PLOS ONE

RESEARCH ARTICLE

Corrección de la clasificación binaria desequilibrada: un enfoque de aprendizaje bayesiano asimétrico

Letí´ciaF.M.Reis,DiegoC.Nascimento,PauloH.Ferreira[3]*, Francisco Louzada[1]

1 Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil, **2** Department of Mathematics, University of Atacama, Copiapó´ , Atacama, Chile, **3** Department of Statistics, Federal University of Bahia, Salvador, Bahia, Brazil

* paulohenri@ufba.br

Citation: Reis LFM, Nascimento DC, Ferreira PH, Louzada F (2024) Fixing imbalanced binary classification: An asymmetric Bayesian learning approach. PLoS ONE 19(10): e0311246. https://doi.org/10.1371/journal.pone.0311246

Editor: Qichun Zhang, Buckinghamshire New University - High Wycombe Campus: Buckinghamshire New University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: April 28, 2024

Accepted: September 16, 2024

Published: October 16, 2024

Data Availability Statement: All data sets and all codes developed for this work can be found on the first author's GitHub: https://github.com/leticiaferreiramurca/Msc.

Funding: Prof Paulo Ferreira received funding in the form of a grant from Conselho Nacional de Desenvolvimento Cientí´fico e Tecnoló´gico, (307221/2022-9). Francisco Louzada received funding in the form of a grant from Fundaç¸ão de

Resumen

La mayoría de los modelos estadísticos y de aprendizaje automático utilizados para el modelado y la clasificación de datos binarios suponen que los datos están equilibrados. Sin embargo, esta suposición puede conducir a un rendimiento predictivo deficiente y a un sesgo en la estimación de parámetros cuando hay un desequilibrio en los datos debido a la elección del umbral para la clasificación binaria. Para abordar este desafío, varios autores sugieren utilizar funciones de enlace asimétricas en la regresión binaria, en lugar de las funciones simétricas tradicionales como logit o probit, con el objetivo de resaltar las características que ayudarían en la tarea de clasificación. Por lo tanto, este estudio tiene como objetivo introducir nuevas funciones de clasificación basadas en la distribución Lomax (y sus variaciones; incluidas las versiones de potencia e inversa). Las funciones bayesianas propuestas han demostrado ser asimétricas y se implementaron en un programa Stan en el flujo de trabajo R. Además, estas funciones mostraron resultados prometedores en aplicaciones de datos del mundo real, superando a las funciones de enlace clásicas en términos de métricas. Por ejemplo, en el primer ejemplo, la comparación del modelo Lomax doble de potencia inversa (RPDLomax) con el enlace logit mostró que, independientemente del desequilibrio de los datos, el modelo RPDLomax asigna efectivamente probabilidades predictivas posteriores medias más bajas para el fracaso y probabilidades más altas para el éxito (21,4% y 63,7%, respectivamente), a diferencia de la regresión logística, que no distingue claramente entre las probabilidades predictivas posteriores medias para estas dos clases (36,0% y 39,5% para el fracaso y el éxito, respectivamente). Es decir, el enfoque Lomax asimétrico propuesto es un modelo competitivo para diferenciar la clasificación de datos binarios en tareas desequilibradas frente al enfoque logístico.

1 Introducción

Diversas tareas de clasificación se llevan a cabo de manera inconsciente a diario. Se ordena la ropa y se coloca en los cajones correspondientes, se priorizan los mensajes y los correos electrónicos, se clasifican los platos como sucios o limpios y se evalúan las tareas en función de los niveles de dificultad. Si bien estas actividades parecen sencillas, no conllevan consecuencias críticas si se clasifican incorrectamente. Por el contrario, el hecho de que un profesional médico no diagnostique un cáncer puede provocar la muerte del paciente en un corto período de tiempo. De manera similar, la clasificación errónea frecuente de los clientes como buenos o malos pagadores por parte de una institución financiera puede dar lugar a pérdidas financieras significativas, que podrían alcanzar miles de millones. La importancia de un modelado y una clasificación precisos se hace evidente en escenarios en los que hay mucho en juego que subraya su papel fundamental en diversos dominios.

PLOS ONE

RESEARCH ARTICLE

Corrección de la clasificación binaria desequilibrada: un enfoque de aprendizaje bayesiano asimétrico

Letí´ciaF.M.Reis,DiegoC.Nascimento,PauloH.Ferreira[2] [3] *, Francisco Louzada[1]

1 Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil, **2** Department of Mathematics, University of Atacama, Copiapó´ , Atacama, Chile, **3** Department of Statistics, Federal University of Bahia, Salvador, Bahia, Brazil

* paulohenri@ufba.br

Citation: Reis LFM, Nascimento DC, Ferreira PH, Louzada F (2024) Fixing imbalanced binary classification: An asymmetric Bayesian learning approach. PLoS ONE 19(10): e0311246. https://doi.org/10.1371/journal.pone.0311246

Editor: Qichun Zhang, Buckinghamshire New University - High Wycombe Campus: Buckinghamshire New University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: April 28, 2024

Accepted: September 16, 2024

Published: October 16, 2024

Data Availability Statement: All data sets and all codes developed for this work can be found on the first author's GitHub: https://github.com/leticiaferreiramurca/Msc.

Funding: Prof Paulo Ferreira received funding in the form of a grant from Conselho Nacional de Desenvolvimento Cientí´fico e Tecnoló´gico, (307221/2022-9). Francisco Louzada received funding in the form of a grant from Fundaç¸ão de

Resumen

La mayoría de los modelos estadísticos y de aprendizaje automático utilizados para el modelado y la clasificación de datos binarios suponen que los datos están equilibrados. Sin embargo, esta suposición puede conducir a un rendimiento predictivo deficiente y a un sesgo en la estimación de parámetros cuando hay un desequilibrio en los datos debido a la elección del umbral para la clasificación binaria. Para abordar este desafío, varios autores sugieren utilizar funciones de enlace asimétricas en la regresión binaria, en lugar de las funciones simétricas tradicionales como logit o probit, con el objetivo de resaltar las características que ayudarían en la tarea de clasificación. Por lo tanto, este estudio tiene como objetivo introducir nuevas funciones de clasificación basadas en la distribución Lomax (y sus variaciones; incluidas las versiones de potencia e inversa). Las funciones bayesianas propuestas han demostrado ser asimétricas y se implementaron en un programa Stan en el flujo de trabajo R. Además, estas funciones mostraron resultados prometedores en aplicaciones de datos del mundo real, superando a las funciones de enlace clásicas en términos de métricas. Por ejemplo, en el primer ejemplo, la comparación del modelo Lomax doble de potencia inversa (RPDLomax) con el enlace logit mostró que, independientemente del desequilibrio de los datos, el modelo RPDLomax asigna efectivamente probabilidades predictivas posteriores medias más bajas para el fracaso y probabilidades más altas para el éxito (21,4% y 63,7%, respectivamente), a diferencia de la regresión logística, que no distingue claramente entre las probabilidades predictivas posteriores medias para estas dos clases (36,0% y 39,5% para el fracaso y el éxito, respectivamente). Es decir, el enfoque Lomax asimétrico propuesto es un modelo competitivo para diferenciar la clasificación de datos binarios en tareas desequilibradas frente al enfoque logístico.

1 Introducción

Diversas tareas de clasificación se llevan a cabo de manera inconsciente a diario. Se ordena la ropa y se coloca en los cajones correspondientes, se priorizan los mensajes y los correos electrónicos, se clasifican los platos como sucios o limpios y se evalúan las tareas en función de los niveles de dificultad. Si bien estas actividades parecen sencillas, no conllevan consecuencias críticas si se clasifican incorrectamente. Por el contrario, el hecho de que un profesional médico no diagnostique un cáncer puede provocar la muerte del paciente en un corto período de tiempo. De manera similar, la clasificación errónea frecuente de los clientes como buenos o malos pagadores por parte de una institución financiera puede dar lugar a pérdidas financieras significativas, que podrían alcanzar miles de millones. La importancia de un modelado y una clasificación precisos se hace evidente en escenarios en los que hay mucho en juego que subraya su papel fundamental en diversos dominios.

Amparo à Pesquisa do Estado de São Paulo, (2013/07375-0).

Competing interests: The authors have declared that no competing interests exist.

El término "datos binarios desequilibrados" denota un conjunto de datos en el que una de las clases supera significativamente a la otra en términos de observaciones. Por ejemplo, en la predicción de impagos, normalmente hay más buenos pagadores que malos pagadores, y en la detección del cáncer, el número de individuos sanos supera con creces a los diagnosticados con la enfermedad. Este desequilibrio plantea un desafío para el modelado y la clasificación de datos binarios, ya que la mayoría de los algoritmos de aprendizaje automático y los modelos estadísticos presuponen una distribución uniforme de las observaciones en ambas categorías [1]. En consecuencia, estos algoritmos a menudo priorizan la clase mayoritaria sobre la clase minoritaria, aunque la clase minoritaria suele ser de mayor interés.

En la regresión binaria, a menudo se eligen funciones de enlace simétricas, como las funciones logit y probit, lo que implica probabilidades iguales para ambas categorías. Sin embargo, investigaciones recientes, ejemplificadas por [2], sostienen que las funciones de enlace asimétricas pueden adaptarse mejor a la tarea de manejar datos desequilibrados al permitir probabilidades distintas para cada categoría. La adopción de funciones de enlace simétricas en tales casos, como se señala en [3], puede resultar en un sesgo sustancial en la estimación de parámetros y la predicción de la probabilidad de éxito.

Para abordar este desafío, [4] propuso un método para transformar distribuciones conocidas para derivar funciones de enlace más flexibles y asimétricas. Este enfoque implica exponenciar las funciones de distribución acumulativas existentes mediante el parámetro positivo λ, que actúa como un parámetro de asimetría y proporciona control sobre la forma de la distribución. Esta transformación introduce asimetría y mantiene una conexión con las funciones de enlace simétricas.

El objetivo principal de este trabajo es introducir nuevas funciones de clasificación asimétrica basadas en la distribución Lomax para mejorar el modelado y la clasificación de datos binarios desequilibrados. Los modelos propuestos ofrecen la ventaja de incorporar solo un parámetro adicional, estableciendo así una alternativa definida paramétricamente a los métodos de regresión binaria tradicionales. Este enfoque garantiza la interpretabilidad, aprovechando el método de generación de asimetría propuesto por [4].

