Gestión De Riesgos De Modelo:

Un Enfoque Desde La Regresión Bayesiana Con Enlaces Asimétricos Para La Predicción De La Morosidad Crediticia

Ciencia Latina Internacional
CID editorial
ALIANZA RED
Crossref

Mayo 2025 – CID - Centro de Investigación y Desarrollo
Copyright © CID - Centro de Investigación y Desarrollo
Copyright del texto © 2025 de Autores
libros.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 26903834

Datos Técnicos de Publicación Internacional
Título: "Gestión de riesgos de modelo: Un enfoque desde la regresión bayesiana con enlaces asimétricos para la predicción de la morosidad crediticia"
Autores: Richard Fernando Fernández Vásquez, Solange Lesly Basualdo Najera
Editor: CID - Centro de Investigación y Desarrollo
Diseño de tapa: CID - Centro de Investigación y Desarrollo
Corrección de Estilo: CID - Centro de Investigación y Desarrollo
Formato: PDF
Páginas: 37 pág.
Tamaño: Sobre C5 162 x 229 mm
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
ISBN: 978-1-967344-81-9
DOI: https://doi.org/10.37811/cli_w1219

1ª. Edición. Año 2025. Editorial CID - Centro de Investigación y Desarrollo.

El contenido del libro y sus datos en su forma, corrección y fiabilidad son responsabilidad exclusiva de los autores. Permite la descarga de la obra y compartir siempre que los créditos se atribuyan a los autores, pero sin la posibilidad de cambiarlo de cualquier forma o utilizarlo con fines comerciales

Prohibida su reproducción por cualquier medio.

Distribución gratuita

Dedicamos este trabajo a todos

los apasionados de la

Ciencia de Datos

ÍNDICE

CAPITULO I **8**
Conceptos clave 8
1.1 Tarjeta de crédito 8
1.2 Modelo 8
1.3 Riesgo de crédito 8
1.4 Reglamento para la Gestión de Riesgos de Modelo 8
1.5 Reglamento para la Evaluación y Clasificación del Deudor 9
CAPITULO II **11**
Pruebas y métodos estadísticos 11
2.1 Prueba Chi-Cuadrado de Pearson 11
2.2 Método de selección de variables mediante al algoritmo de Boruta 11
2.3 Método de Monte Carlo basados en Cadenas de Markov 12
CAPITULO III **14**
Modelo de regresión logística bayesiana **14**
3.1 Regresión Logística 14
3.2 Regresión Logística con variable respuesta dicotómica 14
3.3 Regresión Logística Bayesiana con Enlaces Simétricos 14
3.4 Regresión Logística Bayesiana con Enlaces Asimétricos 15
CAPITULO IV **18**
Medidas de bondad de ajuste 18
4.1 Indicadores estadísticos bayesianos 18
4.2 Tabla de clasificación 19
4.3 Curva ROC 20
CAPITULO V **21**
Caso de aplicación: clasificación de clientes con propensión a caer en mora en una entidad bancaria 21
5.1 Población 21
5.2 Muestra 21
5.3 Unidad de análisis 21
5.4 Fuentes de información 21
5.5 Definición de variables 21
5.6 Muestra de entrenamiento y validación 23
5.7 Análisis de significancia entre la variable dependiente e independientes 23
5.8 Selección de variables independientes 25
5.9 Modelos de regresión logística bayesiana con enlaces asimétricos 25
5.10 Comparación de modelos de regresión logística bayesiana con enlaces asimétricos 29
CAPITULO VI **32**
Conclusiones 32
REFERENCIAS BIBLIOGRAFICAS **33**

