MODELOS PREDICTIVOS EN EL ANÁLISIS DE RIESGO CREDITICIO

https://doi.org/10.22533/at.ed.876122508041

Data de aceite: 23/04/2025

Carlos Alberto Peña Miranda
https://orcid.org/0000-0002-4339-4615

Jesus Adalberto Zelaya Contreras
https://orcid.org/0009-0003-4767-6366

Elizabeth Cosi Cruz
https://orcid.org/0000-0002-0255-7705

RESUMEN: El artículo explica la importancia de los modelos predictivos en el análisis de riesgo crediticio, destacando técnicas como regresión logística, árboles de decisión y Random Forest. Se ha identificado que estos modelos, aunque útiles, presentan limitaciones como la linealidad en regresión logística, sobreajuste en árboles de decisión y complejidad computacional en Random Forest. Además, su rendimiento se ve afectado por datos desbalanceados y variables categóricas. Como conclusión, se sugiere el uso de modelos avanzados como CatBoost, que superan estas limitaciones al manejar variables categóricas de manera eficiente, reducir el sobreajuste y mejorar la interpretabilidad, ofreciendo así una solución más robusta para la predicción de riesgo crediticio.

PALABRAS CLAVE: Modelos predictivos, Riesgo crediticio, Aprendizaje automático

PREDICTIVE MODELS IN CREDIT RISK ANALYSIS

ABSTRACT: The article discusses the importance of predictive models in credit risk analysis, focusing on techniques such as logistic regression, decision trees, and Random Forest. It identifies key limitations, including linearity assumptions in logistic regression, overfitting in decision trees, and computational complexity in Random Forest. Additionally, these models struggle with imbalanced data and categorical variables. The conclusion recommends advanced approaches like CatBoost, which efficiently handles categorical features, reduces overfitting, and improves interpretability, offering a more robust solution for credit risk prediction.

KEYWORDS: Predictive models, Credit risk, Machine learning

INTRODUCCIÓN

En este artículo, se presentan los modelos predictivos como herramientas fundamentales en la gestión del riesgo crediticio, analizando su capacidad para transformar datos históricos en insights accionables que permiten anticipar comportamientos financieros.

En particular, se profundiza en técnicas ampliamente utilizadas como la regresión logística para clasificación binaria, los árboles de decisión para segmentación de reglas claras, y los ensembles como Random Forest que combinan múltiples modelos para mejorar la precisión predictiva. No obstante, cada uno de estos modelos presenta sus propias limitaciones, especialmente en el manejo de variables categóricas, la sensibilidad a datos desbalanceados y la capacidad de modelar relaciones no lineales.

El objetivo de este artículo es proporcionar una comprensión clara de estos modelos clásicos, analizando sus ventajas y desventajas, y justificando la necesidad de enfoques más avanzados, como el algoritmo CatBoost.

A lo largo del artículo, se explorarán los principios matemáticos que sustentan estos modelos, sus aplicaciones prácticas y los criterios clave para evaluar su desempeño en escenarios del mundo real.

DEFINICIÓN DE MODELOS PREDICTIVOS

Un *modelo predictivo* es una función matemática $f : X \to Y$ que mapea un conjunto de variables de entrada X (también llamadas características o *features*) a un conjunto de salidas Y (también llamadas etiquetas o *labels*).

Formalmente, dado un conjunto de datos $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{n}$ donde $x_i \in X$ es un vector de características y $y_i \in Y$ es la etiqueta correspondiente, podemos considerar un caso en que el vector de características x_i este compuesto por tres valores que representen atributos del cliente, como la edad, el promedio de saldo en cuenta bancaria y la cantidad de transacciones en el último mes. Por su parte, la etiqueta y_i puede tomar valores como *aprobado* o *rechazado*.

Ejemplo 1. En un modelo de aprobación de créditos, X podría incluir información como el historial de pagos de un cliente, sus ingresos mensuales y su nivel de endeudamiento. La salida Y sería la decisión del modelo, como *aprobado* o *rechazado*. En este caso, la función f aprende a predecir si un cliente es apto para recibir un crédito en función de sus características financieras.