1.1 Contribuciones

Las principales contribuciones de los modelos propuestos se definen a continuación:
i) Proponer nuevas funciones de clasificación asimétricas que superen a las funciones simétricas tradicionales en la clasificación de datos desequilibrados.

ii) Comparar las funciones de enlace propuestas con las funciones de enlace tradicionales (logit, probit, cauchit, loglog y cloglog), centrándose en el enlace logit debido a su popularidad en la clasificación.

iii) Introducir nuevas funciones asimétricas que requieren solo un parámetro adicional para generar asimetría. Este enfoque reduce la varianza en la estimación de parámetros y mejora la estabilidad del modelo, ofreciendo una alternativa más robusta a otras funciones de clasificación asimétricas.

iv) Proporcionar un modelo que pueda implementarse fácilmente dentro del flujo de trabajo R, facilitando su adopción y aplicación en el análisis de datos del mundo real.

v) Asegurar que los modelos paramétricos sean interpretables, con todos los parámetros claramente definidos. La inclusión del parámetro de asimetría λ permite una relación directa con los modelos de clasificación simétricos, regulando la asimetría del modelo.

vi) Presentar un enfoque bayesiano para interpretar los modelos, ofreciendo una comprensión integral de las distribuciones de parámetros y su impacto en la clasificación.

1.2 Organización del artículo

La estructura de este artículo es la siguiente. En la Sección 2 se discuten trabajos relacionados en la literatura. La Sección 3 presenta las distribuciones asimétricas de Lomax. La Sección 4 presenta el modelo de regresión binaria bayesiana utilizando las distribuciones introducidas en la Sección 3. La Sección 5 proporciona los resultados de las simulaciones y el análisis bayesiano de estas distribuciones. En la Sección 6, la metodología propuesta se ejemplifica en dos aplicaciones: una relacionada con la clasificación de imágenes (wilt) y otra relacionada con la donación de sangre. La Sección 7 analiza los resultados de este trabajo. Finalmente, la Sección 8 contiene algunas observaciones finales y futuras direcciones de investigación.

2 Revisión de literatura

En la literatura académica, varios autores han explorado funciones de enlace asimétricas como alternativas a los modelos simétricos convencionales. Ejemplos notables incluyen [5], que exploró el desempeño de varias funciones de enlace asimétricas para predecir la mortalidad en seguros de vida, y [6], que empleó la distribución t de Student asimétrica para identificar pacientes con enfermedad de Parkinson. Además, [7] investigó las funciones de enlace de Fréchet, Weibull y Gumbel para modelar las ocurrencias de quiebra en pequeñas y medianas empresas. Sin embargo, es crucial reconocer que estos modelos a menudo carecen de un mecanismo para controlar la asimetría a través de un parámetro adicional, lo que dificulta establecer conexiones con modelos simétricos convencionales [2].

Para abordar esta brecha, nuestro estudio apunta a desarrollar un nuevo modelo aplicando la transformación de [4] a la distribución Lomax. Este artículo explora el método de Bazán dentro de este contexto, ofreciendo un enfoque novedoso para modelar y clasificar datos binarios desequilibrados con parámetros adicionales mínimos y manteniendo la interpretabilidad.

3 Distribuciones asimétricas de Lomax

En esta sección, presentaremos los modelos asimétricos Lomax, a saber, el modelo de doble Lomax de potencia (PDLomax) y el modelo de doble Lomax de potencia inversa (RPDLomax). Primero, proporcionaremos la definición de la distribución doble Lomax (DLomax) propuesta por [17], que es una extensión de la distribución Lomax (también conocida como distribución Pareto Tipo II) en la línea real. Para más detalles sobre la distribución Lomax, véase, por ejemplo, [18, 19]. Una variable aleatoria X 2 R sigue una distribución DLomax con parámetros m 2 R y σ > 0 si su función de densidad de probabilidad (pdf) y función de distribución acumulativa (cdf) están dadas, respectivamente, por

$$g(x) = \frac{1}{2\sigma\left(1+\left|\frac{x-\mu}{\sigma}\right|\right)^2} \quad \text{and} \quad G(x) = \begin{cases} \dfrac{1}{2\left(1+\frac{\mu-x}{\sigma}\right)}, & x \le \mu, \\ 1 - \dfrac{1}{2\left(1+\frac{x-\mu}{\sigma}\right)}, & x > \mu. \end{cases}$$

Si μ = 0 y σ = 1, entonces se denomina distribución DLomax estándar. [17] derivaron esta distribución a partir de la relación de dos variables aleatorias clásicas de Laplace estándar independientes e idénticamente distribuidas.

Para construir nuevas funciones de enlace asimétricas utilizando la distribución estándar DLomax como base, consideramos la transformación de potencia propuesta por [4]. Así, si FP es una distribución de potencia con distribución base G, entonces su función de densidad de probabilidad (pdf) y función de distribución acumulada (cdf) se describen mediante el siguiente proceso de exponenciación:

$$f_{\mathrm{P}}(x|\lambda) = \lambda g(x)[G(x)]^{\lambda-1} \quad \text{and} \quad F_{\mathrm{P}}(x|\lambda) = [G(x)]^{\lambda},$$

con λ > 0. Ahora, FRP es una distribución de potencia inversa con distribución base G, y su pdf y cdf se describen de la siguiente manera:

$$f_{\mathrm{RP}}(x|\lambda) = \lambda g(-x)[G(-x)]^{\lambda-1} \quad \text{and} \quad F_{\mathrm{RP}}(x|\lambda) = 1 - [G(-x)]^{\lambda},$$

donde λ > 0. Para varias propiedades que establecen la relación entre las distribuciones de potencia e inversa, véase, por ejemplo, el trabajo de [15].

En consecuencia, una variable aleatoria X tiene una distribución PDLomax estándar con un parámetro de asimetría λ > 0, si su función de densidad de probabilidad se puede escribir de la siguiente manera:

$$f_{\mathrm{P}}(x) = \begin{cases} \frac{\lambda}{2(1+|x|)^2}\left[\frac{1}{2(1-x)}\right]^{\lambda-1}, & x \le 0, \\ \frac{\lambda}{2(1+|x|)^2}\left[1-\frac{1}{2(1+x)}\right]^{\lambda-1}, & x > 0, \end{cases}$$

con cdf dada por

$$F_{\mathrm{P}}(x) = \begin{cases} \left[\frac{1}{2(1-x)}\right]^{\lambda}, & x \le 0, \\ \left[1-\frac{1}{2(1+x)}\right]^{\lambda}, & x > 0. \end{cases}$$

Además, una variable aleatoria X sigue una distribución RPDLomax estándar con un parámetro de asimetría λ > 0, si su función de densidad de probabilidad se describe de la siguiente manera:

$$f_{\mathrm{RP}}(x) = \begin{cases} \frac{\lambda}{2(1+|x|)^2}\left[1-\frac{1}{2(1-x)}\right]^{\lambda-1}, & x \le 0, \\ \frac{\lambda}{2(1+|x|)^2}\left[\frac{1}{2(1+x)}\right]^{\lambda-1}, & x > 0, \end{cases}$$

con cdf dada por

$$F_{\mathrm{RP}}(x) = \begin{cases} 1 - \left[1 - \frac{1}{2(1-x)}\right]^{\lambda}, & x \leq 0, \\ 1 - \left[\frac{1}{2(1+x)}\right]^{\lambda}, & x > 0. \end{cases}$$

Nótese que FRP(x) = 1 – FP(x). Por lo tanto, las distribuciones PDLomax y RPDLomax son distintas pero están estrechamente relacionadas, ya que una refleja a la otra. Además, FP(–x) ¼6 1 – FP(x) y FRP(–x) ¼6 1 – FRP(x), lo que demuestra que FP y FRP no son simétricas.

Hasta donde sabemos, estas dos distribuciones de probabilidad no se han presentado en la literatura. Sin embargo, su caracterización completa (propiedades estadísticas como los momentos, incluida la media, la varianza, la asimetría y la curtosis) no está dentro del alcance de este artículo y se analizará en el futuro.

Las figuras 1 y 2 muestran, respectivamente, los gráficos pdf y cdf de las distribuciones estándar PDLomax y RPDLomax para varios valores de λ.

La figura 1 muestra que la adición del parámetro λ puede introducir asimetría tanto hacia la derecha (positiva) como hacia la izquierda (negativa). En particular, para la distribución estándar PDLomax (panel izquierdo), cuando λ < 1, la curva de densidad se concentra hacia la izquierda; cuando λ > 1, la curva de densidad se concentra hacia la derecha; y en λ = 1, tenemos la distribución original (distribución estándar DLomax). La distribución estándar RPDLomax (panel derecho) exhibe el comportamiento opuesto, ya que ambas funciones son reflejos entre sí.

En la Fig. 2, se puede observar cómo la variación de λ modifica la función de densidad acumulada. Nótese que la variación de este parámetro no solo afecta al intervalo donde se concentra la mayor probabilidad, sino que también afecta a la pendiente de la curva de probabilidad. En particular, para la distribución estándar PDLomax (panel izquierdo), cuando λ < 1, es más probable que X < 0, pero cuando λ > 1, es más probable que X > 0. Es decir, en el contexto de la regresión binaria, en la que se utiliza esta cdf como función de enlace, se espera que cuando λ < 1, haya una mayor proporción de fallos que de éxitos (más 0 que 1); y cuando λ > 1, debe haber una mayor proporción de éxitos. Cuando λ = 1, es un caso atípico en el que éxitos y fallos estarán equilibrados. Para la distribución estándar RPDLomax (panel derecho), se observa un comportamiento inverso.