ÍNDICE DE TABLAS

Tabla 1. Tabla de Clasificación 19
Tabla 2. Prueba Chi-Cuadrado de Pearson entre la variable dependiente y las variables independientes 24
Tabla 3. Modelo de regresión logística bayesiana con enlace asimétrico cloglog 26
Tabla 4. Modelo de regresión logística bayesiana con enlace asimétrico power logit 27
Tabla 5. Modelo de regresión logística bayesiana con enlace asimétrico scobit 28
Tabla 6. Indicadores bayesianos 29
Tabla 7. Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico cloglog 29
Tabla 8. Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico power logit 30
Tabla 9. Tabla de Clasificación para el modelo de regresión logística bayesiana con enlace asimétrico scobit 30
Tabla 10. Área bajo la curva para los modelos de regresión logística bayesiana con enlaces asimétricos 31

ÍNDICE DE FIGURA

Figura. 1 Importancia de las variables independientes 25
Figura. 2 Curva ROC para los modelos de regresión logística bayesiana con enlaces asimétricos 31

GESTIÓN DE RIESGOS DE MODELO: UN ENFOQUE DESDE LA REGRESIÓN BAYESIANA CON ENLACES ASIMÉTRICOS PARA LA PREDICCIÓN DE LA MOROSIDAD CREDITICIA

- **Título:** Gestión de Riesgos de Modelo: Un enfoque desde la Regresión Bayesiana con Enlaces Asimétricos para la predicción de la Morosidad Crediticia
- **Autor:** Richard Fernando Fernández Vásquez, Solange Lesly Basualdo Najera
- **Ciudad:** Lima - Perú
- **Fecha:** Abril de 2025

Dr. Richard Fernando Fernández Vásquez[1]

Mag. Solange Lesly Basualdo Najera[2]

[1] Universidad Nacional de Ingeniería, Docente Universitario, correo electrónico: rfernandezv@uni.edu.pe. ORCID: https://orcid.org/0000-0003-1721-8527

[2] Universidad Nacional de Ingeniería, Docente Universitario, correo electrónico: lbasualdo@uni.edu.pe. ORCID: https://orcid.org/0009-0001-0567-689X

PRÓLOGO

En el dinámico entorno financiero, uno de los mayores desafíos a los que se enfrentan las entidades bancarias es la adecuada gestión de riesgos de modelos con modelos estadísticos para la predicción de la morosidad crediticia. En particular, la morosidad de los clientes se ha convertido en un fenómeno recurrente que impacta directamente en la rentabilidad y estabilidad de las instituciones. Identificar de manera oportuna a los clientes con mayor propensión a incumplir sus obligaciones de pago no solo permite una mejor asignación de recursos, sino también la implementación de estrategias preventivas más efectivas, tomar decisiones informadas y cumplir con la regulación de la gestión de riesgos de modelo.

Este libro se presenta como una contribución rigurosa y aplicada al campo de la gestión del riesgo, enfocándose en el uso de modelos estadísticos avanzados para la clasificación de clientes morosos. Específicamente, se analiza la efectividad de los modelos de regresión binaria bayesiana con enlaces asimétricos —cloglog, power logit y scobit— en la predicción de la morosidad crediticia en una Entidad Bancaria. A través de un enfoque comparativo y basado en evidencia empírica, se demuestra que el modelo cloglog supera significativamente a los demás en términos de sensibilidad, consolidándose como la opción más adecuada para este tipo de clasificación.

El lector encontrará en estas páginas no solo una sólida base teórica, sino también una aplicación práctica respaldada por datos reales, que permite comprender el valor que ofrecen estos modelos en la toma de decisiones estratégicas. Este trabajo aspira a ser útil tanto para profesionales del sector financiero, como para investigadores y estudiantes interesados en la estadística aplicada, la econometría y la ciencia de datos.

Que este libro sirva como una invitación a explorar nuevas herramientas analíticas en un entorno cada vez más cambiante.