El objetivo de un modelo predictivo es aprender una función f que minimice una función de pérdida $L(f(x_i), y_i)$. La función de pérdida L cuantifica la discrepancia entre las predicciones del modelo $f(x_i)$ y las etiquetas reales y_i. Esto se debe a que, mientras menor sea la diferencia entre lo predicho y lo real, mejor será el desempeño del modelo. En otras palabras, la función de pérdida nos permite cuantificar qué tan bien está funcionando nuestro modelo entrenado. Matemáticamente, esto se expresa como:

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} L(f(x_i), y_i)$$

donde $\mathcal{F}$ es el espacio de funciones posibles que el modelo puede aprender.

El operador argmin se utiliza para indicar que buscamos la función f* dentro del espacio de funciones $\mathcal{F}$ que minimiza la suma de la función de pérdida L en el conjunto de datos. En otras palabras, no solo nos interesa el valor mínimo de la suma de pérdidas, sino la función f* que lo alcanza.

Ejemplo 2. En un problema de regresión lineal, si consideramos una función lineal del tipo:

$$f(x) = w_0 + w_1 x$$

donde w_0 y w_1 son parámetros a ajustar, y utilizamos la función de pérdida de error cuadrático medio:

$$L(f(x_i), y_i) = (f(x_i) - y_i)^2,$$

entonces el proceso de entrenamiento del modelo consiste en encontrar los valores óptimos de w_0 y w_1 que minimicen la suma de los errores cuadráticos:

$$(w_0^*, w_1^*) = \underset{w_0, w_1}{\operatorname{argmin}} \sum_{i=1}^{n} (w_0 + w_1 x_i - y_i)^2$$

Aquí, argmin nos proporciona los valores w_0^* y w_1^* que hacen que la función de pérdida sea mínima, es decir, los parámetros que generan la mejor función predictiva dentro del conjunto de funciones lineales posibles.

IMPORTANCIA DE LOS MODELOS PREDICTIVOS EN LA TOMA DE DECISIONES

Los modelos predictivos son fundamentales en la toma de decisiones, ya que permiten anticipar eventos futuros basados en datos históricos. En el contexto del riesgo crediticio, un modelo predictivo puede estimar la probabilidad de que un cliente incumpla con sus obligaciones de pago, lo que permite a las instituciones financieras optimizar la asignación de crédito y minimizar las pérdidas asociadas al incumplimiento.

Según Thomas, Crook y Edelman (2017), en su obra *Credit Scoring and Its Applications*, el uso de modelos predictivos en el análisis de crédito ha demostrado ser una de las herramientas más eficaces para reducir la morosidad y mejorar la rentabilidad de las entidades financieras. La precisión de estos modelos es crucial, ya que decisiones basadas en predicciones erróneas pueden tener consecuencias financieras significativas, afectando tanto a las instituciones como a los clientes.

CLASIFICACIÓN DE MODELOS PREDICTIVOS

Los modelos predictivos se clasifican en dos grandes categorías: *supervisados* y *no supervisados*. Además, dentro de los modelos supervisados, se distinguen dos tipos principales: regresión y clasificación.

Modelos supervisados: En estos modelos, el conjunto de datos $\mathfrak{D}$ incluye tanto las características x_i como las etiquetas y_i. El objetivo es aprender una función f que generalice bien a nuevos datos.

Ejemplo 3. En un sistema de clasificación de correos electrónicos, cada correo puede representarse mediante un conjunto de características x_i, como la cantidad de palabras en el asunto, la presencia de ciertas palabras clave y el remitente del mensaje. La etiqueta y_i indicaría si el correo *spam* o *no spam*. Un modelo supervisado entrenado con datos etiquetados puede aprender a clasificar correctamente nuevos correos basándose en patrones identificados en el conjunto de entrenamiento.

Dentro de los modelos supervisados, se encuentran:

a. **Regresión:** En problemas de regresión, la variable objetivo y_i es continua. El objetivo es predecir un valor numérico, como el precio de una vivienda o el rendimiento de un activo financiero. Un ejemplo clásico es la regresión lineal, donde $f(x_i) = w^T x_i + b$, siendo w un vector de pesos y b un término de sesgo.

Ejemplo 4. Supongamos que queremos predecir el precio de una vivienda en función de su área (en metros cuadrados). Si el modelo aprendido es:

$$f(x) = 5000x + 20000$$

donde x es el área de la vivienda en metros cuadrados, el coeficiente 5000 representa el precio por metro cuadrado y el término independiente 20000 es un ajuste base. Si una vivienda tiene un área de x = 100 m², el precio estimado sería:

$$f(100) = 5000(100) + 20000 = 520\,000$$

b. **Clasificación:** En problemas de clasificación, la variable objetivo y_i es discreta. El objetivo es asignar una etiqueta a cada instancia, como predecir si un cliente incurrirá en default (clasificación binaria) o clasificar un correo electrónico como spam o no spam. Un ejemplo común es la regresión logística, donde $f(x_i) = \sigma(w^T x_i + b)$, con σ siendo la función sigmoide.