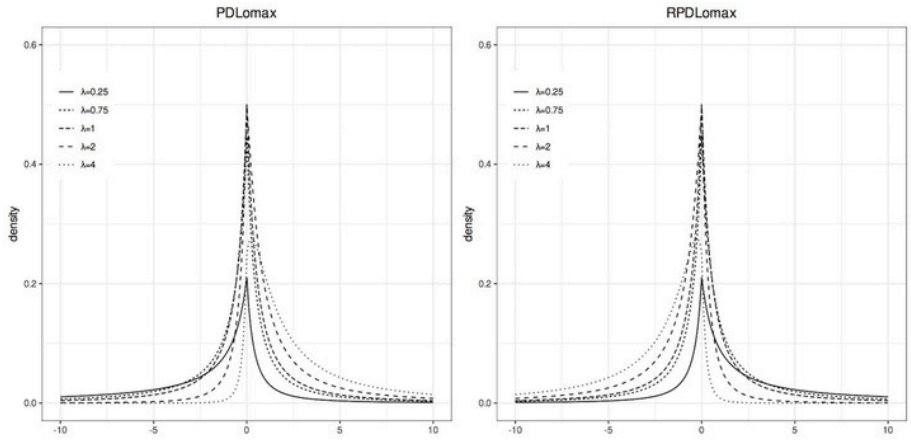

Fig. 1. Funciones de densidad de probabilidad de las distribuciones estándar PDLomax y RPDLomax en varios valores de λ.

https://doi.org/10.1371/journal.pone.0311246.g001

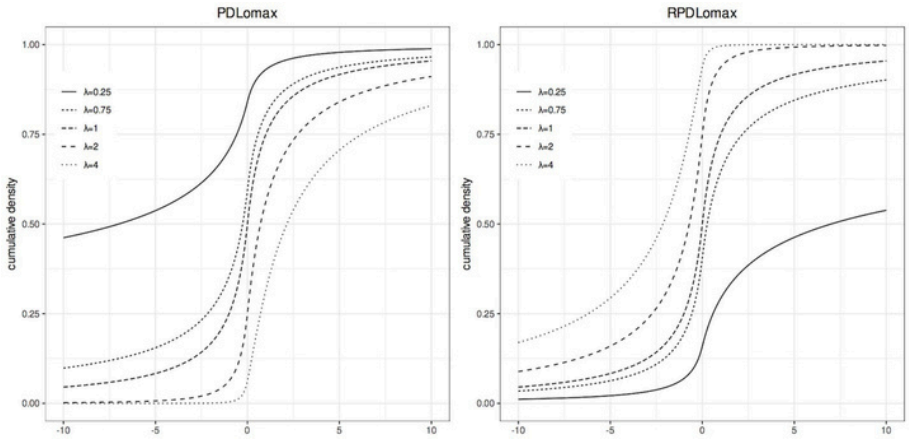

Fig 2. Cumulative distribution functions of the standard PDLomax and RPDLomax distributions at various values of λ.

https://doi.org/10.1371/journal.pone.0311246.g002

El parámetro λ afecta directamente la velocidad a la que aumenta la función de distribución de coeficientes. Un valor λ mayor da como resultado una pendiente más pronunciada de la función de distribución de coeficientes, lo que hace que se acerque a 1 más rápidamente a medida que x aumenta. Por el contrario, un valor λ menor produce una pendiente más suave. En la distribución RPDLomax, una pendiente más suave (λ < 1) corresponde a una mayor proporción de ceros, lo que le permite gestionar mayores desequilibrios de ceros de manera más efectiva que la distribución PDLomax. Por otro lado, la distribución PDLomax exhibe pendientes más suaves (λ < 1) cuando hay más unos que ceros, lo que le permite gestionar mayores proporciones de unos.

4 Modelo de regresión binaria bayesiana

En esta sección, presentaremos el nuevo modelo asimétrico de Lomax para regresión binaria utilizando las distribuciones introducidas en la sección anterior. Utilizando la notación definida anteriormente, este modelo puede describirse mediante el siguiente conjunto de ecuaciones:

$$\begin{aligned} Y_i|\boldsymbol{\beta}, \lambda &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i), \\ p_i &= F_\lambda(\eta_i), \\ \eta_i &= \boldsymbol{x}_i'\boldsymbol{\beta}, \\ (\boldsymbol{\beta}, \lambda) &\sim \pi(\boldsymbol{\beta}, \lambda), \end{aligned}$$

donde Fλ representa las distribuciones introducidas previamente, β es el vector de coeficientes de regresión, λ es el parámetro de asimetría introducido por las transformaciones de potencia e inversa, y π es la distribución previa para los parámetros β y λ.

En este trabajo se asumirá que todos los parámetros son independientes, es decir que la distribución a priori viene dada por π(a, b) = π(a)π(b). Por lo tanto, las distribuciones a priori utilizadas en los modelos se basarán en el estudio de [4]. Adicionalmente, siguiendo la recomendación de los autores, el parámetro λ será reparametrizado como δ = log(λ), ya que esta reparametrización mejora la eficiencia de

Estimaciones de parámetros. El conjunto de ecuaciones que se muestra a continuación describe el modelo utilizado:

$$\begin{aligned} Y_i|\boldsymbol{\beta}, \delta &\overset{\text{ind.}}{\sim} \text{Bernoulli}(p_i), \\ p_i &= F_\delta(\eta_i), \\ \eta_i &= \boldsymbol{x}_i'\boldsymbol{\beta}, \\ \beta_j &\overset{\text{ind.}}{\sim} \text{Normal}(0, 10^2), \quad j = 1, 2, \ldots, k, \\ \delta &\sim \text{Uniform}(-2, 2), \end{aligned}$$

donde Fδ representa la cdf de la distribución PDLomax o RPDLomax reparametrizada.

Nótese que la distribución previa de δ es una distribución uniforme en (−2, 2), es decir, λ está restringida al intervalo (e−2, e2) = (0,14, 7,39). La razón de esta elección es que los valores fuera de este rango tienen una probabilidad muy baja de ocurrencia [16]. Además, la asimetría de las distribuciones de potencia permanece prácticamente constante cuando λ > 6 [13]. Se puede observar que la asimetría de las distribuciones PDLomax y RPDLomax es constante fuera del intervalo establecido en esta distribución previa. A pesar de esta parametrización, los resultados en las siguientes secciones se presentarán en términos de λ para mantener una conexión directa con las funciones de enlace simétricas y la tasa de éxito.

La distribución posterior para los modelos de regresión binaria que tienen el parámetro de asimetría λ está dada por

$$\pi(\boldsymbol{\beta}, \delta|\boldsymbol{X}, \boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{\beta}, \delta|\boldsymbol{X}, \boldsymbol{y})\pi(\boldsymbol{\beta})\pi(\delta),$$

Donde π(β) es la distribución previa de ß, con ßj~Normal(0, 10^2), para j = 1, 2, . . ., k; π(δ) es j la distribución previa de δ, con δ = log(λ) ~ Uniforme(-2, 2); y L(B, δ|X, y) es la función de verosimilitud de los parámetros dado el conjunto de datos, representada por la fórmula:

$$\mathcal{L}(\boldsymbol{\beta}, \delta|\boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} [F_\delta(\eta_i)]^{y_i}[1 - F_\delta(\eta_i)]^{1-y_i}.$$

Así, combinando las expresiones descritas anteriormente, la distribución posterior se puede escribir como:

$$\begin{aligned} \pi(\boldsymbol{\beta}, \delta|\boldsymbol{X}, \boldsymbol{y}) &\propto \prod_{i=1}^{n} [F_\delta(\eta_i)]^{y_i}[1 - F_\delta(\eta_i)]^{1-y_i} \prod_{j=1}^{p} \frac{1}{10\sqrt{2\pi}} \exp\left\{ -\frac{\beta_j^2}{2(10^2)} \right\} \frac{1}{4} \\ &\propto \prod_{i=1}^{n} [F_\delta(\eta_i)]^{y_i}[1 - F_\delta(\eta_i)]^{1-y_i} \prod_{j=1}^{p} \exp\left\{ -\frac{\beta_j^2}{2(10^2)} \right\}. \end{aligned} \tag{1}$$

El enfoque del clasificador bayesiano se considera ventajoso debido a su uso del posterior predictivo, que proporciona una distribución de probabilidad sobre la probabilidad de sensibilidad del resultado i-ésimo, P(Yi = 1|X_1) = p_1. Esto permite una comprensión más matizada de la incertidumbre en las predicciones, lo que lo hace particularmente útil cuando se trabaja con datos limitados (debido a la posibilidad de incorporar anteriores). Por el contrario, el enfoque frecuentista generalmente proporciona estimaciones puntuales sin capturar el rango completo de incertidumbre (solo un criterio de decisión de estimación puntual), lo que puede conducir a predicciones menos sólidas. Por lo tanto, el enfoque bayesiano, al centrarse en el posterior predictivo, ofrece de manera probabilística un marco de incertidumbre más amplio y completo para realizar predicciones.

Toda estimación de parámetros se realizó utilizando el paquete stan en el software R [20].

Este programa realiza sus estimaciones utilizando la técnica No-U-Turn Sampler (NUTS) propuesta por [21]. NUTS es un algoritmo de Monte Carlo de cadena de Markov (MCMC) empleado para la inferencia bayesiana. Es una variante de autoajuste del método de Monte Carlo hamiltoniano (HMC), que explora eficientemente el espacio de parámetros utilizando información de gradiente para evitar paseos aleatorios y proporcionar una convergencia más rápida. NUTS utiliza un algoritmo recursivo para construir un conjunto de puntos candidatos probables, que cubren un amplio rango de la distribución objetivo, y se detiene automáticamente cuando comienza a regresar al mismo lugar. Empíricamente, NUTS funciona al menos tan eficientemente como HMC, y a veces más eficientemente, incluso cuando está bien especificado, con la ventaja de que NUTS opera sin intervención del usuario [21]. Una revisión más detallada del algoritmo NUTS aplicado a modelos de potencia e inversa se puede encontrar en el trabajo de [15].

Para comparar los ajustes del modelo, se utilizaron métricas bayesianas basadas en la media posterior de desviación y la desviación de la media posterior, representadas por las siguientes fórmulas:

$$\overline{D} = \frac{1}{S}\sum_{s=1}^{S} D(\boldsymbol{\beta}^{(s)}, \delta^{(s)}), \quad \text{with} \quad D(\boldsymbol{\beta}^{(s)}, \delta^{(s)}) = -2\log\left(p(\boldsymbol{y}|\boldsymbol{\beta}^{(s)}, \delta^{(s)}\right),$$

and

$$\widehat{D} = D\left(\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\beta}^{(s)}, \frac{1}{S}\sum_{s=1}^{S}\delta^{(s)}\right),$$

para s = 1, . . ., S, donde S es el tamaño de la muestra posterior. A partir de estos valores, se puede calcular el número efectivo de parámetros, $\rho_d = \overline{D} - \widehat{D}$, . Posteriormente, se calculan medidas como el criterio de información de desviación $(\mathrm{DIC} = \overline{D} + \rho_d = 2\overline{D} - \widehat{D})$, el criterio de información de Akaike esperado $(\mathrm{EAIC} = \overline{D} + 2k$, donde k es el número de parámetros del modelo) y el criterio de información bayesiano esperado $(\mathrm{EBIC} = \overline{D} + k\log(n)$, donde n es el tamaño de la muestra, es decir, el número de observaciones de la variable de respuesta Y). Además, se calcula el criterio de información de Watanabe-Akaike

$(\mathrm{WAIC} = -2(\widehat{\mathrm{LPPD}} - \widehat{p}_{\mathrm{WAIC}})$, with $\widehat{\mathrm{LPPD}} = \sum_{i=1}^{n}\log\left(S^{-1}\sum_{s=1}^{S}p(y_i|\boldsymbol{\beta}^{(s)}, \delta^{(s)})\right)$ and $\widehat{p}_{\mathrm{WAIC}} = 2\sum_{i=1}^{n}\left(\log\left(S^{-1}\sum_{s=1}^{S}p(y_i|\boldsymbol{\beta}^{(s)}, \delta^{(s)})\right) - S^{-1}\sum_{s=1}^{S}\log(p(y_i|\boldsymbol{\beta}^{(s)}, \delta^{(s)}))\right)$) and leave-one-out (LOO) metrics are considered.

