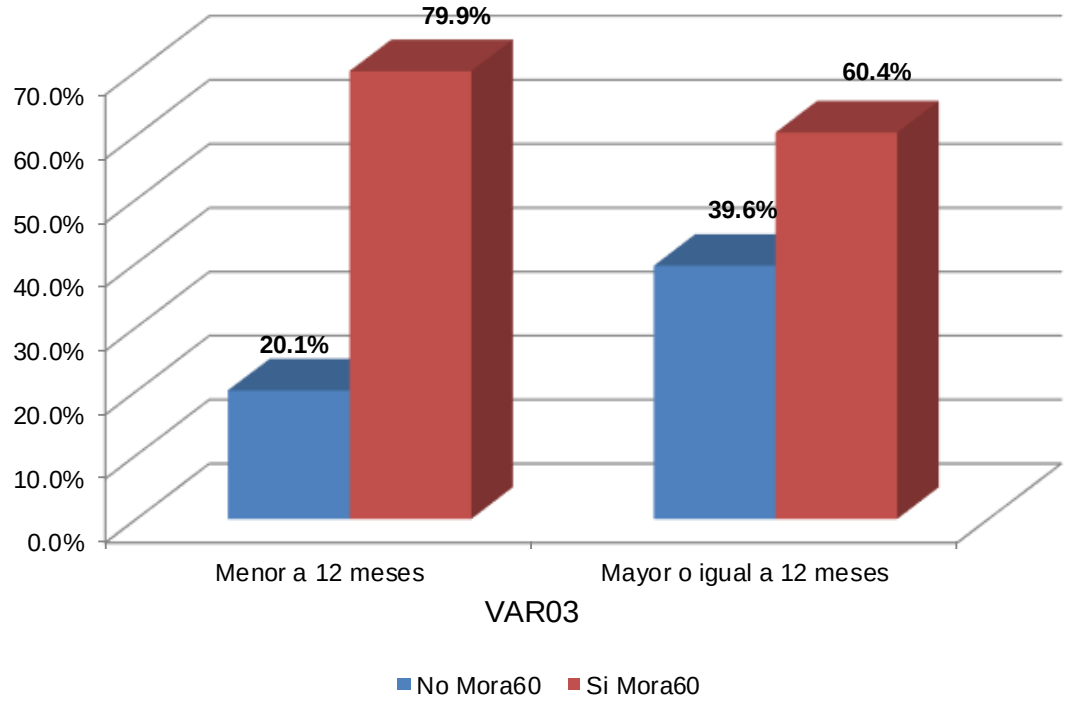
**Figura N° 2:** Gráfico porcentual entre la variable mora 60 y la variable edad del cliente




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

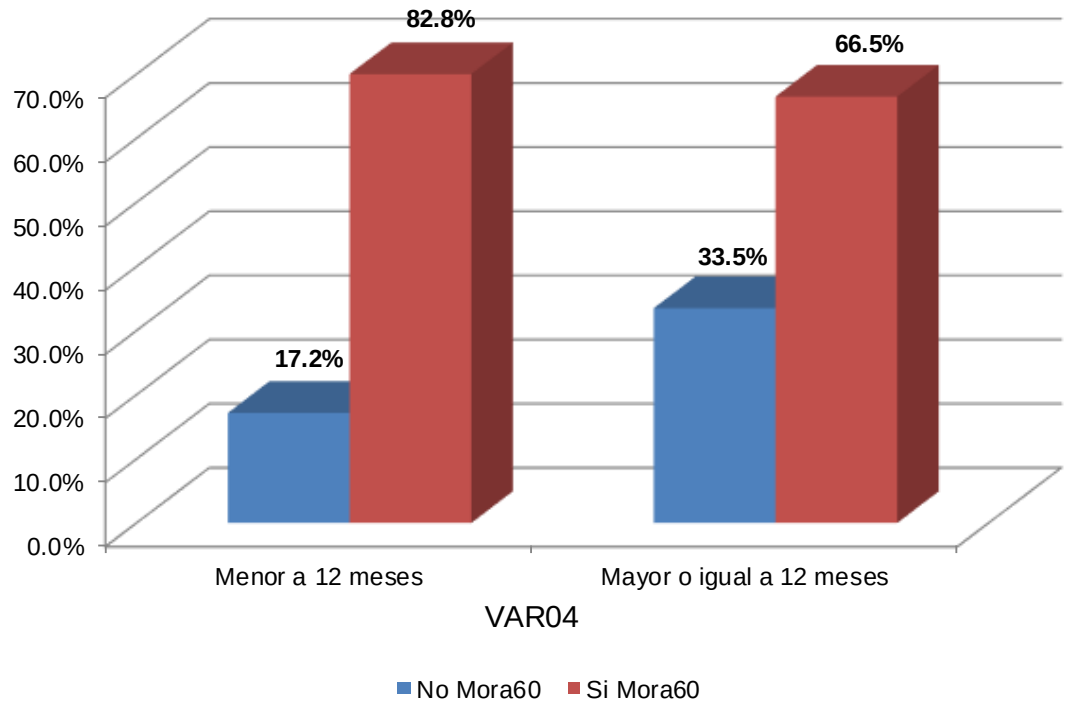
**Figura N° 3:** Gráfico porcentual entre la variable mora 60 y la variable máxima antigüedad con tarjeta de crédito en el sistema financiero