CAPITULO I

Conceptos clave

1.1 Tarjeta de crédito

Mediante el contrato de tarjeta de crédito una entidad bancaria concede una línea de crédito al titular por un plazo determinado y otorga la correspondiente tarjeta, con la finalidad de que el usuario de dicha tarjeta adquiera bienes o servicios en los establecimientos afiliados que los proveen, pague obligaciones o, de así permitirlo la empresa emisora y no mediar renuncia expresa por parte del titular, hacer uso del servicio de disposición de efectivo u otros servicios conexos, dentro de los límites y condiciones pactados, obligándose a su vez, a pagar a la empresa que expide la correspondiente tarjeta, el importe de los bienes y servicios que haya adquirido, obligaciones pagadas, y demás cargos, conforme a lo establecido en el respectivo contrato. (Monografias.com, s.f.)

1.2 Modelo

Un modelo es un método cuantitativo o sistema que aplica teorías matemáticas, técnicas estadísticas, económicas y/o financieras, y supuestos, para procesar información cuantitativa o cualitativa en estimaciones cuantitativas. (Superintendencia de Banca, Seguros y AFP, 2023)

1.3 Riesgo de crédito

El riesgo de crédito se define 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. (Superintendencia de Banca, Seguros y AFP, 2011)

1.4 Reglamento para la Gestión de Riesgos de Modelo

En la Resolución S.B.S. N° 0053 – 2023, se define el riesgo de modelo como la posibilidad de pérdidas o consecuencias adversas derivadas de debilidades en el desarrollo, validación, implementación, uso y seguimiento de los modelos. El riesgo de modelo puede presentarse a partir de especificaciones o metodologías inadecuadas; de estimaciones erradas; de supuestos incorrectos; de errores de cálculo; de datos inexactos, inapropiados o incompletos; del uso inapropiado, impropio o no previsto del modelo; de la falta de entendimiento de las limitaciones del modelo; y del seguimiento y/o controles inadecuados, principalmente. Los

modelos a los que se aplica el presente Reglamento son aquellos utilizados en la gestión de los riesgos de crédito. (Superintendencia de Banca, Seguros y AFP, 2023)

1.5 Reglamento para la Evaluación y Clasificación del Deudor

En la Resolución S.B.S. N° 11356 – 2008, se presentan definiciones respectos a los créditos, tipos de créditos y clasificación del deudor, que a continuación se proceden a detallar.

i) Definiciones

a) Créditos: se refiere a la suma de los créditos directos más indirectos.

b) Créditos directos: representa los financiamientos que, bajo cualquier modalidad, las empresas del sistema financiero otorguen a sus clientes, originando a cargo de éstos la obligación de entregar una suma de dinero determinada, en uno o varios actos.

c) Créditos indirectos: representan los avales, las cartas fianza, las aceptaciones bancarias, las cartas de crédito, los créditos aprobados no desembolsados y las líneas de crédito no utilizadas, otorgados por las empresas del sistema financiero.

d) Créditos revolventes: son aquellos créditos en los que se permite que el saldo fluctúe en función de las decisiones del deudor, por ejemplo, las tarjetas de crédito. Asimismo, se consideran dentro de este tipo de crédito los productos que permiten reutilizaciones parciales, es decir, que tienen un componente revolvente y otro no revolvente.

e) Créditos no revolventes: son aquellos créditos reembolsables por cuotas, siempre que los montos pagados no puedan ser reutilizables por el deudor. En este tipo de crédito no se permite que los saldos pendientes fluctúen en función de las propias decisiones del deudor.

f) Deudor minorista: persona natural o jurídica que cuenta con créditos directos o indirectos clasificados como de consumo (revolventes y no revolventes), a microempresas, a pequeñas empresas o hipotecarios para vivienda.

(Superintendencia de Banca, Seguros y AFP, 2008)

ii) Tipos de crédito

- Créditos de consumo revolvente: son aquellos créditos revolventes otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

- Créditos de consumo no revolvente: son aquellos créditos no revolventes otorgados a personas naturales, con la finalidad de atender el pago de bienes, servicios o gastos no relacionados con la actividad empresarial.