La métrica LOO, similar a WAIC, es una métrica completamente bayesiana. Sin embargo, tiene un alto costo computacional cuando se trabaja con muestras muy grandes. Por lo tanto, [22] propuso el método de validación cruzada PSIS-LOO (Pareto smoothed importance sampling leave-one-out). Esta métrica se puede calcular de la siguiente manera:

$$\widehat{\mathrm{ELPD}}_{\text{PSIS-LOO}} = \sum_{i=1}^{n}\log\left(\sum_{s=1}^{S}\frac{w_i^{(s)}\,p(y_i|\boldsymbol{\beta}^{(s)}, \delta^{(s)})}{\sum_{s=1}^{S}w_i^{(s)}}\right),$$

dónde

$$w_i^{(s)} = \min\left(r_i^{(s)}, \frac{\sqrt{S}}{S}\sum_{s=1}^{S}r_i^{(s)}\right),$$

con $$r_i^{(s)} = \frac{1}{p(y_i|\boldsymbol{\beta}^{(s)}, \delta^{(s)})} \propto \frac{p(\boldsymbol{\beta}^{(s)}, \delta^{(s)}|\boldsymbol{y}_{-i})}{p(\boldsymbol{\beta}^{(s)}, \delta^{(s)}|\boldsymbol{y})},$$

para $s = 1, \ldots, S.$

Cuanto menor sea el valor de todas estas métricas, mejor se ajustará el modelo. Para evaluar la adecuación del modelo, se utilizaron residuos cuantiles, como propone [23]. Cuando se cumplen los supuestos del modelo, estos residuos siguen una distribución normal, independientemente de la distribución de la variable de respuesta.

5 Simulation studies

In this section, we will present simulation studies that aim to verify the accuracy of the Bayesian estimation procedure (Section 5.1), and evaluate whether the proposed models can perform better than the logistic regression model in different imbalanced scenarios (Section 5.2).

5.1 Parameter recovery

In order to assess the ability of the proposed models to estimate their parameters, a simulation study was conducted. In this study, 100 random samples of size n = {500, 1, 000, 2, 000} were generated for each of the proposed models and also for the logistic model (used for compari- son). For models that include the asymmetry parameter (λ), four additional scenarios were considered: λ = {0.25, 0.5, 2, 4}. The covariate X was simulated from a uniform distribution on the interval (−3, 3), and the regression coefficients were fixed at $\boldsymbol{\beta} = (\beta_0, \beta_1) = (0, 1)$. To estimate the parameters, the Stan software, 200 iterations, 4 chains, and 100 samples of warm-up were considered for each chain. Consequently, for each replica, 400 samples of each estimated parameter were generated. The metrics used to assess the models' performance were bias and root mean squared error (RMSE), which are defined, respectively, as follows:

$$\text{Bias}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta) \quad \text{and} \quad \text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2},$$

where $\theta \in \boldsymbol{\theta} = (\lambda, \boldsymbol{\beta})$ (is one of the three parameters), R is the number of replications in the simulation (in this case, R = 100), θ represents the true parameter value, and $\hat{\theta}_r$ stands for the posterior mean of parameter θ in replication r.

Furthermore, the quality of the interval estimation was verified, observing the relative frequency of times in which the true parameter (θ) was contained in percentiles 2.5% and 97.5%, that is, the proportion of times in which the original parameter was between percentiles 2.5% and 97.5% of the samples of the posterior distribution, here called coverage probability (CP). That is,

$$\text{CP} = \frac{1}{R}\sum_{r=1}^{R} I(\theta \in [LL_r, UL_r]),$$

where LL and UL are, respectively, the lower (2.5%) and upper (97.5%) limits of the posterior samples for each replicate. In other words, we calculated the coverage of the 95% credibility interval of MCMC chains. Hence, it is expected that, in 95% of cases, the real value of the parameter will be contained within this range.

Table 1. Bias, RMSE and CP calculated for the parameters $\beta 0$, $\beta 1$, and λ.

Model	*n*	$\beta 0$			β_1			λ		
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	RMSE	CP
Logistic	500	0.009	0.011	0.96	0.016	0.009	0.95	-	-	-
	1,000	0.007	0.003	0.91	0.007	0.004	0.95	-	-	-
	2,000	0.002	0.008	0.91	0.002	0.001	0.97	-	-	-
DLomax	500	0.011	0.016	0.96	0.066	0.037	0.96	-	-	-
	1,000	0.007	0.009	0.96	0.049	0.020	0.92	-	-	-
	2,000	0.001	0.004	0.95	0.020	0.007	0.93	-	-	-
PDLomax(λ=0.25)	500	0.964	9.806	0.95	0.785	1.059	0.88	0.044	0.345	0.93
	1,000	0.558	3.448	0.94	0.410	0.395	0.90	0.023	0.026	0.91
	2,000	0.317	1.862	0.95	0.176	0.128	0.94	0.009	0.011	0.96
PDLomax(λ=0.5)	500	-0.186	0.279	0.93	0.307	0.242	0.90	-0.039	0.009	0.91
	1,000	-0.033	0.140	0.93	0.161	0.081	0.87	-0.012	0.007	0.91
	2,000	-0.028	0.064	0.95	0.062	0.020	0.94	-0.005	0.004	0.91
PDLomax(λ=2)	500	-0.030	0.119	0.92	0.096	0.035	0.96	-0.002	0.166	0.95
	1,000	-0.014	0.027	0.95	0.053	0.016	0.97	0.006	0.040	0.94
	2,000	0.009	0.014	0.93	0.022	0.006	0.96	0.009	0.019	0.95
PDLomax(λ=4)	500	0.103	0.101	0.97	0.382	0.258	0.90	0.746	1.496	0.96
	1,000	0.200	0.110	0.89	0.224	0.105	0.86	0.748	1.545	0.83
	2,000	0.143	0.066	0.86	0.130	0.053	0.87	0.513	0.898	0.87
RPDLomax(λ=0.25)	500	-0.733	7.506	0.95	0.640	0.801	0.89	0.026	0.223	0.95
	1,000	-0.385	2.266	0.95	0.306	0.269	0.91	0.015	0.019	0.93
	2,000	-0.204	0.862	0.93	0.132	0.095	0.96	0.011	0.005	0.96
RPDLomax(λ=0.5)	500	0.112	0.494	0.87	0.297	0.258	0.92	-0.008	0.026	0.89
	1,000	0.028	0.112	0.96	0.118	0.061	0.93	-0.007	0.005	0.94
	2,000	0.032	0.037	0.97	0.046	0.019	0.94	-0.008	0.002	0.95
RPDLomax(λ=2)	500	-0.047	0.102	0.91	0.088	0.032	0.96	0.010	0.167	0.92
	1,000	-0.027	0.042	0.92	0.063	0.024	0.89	0.052	0.094	0.93
	2,000	0.003	0.011	0.99	0.019	0.008	0.96	-0.003	0.016	0.96
RPDLomax(λ=4)	500	-0.182	0.124	0.93	0.338	0.199	0.87	0.906	1.848	0.98
	1,000	-0.187	0.093	0.91	0.225	0.115	0.84	0.740	1.438	0.85
	2,000	-0.135	0.076	0.89	0.140	0.065	0.88	0.454	1.028	0.86

https://doi.org/10.1371/journal.pone.0311246.t001

Table 1 reveals that in most models, the bias of the estimator $\hat{\beta}_0$ decreases as the sample size increases. The models dealing with more imbalanced data, i.e., with λ = 0.25 and λ = 4, show higher bias and higher RMSE. The same holds for $\beta 1$: its bias also decreases as the sample size increases, as does its RMSE. In comparison to the logistic model, in smaller sample sizes, the proposed models exhibit more bias in parameter recovery. However, as the sample size grows, the difference between these models becomes negligible.

Regarding the parameter λ, a slight increase in bias and RMSE is noticeable from n = 500 to n = 1, 000 when λ = 4. This might occur because, as stated by [13], the relationship between an increase in λ and asymmetry is not linear. Beyond a certain point, any increase in λ results in insignificant increments in asymmetry. Nevertheless, a strong downward trend in both bias and RMSE is evident when n = 2, 000, indicating the potential for reducing bias in the model asymptotically. For the remaining models with the parameter λ, both bias and RMSE decrease as the sample size increases.

Table 2. Mean proportion of 1's, for λ = {0.25, 0.5, 2, 4}, of the samples of the power Cauchy model.

	λ=0.25	λ=0.5	λ=2	λ =
Mean proportion of 1's	0.800	0.661	0.32	**4**
https://doi.org/10.1371/journal.pone.0311246.t002			9	0.1
				98

Finally, the credibility intervals for the parameters of all models seem to have a reasonable behavior, given the number of repetitions of the experiment; however, it is notable that the CP is smaller when λ = 4.

5.2 Misspecification

As in the work of [13], imbalanced data were generated based on the power Cauchy model, with fixed regression coefficients $\beta 0$ and $\beta 1$ set to $\boldsymbol{\beta} = (\beta 0, \beta 1) = (0, 1)$. The (sole) covariate X was simulated from a uniform distribution on the interval (−3, 3). Four different scenarios were simulated with varying levels of imbalance, considering the parameter λ = {0.25, 0.5, 2, 4} of the power Cauchy distribution. The binary regression model using the power Cauchy link function is presented below:

$$Y_i \overset{ind.}{\sim} \text{Bernoulli}(p_i);$$
$$p_i = \frac{1}{2} + \frac{1}{\pi}\arctan(\beta_0 + \beta_1 x_i)^{\lambda}.$$

In this experiment, 100 samples with the power Cauchy distribution were generated following the structure outlined above, each containing 5,000 observations. Table 2 displays the degree of imbalance in each sample.

To compare the fit of the proposed models with the logistic regression model, the WAIC and LOO metrics were examined. These metrics were chosen because they tend to perform better in model selection than other metrics such as DIC, as these only consider point estimates, while the WAIC and LOO metrics take into account the entire posterior distribution of the parameters [24]. In addition, they are fully Bayesian measures. From these metrics, the means of LOO and WAIC in each scenario ($\overline{\text{LOO}}$ and $\overline{\text{WAIC}}$, respectively), the percentage of times that the metric of each link is less than the logistic link (%LOO and %WAIC), and the variance of each of these metrics (s^2_{LOO} and s^2_{WAIC}) were calculated. That is,

$$\overline{\text{LOO}} = \frac{1}{R}\sum_{r=1}^{R} \text{LOO}(r); \qquad \overline{\text{WAIC}} = \frac{1}{R}\sum_{r=1}^{R} \text{WAIC}(r);$$

$$\%\,\text{LOO} = \frac{1}{R}\sum_{r=1}^{R} I(\text{LOO}^{(r)} < \text{LOO}^{\text{log}}); \qquad \%\,\text{WAIC} = \frac{1}{R}\sum_{r=1}^{R} I(\text{WAIC}^{(r)} < \text{WAIC}^{\text{log}});$$

$$s^2_{\text{LOO}} = \frac{1}{R-1}\sum_{r=1}^{R} \left(\text{LOO}^{(r)} - \overline{\text{LOO}}\right)^2; \qquad s^2_{\text{WAIC}} = \frac{1}{R-1}\sum_{r=1}^{R} \left(\text{WAIC}^{(r)} - \overline{\text{WAIC}}\right)^2;$$

where R is the number of simulation replicas (in this case, $R = 100$), and I denotes the indicator function.