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

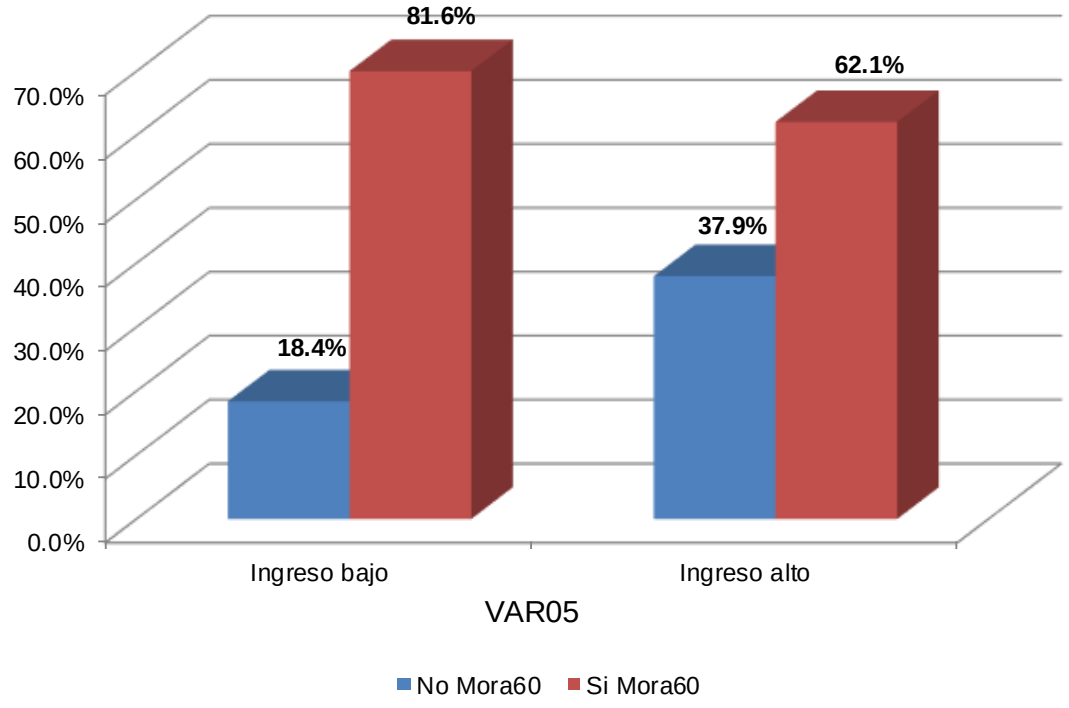
**Figura N° 4:** Gráfico porcentual entre la variable mora 60 y la variable número de meses con algún producto pasivo durante los 12 meses antes de la aprobación del crédito