(Superintendencia de Banca, Seguros y AFP, 2008)

iii) Clasificación del deudor

La clasificación del deudor está determinada principalmente por su capacidad de pago, definida por el flujo de fondos y el grado del cumplimiento de sus obligaciones. Si un deudor es responsable de varios tipos de créditos con una misma entidad, la calificación de basará en la categoría del crédito con mayor riesgo. Las categorías con las que se reportan a los clientes son: Normal (0), Con Problemas Potenciales – CPP (1), Deficiente (2), Dudoso (3) y Pérdida (4). (Superintendencia de Banca, Seguros y AFP, 2008)

CAPITULO II

Pruebas y métodos estadísticos

2.1 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. (Basualdo, 2014)

H_o : Las dos variables son independientes

H_a : Las dos variables no son independientes

La fórmula del estadístico es la siguiente:

$$\chi^2 = \sum_i \frac{(observada_i - teorica_i)^2}{teorica_i}$$

Cuanto mayor sea el valor de χ^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 (gl) vienen dados por:

gl= (r-1)(k-1). Donde r es el número de filas y k el de columnas.

No se rechaza H_0 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.

2.2 Método de selección de variables mediante al algoritmo de Boruta

La selección de variables es un aspecto importante en la construcción de modelos estadísticos 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 modelo más 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 un método de selección de variables mediante algoritmo de Boruta. (Fernández, 2018)

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.

2.3 Método de Monte Carlo basados en Cadenas de Markov

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:

$$I = E[g(\theta)] = \int g(\theta) f(\theta|\mathrm{y}) d\theta \qquad (1)$$

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

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 1 en forma analítica. Por este motivo se hace necesario utilizar métodos aproximados para obtener estas integrales, como por ejemplo los métodos conocidos como métodos de Monte Carlo basados en cadenas de Markov (MCMC). Este método permite generar de manera iterativa una cadena de Markov para θ de tal manera que $f(\theta|\mathrm{y})$ sea una distribución ergódica estacionaria. Empezando en algún estado inicial θo la idea es simular un número suficientemente grande M de transiciones bajo la cadena de Markov y registrar los correspondientes estados simulados θj. En Ross (1995) se muestra que la media ergódica está expresada en la ecuación 2:

$$\hat{I} = \frac{1}{M}\sum_{j=1}^{M} g(\theta_j) \qquad (2)$$

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.

CAPITULO III

Modelo de regresión logística bayesiana

3.1 Regresión Logística

La regresión logística es un modelo de elección discreta en el que la variable dependiente es cualitativa. Este requiere realizar menos supuestos, lo que permite obtener resultados más robustos y es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas. Permite estudiar el impacto que tiene cada una de las variables independientes en la probabilidad de que ocurra el suceso de estudio. (Gujarati, 2004)

3.2 Regresión Logística con variable respuesta dicotómica

Para este modelo se considera que la variable respuesta, es una variable dicotómica que tomas dos valores. (Menard, 2002)

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.

Para estos modelos dicotómicos, las dos categorías deben de ser mutuamente excluyentes. La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \Pr ob(Y_i = 1) = P_i \\ 0, \Pr ob(Y_i = 0) = 1 - P_i \end{cases} \qquad (3)$$

3.3 Regresión Logística Bayesiana con Enlaces Simétricos

Se considera un modelo de regresión binaria:

$Y_i \sim Bernoulli(p_i)$ (4)

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

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

donde t = $x_i'\beta$

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

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

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

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

Luego la distribución a posteriori estará dada por:

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

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

3.4 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:

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

donde t = $x_i'\beta$

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

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

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

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

Luego la distribución a posteriori estará dada por:

$$f(\beta|\mathrm{y}) \propto L(\beta)f(\beta) \qquad (14)$$

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

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 Monte Carlo basados en cadenas de Markov. (Chen, 2001)

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

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

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

donde t = $x_i'\beta$

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

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

Se considera que la priori β que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ y la priori para λ que sigue una distribución gamma con parámetros α, δ; 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:

$$f(\beta, \lambda | \mathrm{y}) \; \alpha \; L(\beta, \lambda) f(\beta) f(\lambda) \quad (20)$$

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

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:

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

Se considera que la priori β que sigue una distribución normal multivariada p-dimensional con vector de medias 0 y matriz de covarianza Σ y la priori para λ que sigue una distribución gamma con parámetros α, δ; 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:

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

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

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.