Table 3 shows that in all cases, at least one of the proposed models (PDLomax and RPDLomax) performed better than the logistic regression. When λ = 0.25, the RPDLomax model, even with LOO and WAIC values higher than logistic regression, still outperforms it in 58% of cases in both metrics. When λ = 0.5, it can be observed that the PDLomax and RPDLomax models have lower WAIC and LOO values than the logistic model, performing better in over

Table 3. Comparative Bayesian measures of the proposed models and binary logistic regression.

Link	$\overline{LOO}$	s^2_{LOO}	%LOO	$\overline{WAIC}$	s^2WAIC	%WAIC
λ = 0.25						
Logit	5,007.371	71.147	-	5,007.34	71.14	-
DLomax	5,007.349	71.169	49	3	7	49
PDLomax	5,008.029	70.471	44	5,007.32	71.16	45
RPDLomax	5,008.471	70.813	58	0	9	58
λ = 0.5				5,008.00	70.47	
Logit	6,403.237	39.623	-	2 6,403.20	1 39.62	-
DLomax	6,403.268	39.645	40	5,008.44	70.81	41
PDLomax	6,403.053	39.570	66	8 6,403.23	2 1 39.64	66
RPDLomax	6,403.113	39.602	62	9	5	62
λ = 2				6,403.02	39.57	
Logit	6,334.080	38.417	-	5 6,334.05	1 38.41	-
DLomax	6,333.970	38.384	60	6,403.08	39.60	60
PDLomax	6,333.958	38.358	60	1 6 6,333.94	7 2 38.38	60
RPDLomax	6,333.960	38.404	59	2	4	59
λ = 4				6,333.93	38.35	
Logit	4,977.608	65.323	-	1 4,977.57	9 65.32	-
DLomax	4,977.568	65.304	51	6,333.93	38.40	52
PDLomax	4,979.453	65.368	65	2 4,977.54	2 4 65.30	62
RPDLomax	4,978.640	65.189	42	0	5	43
				4,979.43	65.37	
				3	3	
				4,978.61	65.18	
				6	9	

https://doi.org/10.1371/journal.pone.0311246.t003

60% of cases, with the PDLomax model outperforming all (66% of success compared to logistic regression). When λ = 2, the PDLomax and RPDLomax models present lower average LOO and WAIC values than logistic regression and also outperform it most of the time; in this case, the PDLomax model had better results (LOO and WAIC were lower than the logistic model in 60% of cases). Finally, at λ = 4, the PDLomax model, even with WAIC and LOO values higher than the logistic model, has lower LOO in 65% of cases and lower WAIC in 62% of cases.

6 Applications

This section presents two applications that were developed in order to illustrate the performance of the DLomax, PDLomax, and RPDLomax link functions on real data. First, the proposed models were applied to a database with images of diseased trees (Section 6.1), and then the effect of these new link functions was studied on a blood donation database (Section 6.2).

6.1 Wilt dataset

This study considers a dataset comprised of image segments resulting from the pansharpening technique, as described in the work of [25], for the detection of diseased pine and oak trees. This dataset was created because in Japan, beetles that feed on pine and oak trees are responsible for the majority of damage to forested areas, as they transmit diseases to the trees, causing them to wither. Hence, rapid detection, removal, or treatment of newly infected trees is necessary to prevent the beetles from emerging the following year and spreading their diseases. The discoloration of the foliage is a clear sign of infection, so the detection of a diseased tree is usually associated with the detection of a discolored tree. Due to the fact that the number of diseased trees was much smaller compared to the number of healthy trees in the study area,

Table 4. Descriptive measures of the Wilt dataset. SD = standard deviation.

Variable *GLCM_Pan*	**Mean**	**Min.**	**Max.**	**SD**
Mean_G	127.07	81.12	167.94	10.67
Mean_R	209.80	117.20	1848.90	78.68
Mean_NIR	107.74	50.58	1594.58	71.77
SD_Pan	453.70	144.90	1597.30	156.20
https://doi.org/10.1371/journal.pone.0311246.t004	20.64	5.77	62.39	6.76

collecting images of diseased trees was more challenging and time-consuming. Consequently, an imbalanced dataset was constructed.

This dataset was introduced in the paper [25] and is available in the UCI repository [26]. For this study, the validation dataset, which consists of 500 observations, was considered. Below is a brief description of the variables in the dataset (given the complexity of the subject, more information about the variables can be found in [25]):

- *GLCM_Pan*: gray-level co-occurrence matrix (GLCM) mean texture;
- *Mean_G*: mean green value;
- *Mean_R*: mean red value;
- *Mean_NIR*: mean near-infrared (NIR) value;
- *SD_Pan*: standard deviation of the panchromatic (Pan) band;
- *Y*: binary variable indicating whether the tree is diseased (1) or not (0).

Table 4 presents the descriptive statistics for each of the variables. This dataset exhibits a 37.4% success (observed diseased trees, $Y = 1$) rate.

When analyzing the data, it was found that the variables *Mean_G* and *Mean_R* had a correlation of 0.98. We then removed the variable *Mean_G* from the subsequent analysis, since it was more correlated with the other variables. Therefore, the models were adjusted using the standardized variables *GLCM_Pan*, *Mean_R*, *Mean_NIR* and *SD_Pan*. Once again, the models were fitted using the stan package of the R software. In each case, 5,000 iterations, 4 chains, and 2,500 warm-up iterations were considered. In almost all cases (except for the PDLomax distribution), convergence was achieved based on the potential scale reduction statistic ($\hat{R}$b) of [27]. The PDLomax model encountered convergence issues, potentially influenced by several factors, particularly the choice of priors for λ and $\boldsymbol{\beta}$.

It is evident from Table 5 that the RPDLomax model stands out when compared to the other models (including the well-known logistic, probit, cauchit, loglog, and complementary log-log or cloglog models). This model had significantly lower values for DIC, EAIC, EBIC, LOO, and WAIC, demonstrating its superior suitability for the dataset compared to the other models presented. The DLomax model, on the other hand, was the second-best performing model, obtaining the second lowest measures for DIC, EAIC, EBIC, LOO, and WAIC. The cauchit model, although achieving metrics similar to the DLomax model, did not exhibit any measures superior to the RPDLomax and DLomax models.

As the RPDLomax model was chosen, Table 6 provides descriptive statistics of the posterior distribution samples for this model. In this table, it can be observed that the 90% credibility interval for the skewness parameter λ does not encompass the one value, suggesting that the current model cannot be reduced to the base model (DLomax). Additionally, the 90% credibility intervals for the $\beta 1$ and $\beta 4$ parameters encompass the zero value; thus, the *GLCM_Pan* and

Table 5. Comparison metrics of the models fitted to the Wilt dataset.

Model	ρ_d	$\overline{D}$	$D(\bar{\theta})$	DIC	EAIC	EBIC	LOO	WAIC
DLomax	4.658	532.591	527.933	537.249	542.591	563.664	537.874	537.863
PDLomax	-	-	-	-	-	-	-	-
RPDLomax	3.985	428.726	424.741	432.711	440.726	466.013	441.044	441.031
Logistic	5.059	657.361	652.302	662.420	667.361	688.434	688.290	700.992
Probit	5.003	660.494	655.492	665.497	670.494	691.567	675.818	672.110
Cauchit	4.916	543.991	539.076	548.907	553.991	575.064	549.015	549.008
loglog	5.060	635.796	630.735	640.856	645.796	666.869	656.906	667.428
cloglog	4.919	662.544	657.625	667.463	672.544	693.617	670.433	668.231

https://doi.org/10.1371/journal.pone.0311246.t005

SD_Pan covariates are not significant. Despite that, these covariates were kept in the model as they are important for ensuring that the model's residuals meet the assumption of normality. Using the posterior mean as the point estimate for the parameters, the adopted model can be represented by the formula:

$$\hat{\eta}_i = 26.110 + 1.142\,X_{i1} + 137.094\,X_{i2} - 19.340\,X_{i3} + 0.206\,X_{i4};$$

$$\hat{p}_i = \begin{cases} 1 - \dfrac{1}{2(1+\hat{\eta}_i)}, & \hat{\eta}_i \geq 0; \\ \dfrac{1}{2(1+\hat{\eta}_i)}, & \hat{\eta}_i < 0; \end{cases}$$

$$Y_i \overset{ind.}{\sim} \text{Bernoulli}(\hat{p}_i);$$

where $X1$, $X2$, $X3$, and $X4$ are, respectively, the standardized *GLCM_Pan*, *Mean_R*, *Mean_NIR*, and *SD_Pan* variables.

Observing the signs of the parameters, it can be interpreted that:

- As the variables *GLCM_Pan*, *Mean_R* and *SD_Pan* increase, the probability of the tree being diseased also increases;
- As the variable *Mean_NIR* increases, the probability of the tree being diseased decreases;
- The variable *Mean_R* plays a significant role in calculating the probability of the tree being diseased, given the magnitude of its associated parameter. This makes sense because dry trees lose their green color and exhibit more reddish tones;

Table 6. Descriptive measures of the parameters of the RPDLomax model fitted to the Wilt dataset.