Ejemplo 5. Supongamos que queremos predecir si un cliente aprobará un crédito (y = 1) o no (y = 0) en función de su ingreso mensual (x, en miles de dólares). Si el modelo aprendido es:

$$f(x) = \sigma(-3 + 2x),$$

donde

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

es la función sigmoide. Si un cliente tiene un ingreso de x = 3mil dólares, la probabilidad de que se le apruebe el crédito es:

$$f(3) = \frac{1}{1 + e^{-(-3 + 2(3))}} = \frac{1}{1 + e^{-3}} \approx 0.95$$

Esto indica que el cliente tiene una alta probabilidad de ser aprobado para el crédito.

c. Modelos no supervisados: En estos modelos, el conjunto de datos solo incluye las características x_i, sin etiquetas y_i. El objetivo es descubrir patrones o estructuras subyacentes en los datos.

Por ejemplo, en la segmentación de clientes de un banco, cada cliente puede describirse con características como el monto promedio de sus transacciones, la frecuencia de uso de la tarjeta de crédito y su historial de depósitos. Sin embargo, no se cuenta con etiquetas que indiquen a qué tipo de cliente pertenece cada uno. Un modelo no supervisado, como el algoritmo de k-means, puede agrupar a los clientes en diferentes segmentos según sus comportamientos financieros, permitiendo a la institución ofrecer estrategias personalizadas para cada grupo.

Ejemplos comunes incluyen:

- **Clustering:** Agrupa los datos en clusters basados en similitudes, como el algoritmo K-means.

Ejemplo 6. Supongamos que un banco quiere segmentar a sus clientes en tres grupos según sus hábitos de gasto. Se tienen los siguientes datos de clientes representados por dos características: ingreso mensual ($1000 a $5000) y gasto promedio en tarjeta de crédito ($200 a $2000).

Aplicando el algoritmo K-means con k = 3, se obtienen tres grupos:

- Cluster 1: Clientes con ingresos bajos y gastos moderados.
- Cluster 2: Clientes con ingresos altos y gastos elevados.
- Cluster 3: Clientes con ingresos medios y gastos balanceados.

Con esta información, el banco puede ofrecer estrategias personalizadas para cada grupo, como promociones o límites de crédito adecuados.

- **Reducción de dimensionalidad:** Reduce el número de características mientras preserva la estructura de los datos, como el Análisis de Componentes Principales (PCA).

Ejemplo 7. Supongamos que tenemos un conjunto de datos con 10 variables financieras para cada cliente, como ingresos, deudas, historial de pagos, uso de tarjetas, etc. Aplicando PCA, descubrimos que solo dos combinaciones lineales de estas variables explican el 90% de la variabilidad de los datos.

En lugar de trabajar con las 10 variables originales, podemos representar a cada cliente con solo dos valores (las dos primeras *componentes principales*), reduciendo la complejidad del modelo y facilitando la visualización de los datos en un gráfico bidimensional.

EJEMPLOS DE MODELOS PREDICTIVOS

Entre los ejemplos de modelos clásicos utilizados en la predicción de riesgo crediticio se encuentran:

a. **Regresión logística.** La *regresión logística* es un modelo de clasificación binaria que estima la probabilidad de que una instancia pertenezca a una clase específica. Dado un vector de características x_i, el modelo predice la probabilidad p_i de que la etiqueta y_i sea 1 (default) mediante la función logística:

$$p_i = \sigma(w^T x_i + b) = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

donde w es un vector de pesos, b es el término de sesgo, y σ es la función sigmoide. La función de pérdida utilizada en la regresión logística es la *pérdida logarítmica* (log loss), definida como:

$$L(y_i, p_i) = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)].$$

Para esta fórmula, los logaritmos hacen referencia al logaritmo neperiano en base e. El objetivo es encontrar los parámetros w y b que minimicen la pérdida logarítmica sobre el conjunto de entrenamiento.