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

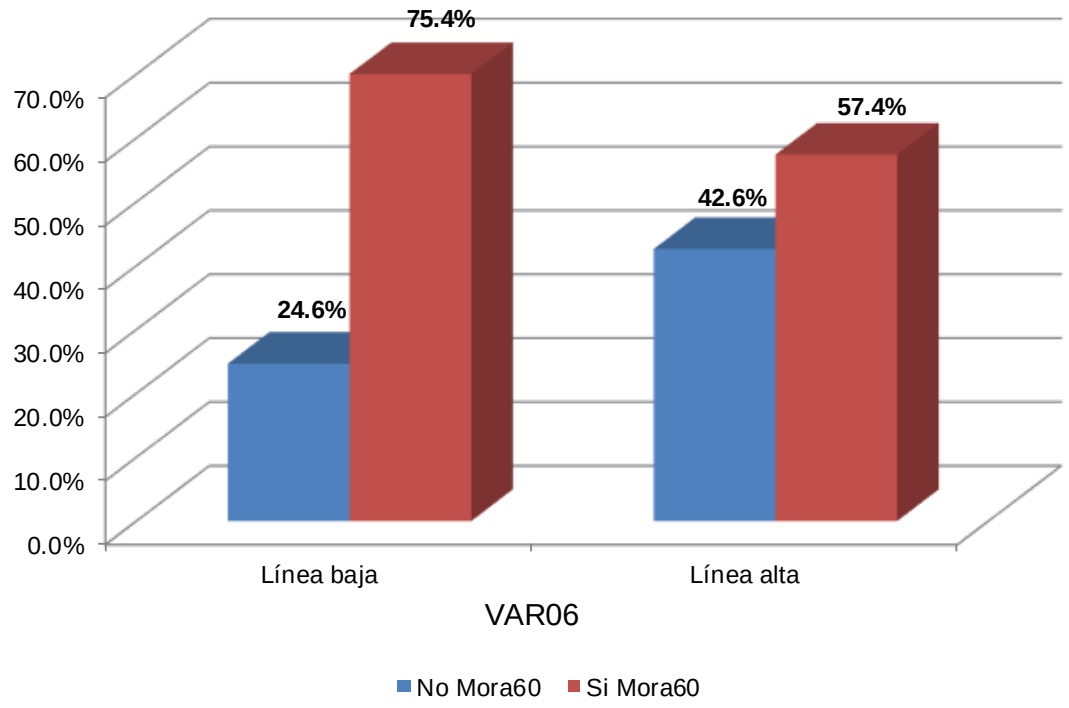
**Figura N° 5:** Gráfico porcentual entre la variable mora 60 y la variable ingreso mensual del cliente




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

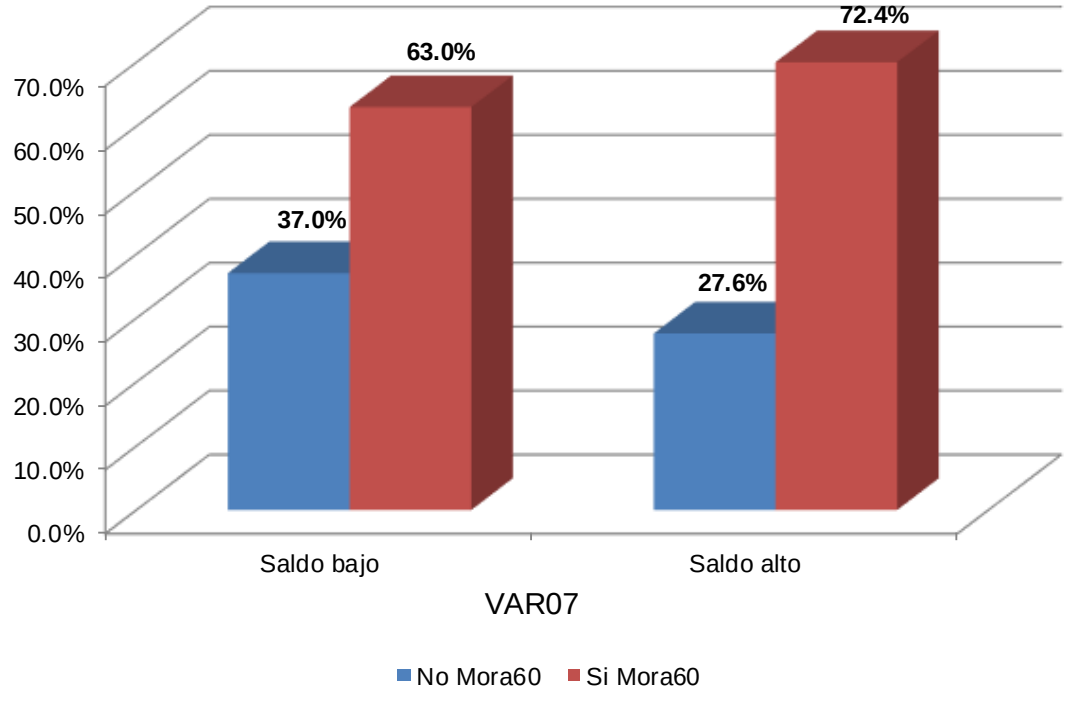
**Figura N° 6:** Gráfico porcentual entre la variable mora 60 y la variable monto de línea de tarjeta de crédito en el sistema financiero