Covariate	Parameter	Mean	SD	Median	Percentile 5%	Percentile 95%
Intercept	$\beta 0$	26.110	11.423	24.445	10.738	47.329
GLCM_Pan	$\beta 1$	1.142	1.732	1.026	-1.464	4.141
Mean_R	$\beta 2$	137.094	53.400	128.895	64.941	236.957
Mean_NIR	$\beta 3$	-19.340	7.268	-18.330	-32.818	-9.549
SD_Pan	$\beta 4$	0.206	2.120	0.332	-3.300	3.307
Skewness	λ	0.256	0.038	0.251	0.203	0.327

https://doi.org/10.1371/journal.pone.0311246.t006

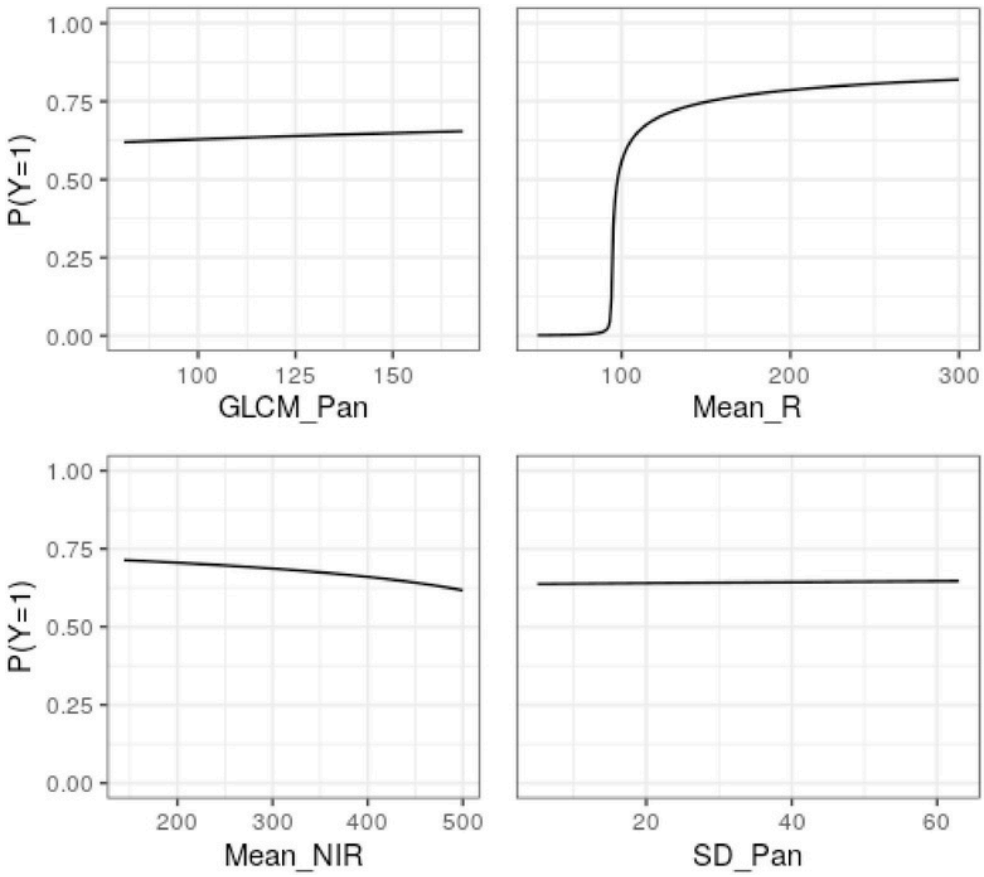

Fig 3. Nonlinear effect of each variable on the probability that a tree is diseased (PðŸ¼1Þ), on average, when the other variables are constant, based on the adjusted RPDLomax model (Wilt dataset).
https://doi.org/10.1371/journal.pone.0311246.g003

In addition to interpreting the signs of the parameters, the impact can be observed of the variation of each variable on the probability of a tree being diseased. In Fig 3, the effect of the variation of each variable is presented, maintaining the others at their average value.

From Fig 3, the significant impact of the variable *Mean_R* can be seen, in which there is a sharp increase in the probability of success (diseased tree) in the range between 90 and 100. When *Mean_R* equals 90, the probability of success is 0.015, while for *Mean_R* equals 100, the probability of success is 0.558. On the other hand, the other variables seem to have a smaller influence on the probability of success as their curves show little variation.

The interpretations provided above, Bayesian Statistics also allows us to interpret the probabilities of success ($Y = 1$) of each of the observations. In Fig 4, it is noted that some observations have a very low probability of success, and their distributions are concentrated in an interval close to 0, while other observations, in turn, are concentrated in points closer to the center, or have a more flattened distribution. Thus, observations can be identified that have a greater degree of uncertainty in their classification.

The residuals of this model are displayed graphically in Fig 5. It can be observed that the residuals behave as expected, showing no signs that they do not follow a normal distribution (left panel). Additionally, it can be noted that the residuals appear to be distributed randomly, with no strong evidence of a trend or changes in variance (right panel).

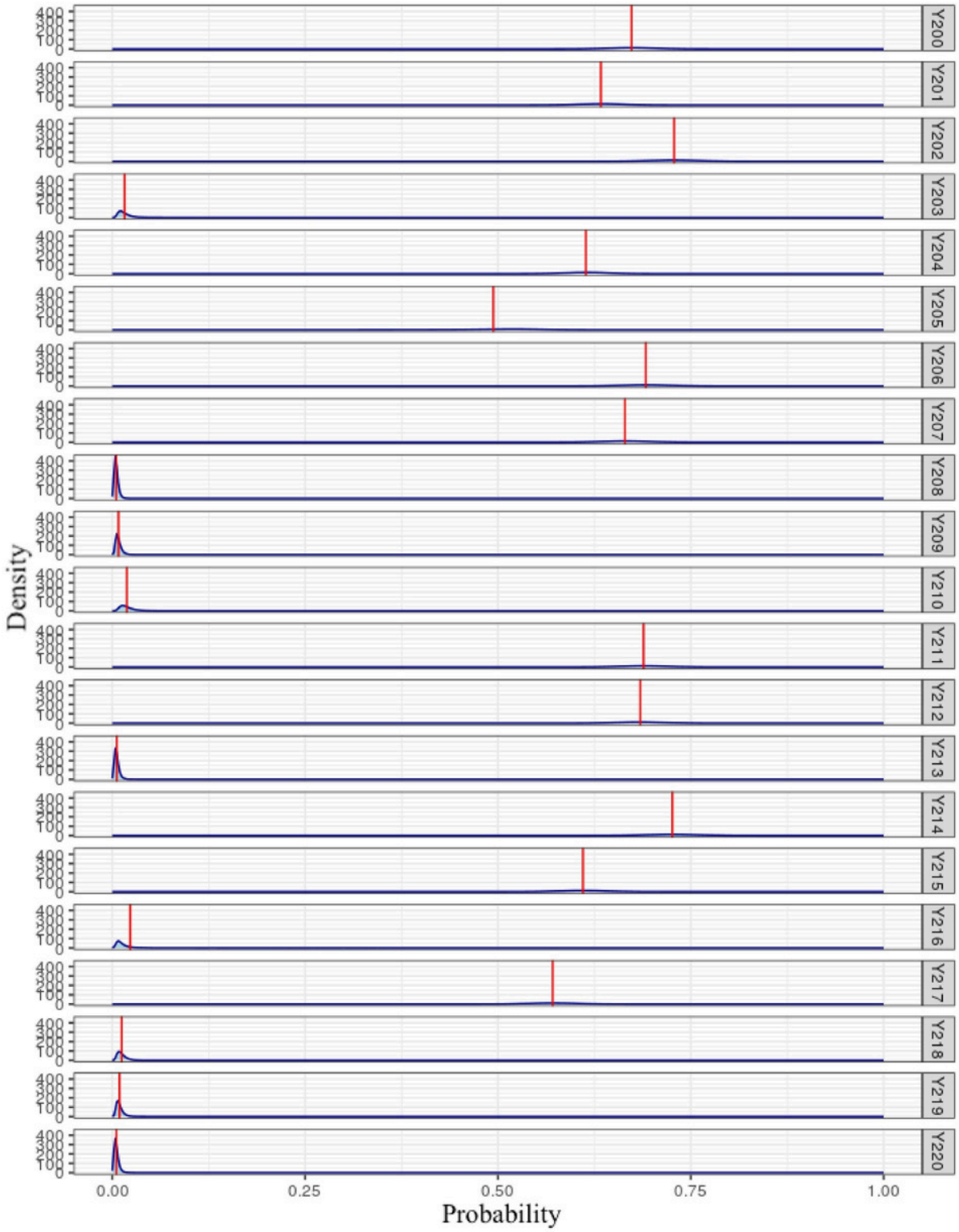

Fig 4. Density of the predictive probabilities estimated by the RPDLomax model adopted in the first application (Wilt dataset), for observations #200 to 220. The red line represents the average of the estimated diseased tree probabilities, $\hat{P}(Y_i = 1 \mid \boldsymbol{x}_i) = \hat{p}_i$.

https://doi.org/10.1371/journal.pone.0311246.g004

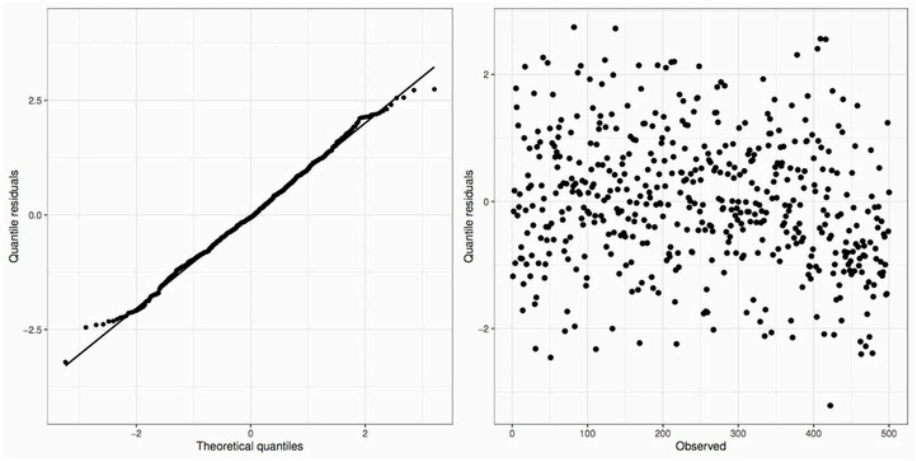

Fig 5. Plots of the quantile residuals of the RPDLomax model fitted to the Wilt dataset.
https://doi.org/10.1371/journal.pone.0311246.g005

In Fig 6, it is clear that the RPDLomax model assigns low mean posterior predictive probabilities to failure and higher mean posterior predictive probabilities to success (left panel). This is in contrast to the logistic regression model, which does not clearly distinguish between the mean predictive probabilities for the two classes (right panel). In Table 7, the median posterior predictive probability for the not diseased tree group, PðYb ¼ 1jY ¼ 0; **X**Þ, based on the RPDLomax model was 1.2%, while for the Logistic model was 34.4%. In comparison, the median predictive probability for the diseased tree, PðYb ¼ 1jY ¼ 1; **X**Þ, based on the RPDLomax model was 66.9%, versus 39.2% for the Logistic model.