CAPITULO IV

Medidas de bondad de ajuste

4.1 Indicadores estadísticos bayesianos

En esta sección se revisarán diferentes criterios para la comparación de modelos estadísticos 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 Louis (2001) y Spiegelhalter et al. (2002).

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

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

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

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 ρ_D el número efectivo de parámetros, se define ρ_D como:

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

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:

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

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):

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

- El esperado del criterio de información de Akaike (EAIC), propuesto por Carlin y Louis (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.

4.2 Tabla de clasificación

La tabla de clasificación muestra la distribución de valores observados y pronosticados. Los valores observados son los valores reales y los valores pronosticados se obtienen a partir del modelo estadístico bayesiano. Esto se aprecia en la tabla 1.

López I. y Píta S. (2001) mencionaron 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.

Tabla 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)

Haciendo uso de la tabla 1, la sensibilidad y especificidad se calculan 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.

4.3 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.

CAPITULO V

Caso de aplicación: clasificación de clientes con propensión a caer en mora en una entidad bancaria

5.1 Población

La población a estudiar fueron los clientes de la Entidad bancaria que tienen una tarjeta de crédito.

5.2 Muestra

La muestra a estudiar fueron los clientes de la Entidad bancaria que tienen una tarjeta de crédito con una antigüedad de 12 meses, la cual asciende a 5,000 personas naturales.

5.3 Unidad de análisis

La unidad de análisis fue el cliente de la Entidad Bancaria que posee el producto tarjeta de crédito.

5.4 Fuentes de información

a) Data Warehouse de la Entidad Bancaria

Es el repositorio central de información de la Entidad Bancaria, que tiene información sociodemográfica, socioeconómica y financiera de los clientes.

b) Reporte Crediticio Consolidado

El Reporte Crediticio Consolidado (RCC), es el reporte que emite la Superintendencia de Banca, Seguros y AFP (SBS) de manera mensual de todos los deudores de las Entidades Financieras que reportaron información sobre las deudas, líneas y otras variables relativas a la morosidad de sus clientes.

5.5 Definición de variables

a) Variable dependiente

La variable dependiente es la Mora60 que 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).

$$Y_i = f(x) = \begin{cases} 1, & si\ P(Y_i = 1) = P_i, si\ tiene\ mora\ 60\ dias \\ 0, & si\ P(Y_i = 0) = 1 - P_i, si\ no\ tiene\ mora\ de\ 60\ dias \end{cases} \quad (34)$$

b) Variables independientes

Las variables independientes son las siguientes:

- 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 cliente 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.

5.6 Muestra de entrenamiento y validación

Se realizó la aplicación mediante el análisis de 5,000 personas naturales, correspondiente a la base de datos de una 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.

5.7 Análisis de significancia entre la variable dependiente e 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.

La tabla 2 presenta la significancia entre la variable Mora60 y las variables independientes haciendo uso de la prueba Chi-Cuadrado de Pearson. En esta tabla 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 entres las variables independientes 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.