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

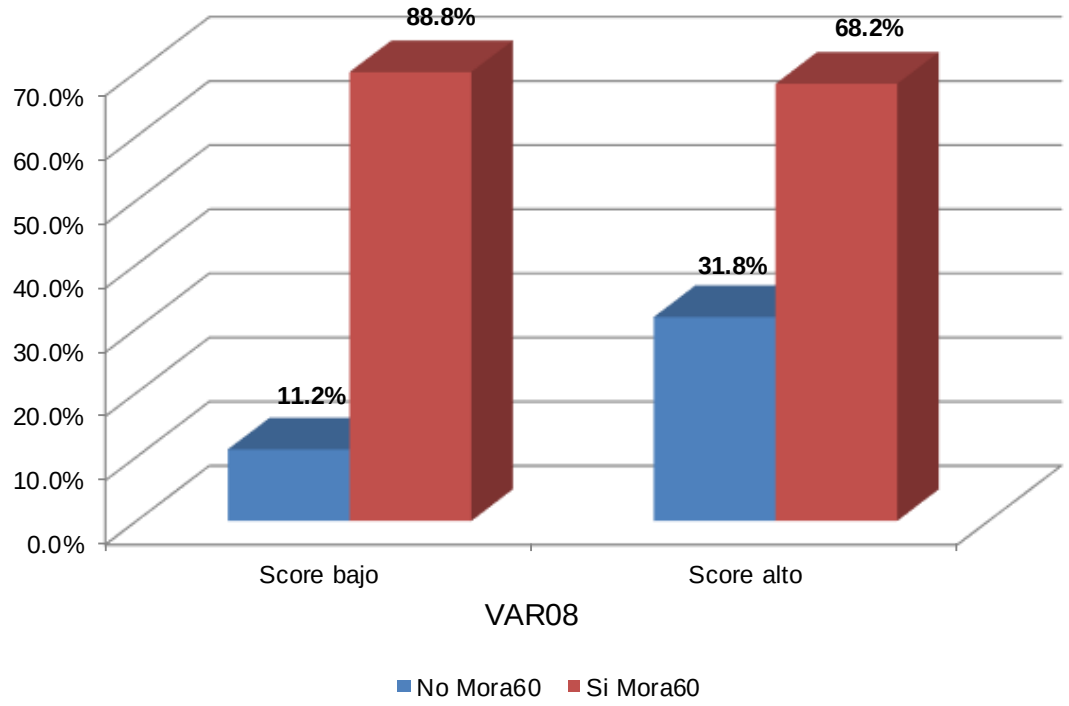
**Figura N° 7:** Gráfico porcentual entre la variable mora 60 y la variable monto de saldo deudor promedio total en el sistema financiero