6.2 Blood donation dataset

The database used was introduced and analyzed by [28], and is available in the UCI repository [26]. This dataset comprises 748 random samples of blood donor data from the Blood Transfusion Center in Hsin-Chu City, Taiwan, with the following variables:

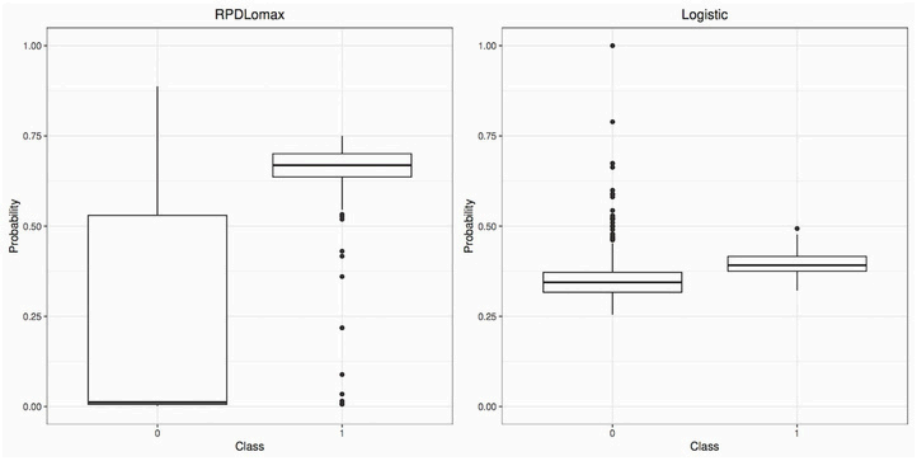

Fig 6. Boxplots of the estimated mean predictive probabilities for each observation, based on the RPDLomax model and the Logistic model for each class (Wilt dataset).
https://doi.org/10.1371/journal.pone.0311246.g006

Table 7. Descriptive statistics (minimum, first quartile, median, mean, third quartile, maximum) of probabilities for the RPDLomax and Logistic models by class (Wilt dataset).

Model	Class	Min.	1stQu.	Median	Mean	3rdQu.	Max.
RPDLomax	Success	0.006	0.636	0.669	0.637	0.701	0.750
	Failure	0.002	0.006	0.012	0.214	0.530	0.887
Logistic	Success	0.322	0.375	0.392	0.395	0.416	0.493
	Failure	0.255	0.317	0.344	0.360	0.372	1.000

https://doi.org/10.1371/journal.pone.0311246.t007

- *Recency*: number of months since the last donation;
- *Frequency*: total number of donations made by the donor;
- *Time*: time, in months, since the first donation;
- *Monetary*: total, in milliliters (ml), of blood donated since the first donation;
- *Y*: binary variable indicating whether he/she donated blood (1—yes, 0—no) in March 2007.

Table 8 shows the descriptive statistics for each of these variables. This dataset exhibits a 23% success (1's) rate.

In the initial analysis, it was observed that the variables *Frequency* and *Monetary* had a correlation of 1 (perfect positive correlation). This occurs because 250 ml of blood is donated with each donation, therefore the *Monetary* variable, which represents the total donated blood, is nothing more than the *Frequency* variable multiplied by 250. Therefore, the decision was made to exclude the *Monetary* variable from the model. In addition to the correlation between these two variables, no other correlations were found that would hinder the model fitting.

Thus, the models were adjusted considering the standardized covariates *Recency*, *Frequency*, and *Time* to predict the variable *Y*. The stan package in the R software was used for parameter estimation. For each distribution, 5,000 iterations, 4 chains, and 2,500 warm-up iterations were considered. Convergence was achieved in all distributions based on the potential scale reduction statistic ($\hat{R}$b) by [27].

In Table 9, it can be observed that the RPDLomax model achieved the lowest values in the DIC, EAIC, LOO, and WAIC metrics, proving to be the model that performed better in most of the metrics. The cauchit model also showed satisfactory performance; however, it only outperformed the RPDLomax model in the EBIC metric. On the other hand, the other models proposed in this work, PDLomax and DLomax, although not performing as well as the RPDLomax model, demonstrated superiority over most traditional models, as they showed lower DIC, EAIC, LOO, and WAIC values than the logistic, probit, loglog, and cloglog models.

Considering the model comparison criteria and predictive evaluation, the RPDLomax model was chosen. Therefore, Table 10 presents the descriptive measures of the posterior samples of this model's parameters. Note that all parameters (coefficients) β's are significant, as none of the 90% credibility intervals for them include the zero value. Additionally, the 90%

Table 8. Descriptive measures of the Blood Donation dataset.

Variable	Mean	Min.	Max.	SD
Recency	9.5	0	74	8.1
Frequency	5.5	1	50	5.8
Time	34.3	2	98	24.4
Monetary	1,378.7	250	12,500	1,459.8

https://doi.org/10.1371/journal.pone.0311246.t008

Table 9. Comparison metrics of the models fitted to the Blood Donation dataset.

Model	ρ^d	$\overline{D}$	$D(\bar{\theta})$	DIC	EAIC	EBIC	LOO	WAIC
DLomax	4.179	708.800	704.620	712.979	716.800	735.269	712.960	712.953
PDLomax	3.545	707.867	704.322	711.412	717.867	740.954	713.474	713.474
RPDLomax	4.963	706.123	701.159	711.086	716.123	739.210	711.542	711.532
Logistic	3.976	711.861	707.885	715.837	719.861	738.330	716.279	716.271
Probit	3.962	713.551	709.589	717.513	721.551	740.020	718.261	718.201
Cauchit	4.111	708.511	704.400	712.622	716.511	734.981	712.855	712.852
loglog	3.995	715.224	711.229	719.219	723.224	741.694	719.814	719.793
cloglog	4.016	715.178	711.162	719.193	723.178	741.647	721.906	720.949

https://doi.org/10.1371/journal.pone.0311246.t009

credibility interval for λ does not include the one value, indicating that this parameter is important for the fit, and the current model cannot be reduced to the base model (DLomax).

It was chosen to use the posterior mean as the point estimate for the parameters. Thus, the adopted model can be represented as follows:

$$\begin{aligned} \hat{\beta} &= (-1.410,\ -1.381,\ 1.422,\ -0.889)^\top; \\ \hat{p}_i &= \begin{cases} 1 - \left(1 + \dfrac{1}{2(1 + z_i)}\right)^{-0.679}, & z_i > 0; \\ \dfrac{1}{2(1 + z_i)}, & z_i \le 0; \end{cases} \\ Y_i &\overset{ind.}{\sim} \text{Bernoulli}(\hat{p}_i), \end{aligned}$$

where $X1$, $X2$, and $X3$ are, respectively, the standardized *Recency*, *Frequency*, and *Time* variables.

Observing the signs of the parameters, it can be interpreted that:

- As the number of months since the last donation (*Recency*) increases, the probability of the donor donating blood on the specified date decreases;
- When the number of donations made by the donor (*Frequency*) increases, the probability of he/she donating in the period also increases;
- As the time in months since the first donation (*Time*) increases, the probability of donation in March 2007 decreases.

These interpretations make sense, as donors who made their first donation a long time ago, donated infrequently, and have not donated blood for a long time; they represent a profile of

Table 10. Descriptive measures of the parameters of the RPDLomax fitted to the Blood Donation dataset.

Covariate	Parameter	Mean	SD	Median	Percentile 5%	Percentile 95%
Intercept	$\beta 0$	-1.410	0.479	-1.352	-2.281	-0.747
Recency	$\beta 1$	-1.381	0.356	-1.343	-2.021	-0.872
Frequency	$\beta 2$	1.422	0.349	1.396	0.902	2.029
Time	$\beta 3$	-0.889	0.246	-0.870	-1.319	-0.520
Skewness	λ	0.679	0.148	0.656	0.476	0.955

https://doi.org/10.1371/journal.pone.0311246.t010

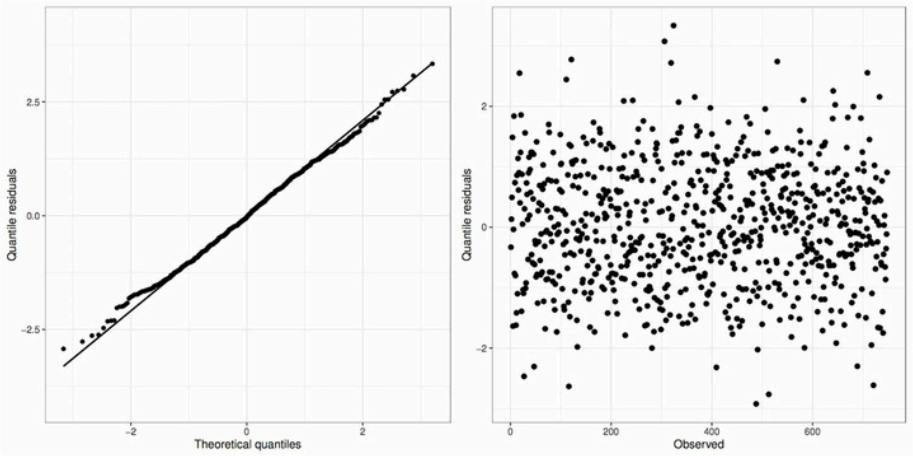

Fig 7. Plots of the quantile residuals of the RPDLomax model fitted to the Blood Donation dataset.
https://doi.org/10.1371/journal.pone.0311246.g007

sporadic donors. On the other hand, donors with a high number of donations represent the profile of regular donors.

In Fig 7, the quantile residuals exhibit the expected behavior, showing no indications of deviating from a normal distribution (left panel). Furthermore, the residuals appear to be distributed randomly, lacking compelling evidence of any discernible trend or variations in variance (right panel).

Turning our attention to Fig 8, one can discern a shift in the mean predictive distribution of success (1) versus the failure (0) mean predictive probabilities when transitioning from the logistic model (right panel) to the RPDLomax model (left panel). In the RPDLomax model, the medians of these predictives are more distinctly (separable), resulting in a heightened distinction between the probabilities associated with each class.

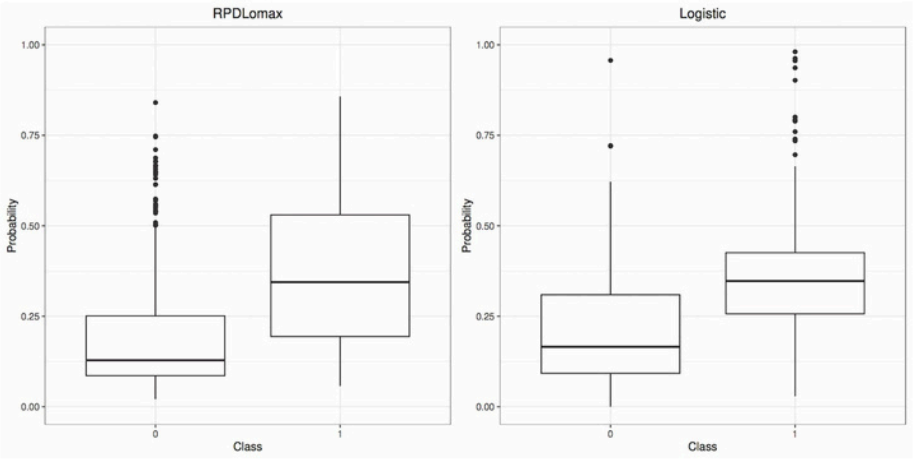

Fig 8. Boxplots of the estimated mean predictive probabilities for each observation, based on the RPDLomax model and the Logistic model for each class (Blood Donation dataset).
https://doi.org/10.1371/journal.pone.0311246.g008

7 Discussion

In this study, the RPDLomax model was identified as the optimal choice due to its superior selection metrics (LOO, WAIC, and DIC) and satisfactory fit, with residuals conforming to necessary assumptions. This finding is consistent with literature indicating that models with asymmetric links often outperform those with symmetric links in real-world data applications [5, 7].