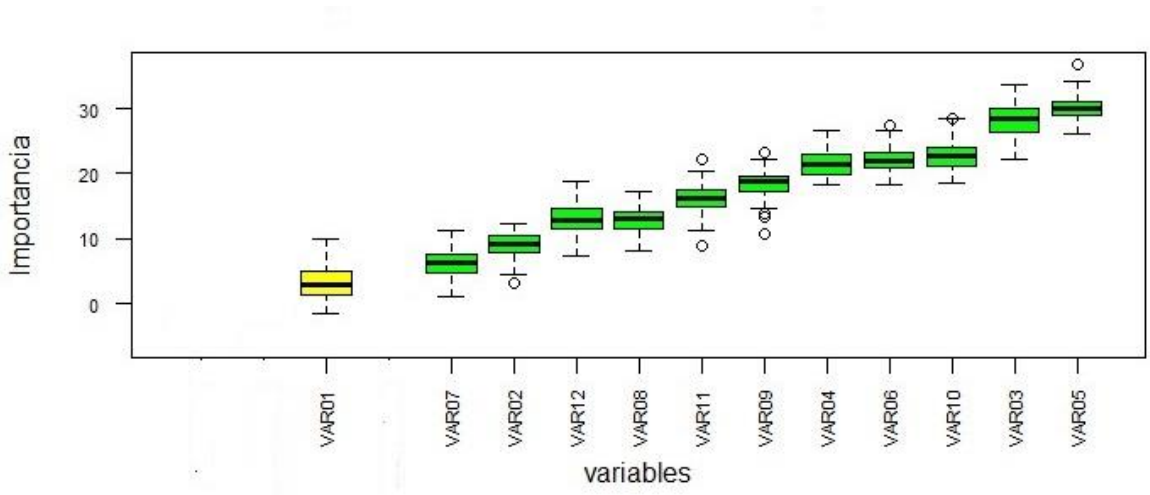
Tabla 2. *Prueba Chi-Cuadrado de Pearson entre la variable dependiente y las variables independientes*

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

5.8 Selección de variables independientes

Se usó el algoritmo de boruta para la selección de variables y se encontró según la figura 1, 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 estadístico de regresión logística con enlaces asimétricos.

Figura. 1 *Importancia de las variables independientes*



5.9 Modelos de regresión logística bayesiana con enlaces asimétricos

a) Modelos de regresión logística bayesiana con enlace asimétrico cloglog

En la tabla 3, se muestra el resultado del modelo de regresión logística bayesiana 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 mora60. 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.

Tabla 3. *Modelo de regresión logística bayesiana 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

De este modo, el modelo de regresión logística bayesiana con enlace asimétrico cloglog se presentó 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 haciendo uso de las variables más importantes, fue 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 de la tabla 3, 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:

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

b) Modelos de regresión logística bayesiana con enlace asimétrico power logit

En la tabla 4, se muestra el resultado del modelo de regresión logística bayesiana 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 mora60. 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.

Tabla 4. *Modelo de regresión logística bayesiana 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

De este modo, el modelo de regresión logística bayesiana con enlace asimétrico power logit se presentó 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 haciendo uso de las variables más importantes, fue 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 de la tabla 4, 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:

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

$\lambda = 0.7$

c) Modelos de regresión logística bayesiana con enlace asimétrico scobit

En la tabla 5, se muestra el resultado del modelo de regresión logística bayesiana 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 mora60. 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.

Tabla 5. *Modelo de regresión logística bayesiana 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

De este modo, el modelo de regresión logística bayesiana con enlace asimétrico scobit se presentó 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 haciendo uso de las variables más importantes, fue el siguiente:

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

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

donde t = $x_i'\beta$

Por lo tanto, usando los coeficientes Mean B de la tabla 5, 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:

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

$\lambda = 2.2$

5.10 Comparación de modelos de regresión logística bayesiana con enlaces asimétricos

a) Uso de indicadores bayesianos

En tabla 6, se presentan los indicadores bayesianos para los modelos de regresión logística bayesiana con enlaces asimétricos cloglog, power logit y scobit.

De la tabla 6, 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.

Tabla 6. *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

b) Uso de la sensibilidad de la tabla de clasificación

Las tablas 7, 8 y 9, presentan las tablas de clasificación de los modelos de regresión logística bayesiana con enlaces asimétricos cloglog, power logit y scobit, respectivamente.

Tabla 7. *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%

Tabla 8. *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%

Tabla 9. *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%

De las tablas 7, 8 y 9, 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 fue el modelo de regresión logística bayesiana con enlace asimétrico cloglog.

c) Uso de la curva ROC

La tabla 10, presenta el área bajo la 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.