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

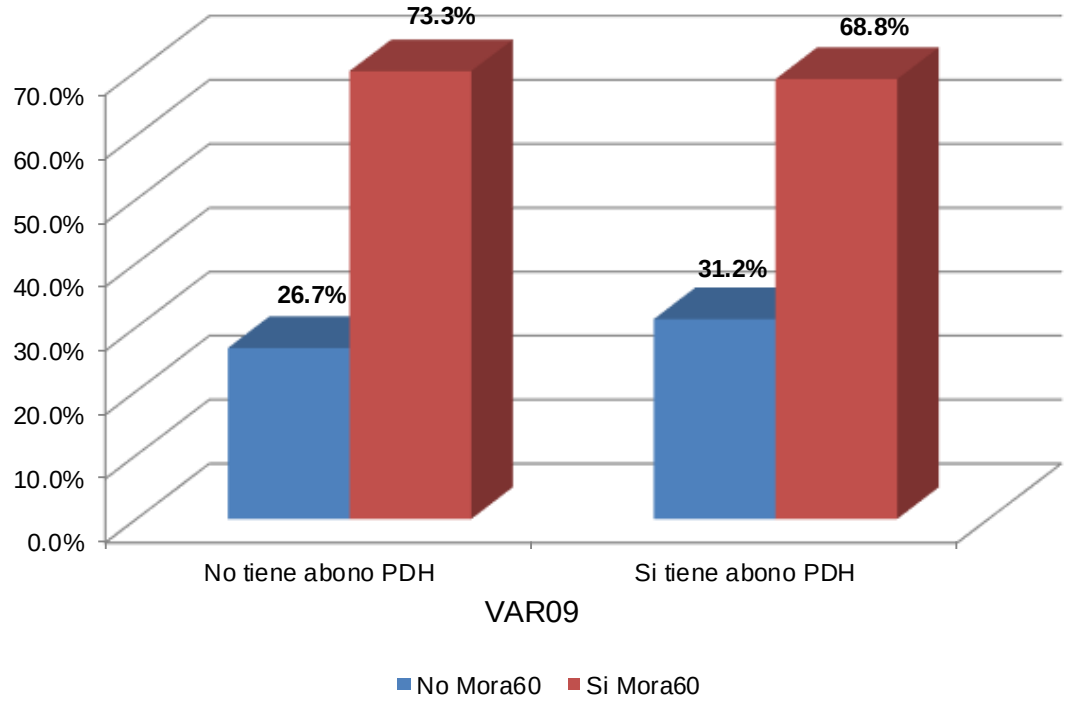
**Figura N° 8:** Gráfico porcentual entre la variable mora 60 y la variable score con el que fue aprobada la tarjeta de crédito en el banco




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

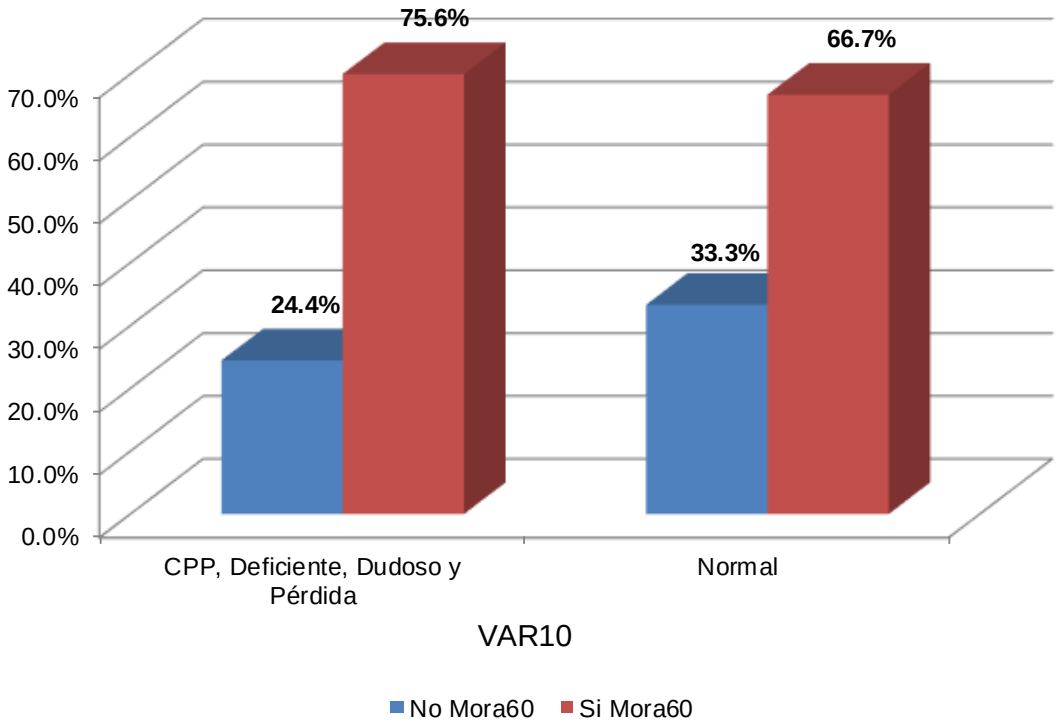
**Figura N° 9:** Gráfico porcentual entre la variable mora 60 y la variable si tuvo abono de pago de haberes en el banco durante los 12 meses antes de la aprobación del crédito