The reverse transformation exhibited outstanding performance in predicting real data, supported by multiple studies [2, 4, 14, 16]. In most cases, the reverse power transformation outperformed the power transformation, except in [10], where the latter achieved better results. Therefore, while both transformations yield similar outcomes, the reverse transformation generally offers a slight performance advantage.

Furthermore, the Bayesian approach for parameter fitting and estimation was explored. However, it is noted that many studies in this field do not fully utilize all the available information from this approach, such as the probability distribution of success for each sample, as illustrated in Fig 4.

7.1 Implication

The significance and novelty of this work lie in the introduction of new asymmetric classification functions that outperform traditional link functions such as logit, cauchit, probit, loglog, and cloglog. These novel functions offer a robust option for classifying binary imbalanced data and can be integrated into the R workflow. The simulation studies conducted in this paper demonstrated that these models can surpass logistic regression in various scenarios. Moreover, the simulations indicated that the parameters of these models can be recovered with low bias and reduced variance, particularly in larger samples.

The models were implemented using a Bayesian approach, which provides several advantages. Bayesian methods facilitate the incorporation of prior knowledge into the analysis through the use of prior distributions. They also offer a natural mechanism for quantifying uncertainty in parameter estimates and predictions. Additionally, Bayesian methods can exhibit greater robustness than frequentist methods when dealing with small sample sizes, owing to the ability to utilize informative prior distributions.

7.2 Limitations

Some limitations of this study should be considered. First, the number of replications and iterations in the simulations was limited. Increasing these in future work could improve the reliability of the results. The PDLomax model demonstrated promising performance with the second database but failed to converge in the first application. This suggests that the choice of priors impacts model performance. Future studies should investigate the use of less restrictive priors to improve convergence.

Moreover, the models were tested only on small datasets with a limited number of numerical features. The performance of these models on larger datasets with more features, including categorical variables, remains uncertain. Additionally, while the Bayesian approach offers a rich interpretative framework, it is computationally intensive, particularly for large-scale datasets. Future research should aim to optimize computational efficiency, or alternatively, consider a frequentist approach for handling big data.

7.3 Future work

In future studies, these presented models can be explored for new applications, especially in larger datasets. Additionally, the asymmetric link functions can be used as activation functions in neural networks, as [29] demonstrated that asymmetric activation functions can improve time series prediction with neural networks in their work. Furthermore, a performance comparison (also using, for instance, the evaluation metrics obtained from the confusion matrix, as well as an appropriate validation scheme) of the presented models with other asymmetric links (e.g., the ones based on other power and reverse power distributions) can be conducted. Moreover, implementing these models in R packages would enhance their accessibility and usability within the research community. Additionally, exploring alternative priors for λ could provide further refinement to the models.

8 Conclusion

In this work, new approaches to modeling imbalanced data were presented by introducing new link functions for binary regression. These novel link functions were created by applying the transformation proposed by [4] to the double Lomax distribution (DLomax) [17], in order to generate the power double Lomax (PDLomax) and reverse power double Lomax (RPDLomax) distributions. Despite their various applications, the Lomax distribution and its extensions had not been explored in the context of binary regression until now, making this work a novel and significant contribution to the literature. Additionally, the evidence pointed out the advantages of using the Bayesian classification approach by associating each event with a predictive posterior probability. Then, the Lomax Bayesian learning overcame the Logistic classification for differentiating binary data in imbalanced tasks.

Two simulation studies were carried out to assess the models' ability to recover parameters and their fit quality under misspecification scenarios. The first study indicated that the proposed models and Bayesian estimation procedure are efficient at parameter recovery and exhibit reduced estimation bias as the sample size increases. In the second study, it was observed that the proposed models outperformed logistic regression in terms of model fit quality as evaluated by LOO and WAIC metrics, both in scenarios with moderate and severe asymmetry.

The proposed models were also applied to two imbalanced real datasets. The first database pertains to the classification of potential blood donors, and the second database involves the classification of image segments to identify diseased trees. In both databases, the RPDLomax model outperformed conventional link functions such as logit, probit, cauchit, loglog, and cloglog, showing the lowest fit metrics (WAIC, LOO, DIC, EAIC, and EBIC).

Finally, it is worth mentioning that all codes developed for this work can be found on the first author's GitHub: https://github.com/leticiaferreiramurca/Msc.

Author Contributions

Conceptualization: Paulo H. Ferreira, Francisco Louzada.

Formal analysis: Letí´cia F. M. Reis.

Investigation: Letí´cia F. M. Reis, Diego C. Nascimento, Paulo H. Ferreira.

Methodology: Letí´cia F. M. Reis, Diego C. Nascimento, Paulo H. Ferreira, Francisco Louzada.

Project administration: Francisco Louzada.

Software: Letí´cia F. M. Reis.

Supervision: Paulo H. Ferreira, Francisco Louzada.

Validation: Letí´ciaF.M.Reis,DiegoC.Nascimento,PauloH.Ferreira,FranciscoLouzada.

Visualization: Letí´cia F. M. Reis.

Writing–originaldraft: Letí´ciaF.M.Reis.

Writing – review & editing: Diego C. Nascimento, Paulo H. Ferreira, Francisco Louzada.

References

1. Haibo H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
2. Alves JSB, Bazá´ n JL, Arellano-Valle RB. Flexible cloglog links for binomial regression models as an alternative for imbalanced medical data. Biometrical Journal. 2023; 65(3):2100325. https://doi.org/10.1002/bimj.202100325 PMID: 36529694
3. Czado C, Santner TJ. The effect of link misspecification on binary regression inference. Journal of Statistical Planning and Inference. 1992; 33(2):213–231. https://doi.org/10.1016/0378-3758(92)90069-5
4. Bazá´nJL,RomeoJS,RodriguesJ.Bayesianskew-probitregressionforbinaryresponsedata.Brazilian Journal of Probability and Statistics. 2014; 28(4):467–482.
5. Yin S, Dey DK, Valdez EA, Gan G, Vadiveloo J. Skewed link regression models for imbalanced binary response with applications to life insurance. arXiv. 2020.
6. NaranjoL,Pé´rezC,Martí´nJ,Calle-AlonsoF.Anewasymmetriclink-basedbinaryregressionmodelto detect parkinson's disease by using replicated voice recordings. 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy. 2018;1182–1186.
7. Calabrese R, Osmetti SA. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. Journal of Applied Statistics. 2013; 40(6):1172–1188. https://doi.org/10.1080/02664763.2013.784894
8. Golet I. Symmetric and asymmetric binary choice models for corporate bankruptcy. Procedia—Social and Behavioral Sciences. 2014; 124:282–291. https://doi.org/10.1016/j.sbspro.2014.02.487
9. Prasetyo RB, Kuswanto H, Iriawan N, Ulama BSS. Binomial regression models with a flexible generalized logit link function. Symmetry. 2020; 12(2):221. https://doi.org/10.3390/sym12020221
10. Huayanay AC, Bazá´n JL, Cancho VG, Dey DK. Performance of asymmetric links and correction methods for imbalanced data in binary regression. Journal of Statistical Computation and Simulation. 2019; 89(9):1694–1714. https://doi.org/10.1080/00949655.2019.1593984
11. Stukel TA. Generalized logistic models. Journal of the American Statistical Association. 1988; 83 (402):426–431. https://doi.org/10.1080/01621459.1988.10478613
12. Taylor JMG. The cost of generalizing logistic regression. Journal of the American Statistical Association. 1988; 83(404):1078–1083. https://doi.org/10.1080/01621459.1988.10478704
13. Huayanay AC. Modelos de regressão para resposta biná´ ria na presenç¸a de dados desbalanceados. Master's thesis, UFSCAR-USP, São Carlos. 2019.
14. Lemonte AJ, Bazá´n JL. New links for binary regression: an application to coca cultivation in Peru. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research. 2018; 27 (3):597–617. https://doi.org/10.1007/s11749-017-0563-1
15. Anyosa SAC. Regressão biná´ ria usando ligaç¸ões potência e reversa de potência. Master's thesis, UFSCAR-USP, São Carlos. 2017.
16. Bazá´nJ,Torres-Avilé´sF,SuzukiA,LouzadaF.Powerandreversalpowerlinksforbinaryregressions: An application for motor insurance policyholders. Applied Stochastic Models in Business and Industry. 2017; 33(1):22–34. https://doi.org/10.1002/asmb.2215
17. Bindu P, Sangita K. Double lomax distribution and its applications. Statistica. 2015; 75(3):331–342.
18. Arnold BC. Pareto Distributions. Fairland, MD: International Cooperative Publishing House; 1983.
19. Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions. 2nd edn. New York, NY: John Wiley & Sons; 1994.
20. Stan Development Team. RStan: the R interface to Stan. R package version 2.21.7; 2022. Available from: https://mc-stan.org/.
21. Homan DM, Gelman A. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014; 15(1):1593–1623.

22. Vehtari A, Gelman A, Gabry J. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. 2016; 27(5):1413–1432. https://doi.org/10.1007/s11222-016-9696-4

23. Dunn PK, Smyth GK. Generalized Linear Models With Examples in R. New York, NY: Springer; 2018.

24. Yong L. LOO and WAIC as Model Selection Methods for Polytomous Items. arXiv. 2018.

25. Johnson BA, Tateishi R, Hoan NT. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. International Journal of Remote Sensing. 2013; 34(20):6969–6982. https://doi.org/10.1080/01431161.2013.810825

26. Dua D, Graff C. UCI Machine Learning Repository; 2027. http://archive.ics.uci.edu/ml.

27. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992; 7(4):457–472. https://doi.org/10.1214/ss/1177011136

28. Yeh I, Yang K, Ting T. Knowledge discovery on rfm model using Bernoulli sequence. Expert Systems with Applications. 2009; 36(3, Part 2):5866–5871. https://doi.org/10.1016/j.eswa.2008.07.018

29. Gomes GSS, Ludermir TB. Optimization of the weights and asymmetric activation function family of neural network for time series forecasting. Expert Systems with Applications. 2013; 40(16):6438–6446. https://doi.org/10.1016/j.eswa.2013.05.053