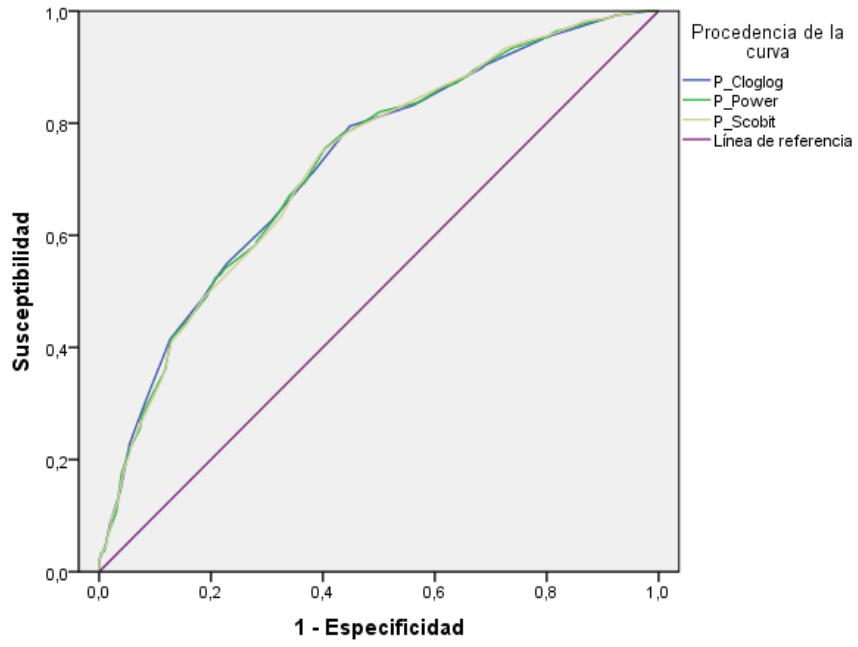
Tabla 10. *Área bajo la curva para los modelos de regresión logística bayesiana con enlaces asimétricos*

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

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. 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 2 las curvas ROC muy similares para los modelos cloglog, power logit y scobit.

Figura. 2 *Curva ROC para los modelos de regresión logística bayesiana con enlaces asimétricos*



CAPITULO VI

Conclusiones

En esta parte se exponen las conclusiones obtenidas como producto del desarrollo del caso de aplicación, las cuales son las siguientes:

1. El modelo de regresión bayesiana 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 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.

2. Los modelos de regresión 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. 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.

3. 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 según su probabilidad de mora.

REFERENCIAS BIBLIOGRAFICAS

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.

Basualdo Najera, S. L. (2014). *Aplicación de la regresión logística para la clasificación de personas naturales bancarizadas con propensión a adquirir una tarjeta de crédito de una entidad bancaria*. Tesis, Universidad Nacional de Ingeniería. http://hdl.handle.net/20.500.14076/5999

Bazán, J. & 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.

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., & Louis, T. A. (2001). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Chapman & Hall.

Chen, M. H., Dey, D. K., & Shao, Q-M. (1999). *A new skewed link model for dichotomous quantal response data*. Journal of the American Statistical Association, 94, 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, 201-202.

Fernández Vásquez, R. F. (2018). *Regresión bayesiana con enlaces asimétricos para la clasificación de clientes con propensión a caer en mora en una entidad bancaria*. Tesis, Universidad Nacional Agraria La Molina. https://hdl.handle.net/20.500.12996/3093

Gujarati, D. N. (2004). *Econometría* (4ta ed.). McGraw-Hill Interamericana.

Kursa, M., & Rudnicki, W. (2010). *Feature Selection with Boruta Package*. Journal of Statistical Software, 36 (11).

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

López, I., & Pita, S. (2001). *Curvas ROC*. Unidad de Epidemiología Clínica y Bioestadística. España.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). SAGE Publications.

Monografias.com. (s.f.). *Tarjetas de crédito*. https://acortar.link/9CWVUH

Novales, A. (1993). *Econometría* (2nd ed.). McGraw-Hill.