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

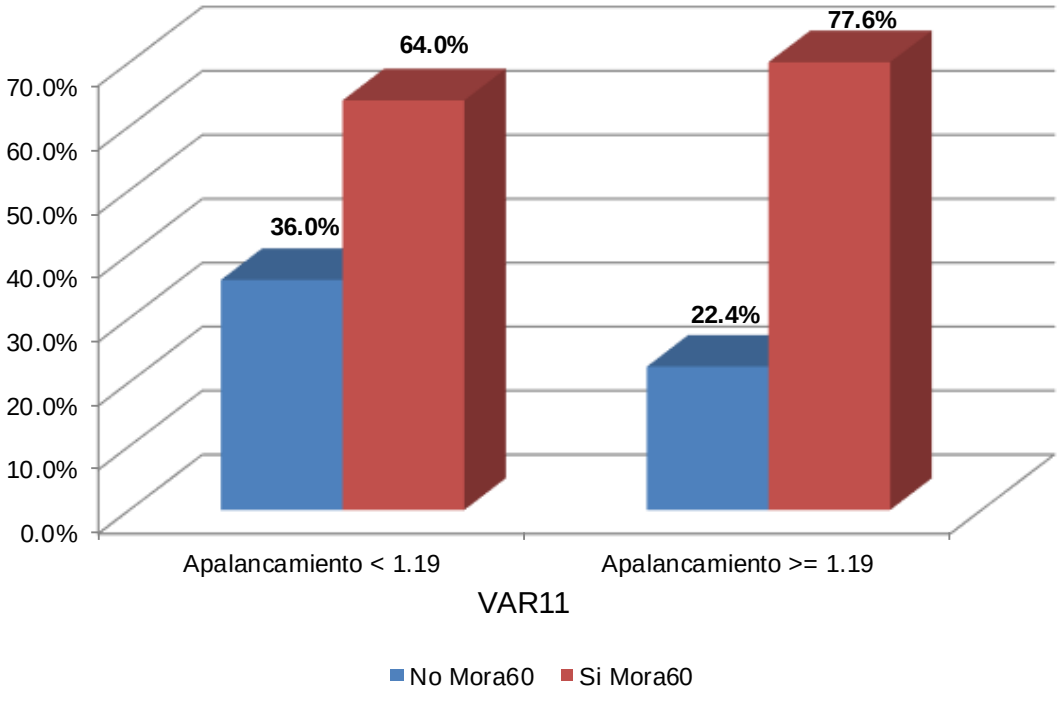
**Figura N° 10:** Gráfico porcentual entre la variable mora 60 y la variable máxima clasificación de riesgos SBS durante los 12 meses antes de la aprobación del crédito