Pérez, J. M., Negrín, M. A., García, C., & Gómez, E. (2014). *Bayesian Asymmetric Logit Model for Detecting Risk Factors in Motors Ratemaking*. Astin Bulletin, 44 (2), 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, 761-768.

Ross, S. (1995). *Stochastic Processes*. Wiley.

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

Superintendencia de Banca, Seguros y AFP. (2008). *Resolución S.B.S. N° 11356 - 2008*. https://acortar.link/UA9in3

Superintendencia de Banca, Seguros y AFP. (2011). *Resolución S.B.S. N° 3780 - 2011*. https://acortar.link/WRo5kv

Superintendencia de Banca, Seguros y AFP. (2023). *Resolución S.B.S. N° 0053 - 2023*. https://acortar.link/0PyNdb

DR. RICHARD FERNANDO FERNÁNDEZ VÁSQUEZ.

AUTOR PRINCIPAL

ORCID: https://orcid.org/0000-0003-1721-8527

Peruano, actualmente se desempeña como docente de Pregrado y Posgrado en diferentes universidades en el Perú, entre la cual destaca la Universidad Nacional de Ingeniería. Asesor y revisor de Tesis en Pregrado y Posgrado.

Doctor en Estadística y Matemática en temas de Deep Learning e IA. Magister en Marketing y Gestión Comercial por la Universidad Peruana de Ciencias Aplicadas y la EOI de España. Magister Scientiae en Estadística Aplicada por la Universidad Nacional Agraria La Molina. Ingeniero Estadístico por la Universidad Nacional de Ingeniería. Con Especialización en Deep Learning por la Secretaria de Trabajo de México, Especialización en Gestión Integral de Riesgos por Equilibrium Financiero, Especialización en Ciencia de Datos por la Universidad Nacional Autónoma de México, Especialización en Innovación y Tecnología por San Francisco State University.

Con más de 12 años de experiencia liderando proyectos de Ciencia de Datos. Ha sido SubDirector de Inteligencia y Prospectiva Turística en PROMPERÚ, Head de Modelos & Data Analytics en Agrobanco, Data Science Head en el Grupo El Comercio, Data Science Lead en SUNAT, Subgerente Adjunto de Gestión de Portafolio de Productos en el Banco de Crédito del Perú, Analista de Business Analytics en Banco Financiero y Banco Falabella. Ganador del premio ORO en la categoría de Big Data, Analítica e Inteligencia Artificial de los Premios País a los Innovadores Financieros en las Américas 2023 por Fintech Americas. Reconocido por San Francisco State University en los Estados Unidos por contribuir de manera valiosa en el Programa de Innovación y Tecnología en San Francisco - Silicon Valley 2019 para los estudiantes del MBA de la Universidad de Lima.

MAG. SOLANGE LESLY BASUALDO NAJERA.

AUTOR PRINCIPAL

ORCID: https://orcid.org/0009-0001-0567-689X

Peruana, se desempeña como docente de Pregrado y Posgrado en diferentes universidades en el Perú, destacando la Universidad Nacional de Ingeniería.

Magister en Marketing y Gestión Comercial por la Universidad del Pacífico. Ingeniera Estadística por la Universidad Nacional de Ingeniería. Con Especialización en Data Storytelling por INCAE Business School, Especialización en Data Science and Machine Learning por Massachusetts Institute Technology.

Con más de 14 años de experiencia liderando proyectos de Ciencia de Datos. Se desempeña como Lead Data Analytics en Zubale México, ha sido Head de Business Analytics en Bankea, Head de Data Analytics en Ideas Diferentes, Analista de Modelos Analíticos en Scotiabank, Analista de Inteligencia de Negocios en Banco Falabella y Analista de Riesgos en MiBanco.

Enero 2025 - CID - Centro de Investigación y Desarrollo
Copyright © - CID - Centro de Investigación y Desarrollo
Copyright del texto © 2024 de Autores
Formato: PDF
Tamaño: A4 210 x 297 mm
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
biblioteca.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 2690 3834