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

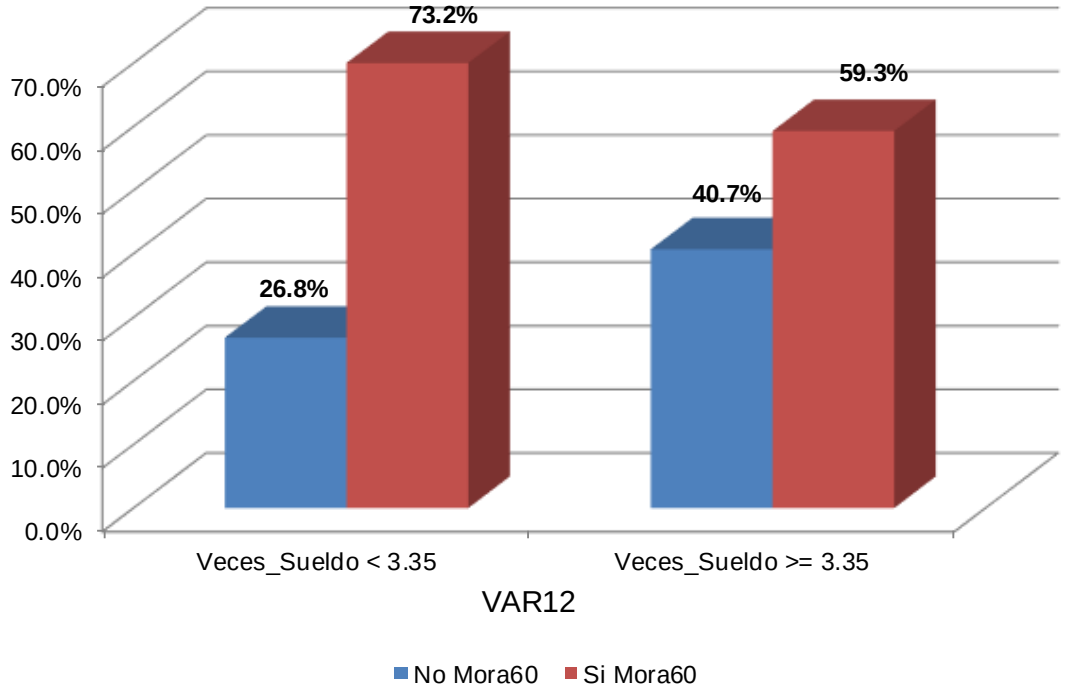
**Figura N° 11:** Gráfico porcentual entre la variable mora 60 y la variable apalancamiento




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

**Figura N° 12:** Gráfico porcentual entre la variable mora 60 y la variable número de veces sueldo




FUENTE: Entidad Bancaria UNIBANK AL 31 DE DICIEMBRE DE 2015

ELABORACIÓN: Propia

## Anexo 3: Códigos en R de los modelos cloglog, power logit y scobit

```
#Instalando librería
install.packages("R2WinBUGS")
install.packages("BRugs")
install.packages("R2OpenBUGS")
install.packages("rbugs")

library(R2WinBUGS)
library(BRugs)
library(R2OpenBUGS)
library(rbugs)

##################
#Modelos Cloglog#
##################

#Generando valores iniciales
inits<-function(){list(beta=c(0,0,0,0,0,0,0,0,0,0,0,0,0))}

#Implementación de la estimación bayesiana y las simulaciones
salida1<-bugs(data,inits,parameters.to.save=c("beta"),
        model.file="G:/Maestría UNALM/Tesis/Base/Modelo - 3 - Cloglog.txt",
n.chains=1, n.iter=2000,
        n.burnin=1000)

######################
#Modelos Power Logit#
######################

#Generando valores iniciales
inits<-function(){list(beta=c(0,0,0,0,0,0,0,0,0,0,0,0,0),lambda=0.5)}

#Implementación de la estimación bayesiana y las simulaciones
```

```
salida2<-bugs(data,inits,parameters.to.save=c("beta","lambda"),
        model.file="G:/Maestría UNALM/Tesis/Base/Modelo - 4 - Power Logit.txt",
n.chains=1, n.iter=2000,
        n.burnin=1000)

#################
#Modelos Scobit#
#################

#Generando valores iniciales
inits<-function(){list(beta=c(0,0,0,0,0,0,0,0,0,0,0,0,0),lambda=0.5)}

#Implementación de la estimación bayesiana y las simulaciones
salida3<-bugs(data,inits,parameters.to.save=c("beta","lambda"),
        model.file="G:/Maestría UNALM/Tesis/Base/Modelo - 5 - Scobit.txt", n.chains=1,
n.iter=2000,
        n.burnin=1000)
```