Ejemplo 8. Supongamos que un banco quiere predecir si un cliente entrará en *default*, lo que significa que no pudo pagar su deuda o incumplió con el pago de un crédito. Se define y = 1 si esto ocurrió, e y = 0 si no ocurrió, en función de su ingreso mensual (x, en miles de dólares). Si el modelo aprendido es:

$$p_i = \frac{1}{1 + e^{-(-2 + 3x_i)}}$$

y un cliente tiene un ingreso de xi = 2.5 mil dólares, la probabilidad de *default* es:

$$p_i = \frac{1}{1 + e^{-(-2 + 3(2.5))}} = \frac{1}{1 + e^{-5.5}} \approx 0.995$$

Esto indica que el modelo predice una probabilidad del 99.5% de que el cliente entre en *default*.

Ahora, si el valor real de y_i es 1 (es decir, el cliente efectivamente entró en default), la pérdida logarítmica correspondiente es:

$$L(1, 0.995) = -[1 \log(0.995) + (1 - 1) \log(1 - 0.995)].$$

Dado que el segundo término se anula, queda:

$$L(1, 0.995) = -\log(0.995) \approx 0.005$$

Como la probabilidad predicha es cercana a 1 y el valor real también es 1, la pérdida es pequeña, lo que indica que el modelo hizo una buena predicción.

b. Árboles de decisión. Un *árbol de decisión* es un modelo no paramétrico que divide recursivamente el espacio de características en regiones más simples, basándose en reglas de decisión. Un modelo es no paramétrico porque no asume una forma matemática fija para los datos (como una ecuación lineal). En su lugar, aprende patrones directamente de los datos sin hacer suposiciones previas sobre cómo se relacionan las variables. El espacio de características es el conjunto de todas las posibles combinaciones de valores que pueden tomar las variables de entrada. Por ejemplo, si se evalúan clientes para un préstamo en función de su ingreso y su historial de pagos, el espacio de características incluye todas las combinaciones posibles de ingresos e historiales.

El *árbol de decisión* está formado por nodos, donde cada nodo representa una prueba sobre una característica de los datos. Es decir, en cada nodo se formula una pregunta, como: ¿El ingreso es mayor a 3000 dólares?}. También contiene ramas, que representan los posibles resultados de la prueba. Siguiendo el ejemplo anterior, si el nodo pregunta ¿El ingreso es mayor a 3000 dólares?, habrá dos ramas: una para *Sí* y otra para *No*. Finalmente, el árbol tiene hojas, que representan una etiqueta de clase o una decisión final después de seguir todas las reglas de decisión. En el caso del modelo de préstamos, una hoja puede indicar si el préstamo es *aprobado* o *rechazado.*

Formalmente, un árbol de decisión puede expresarse como una $f(x_i)$ que asigna una etiqueta y_i basada en una serie de condiciones:

$$f(x_i) = \sum_{m=1}^{M} c_m \mathbb{I}(x_i \in R_m)$$

donde R_m son regiones disjuntas del espacio de características, c_m es la etiqueta asociada a la región R_m.

$\mathbb{I}$ es la *función indicadora*, la cual toma el valor de 1 si se cumple una determinada condición y 0 en caso contrario. Matemáticamente, se define como:

$$\mathbb{I}(x \in R) = \begin{cases} 1, & \text{si } x \in R \\ 0, & \text{si } x \notin R \end{cases}$$

En el contexto de los árboles de decisión, esta función se usa para indicar si una instancia x_i pertenece a una región específica R_m, lo que determina la asignación de una etiqueta correspondiente.

Ejemplo 9. Supongamos que un banco quiere decidir si aprueba (y = 1) o rechaza (y = 0) un préstamo en función del ingreso mensual (x_1, en miles de dólares) y el historial de pagos (x_2, donde 1 significa buen historial y 0 significa mal historial). Definimos tres regiones R_1, R_2 y R_3, con etiquetas $c_1 = 0$ (rechazo), $c_2 = 1$ (aprobado) y $c_3 = 0$ (rechazo), respectivamente.

Las reglas del árbol dividen el espacio de características en las siguientes regiones:

$R_1 = \{(x_1, x_2): x_1 < 3\}$, $R_2 = \{(x_1, x_2): x_1\ 3,\ x_2 = 1\}$, $R_3 = \{(x_1, x_2): x_1\ 3,\ x_2 = 0\}$.

El árbol de decisión se expresa como:

$$f(x_i) = c_1 \cdot \mathbb{I}(x_i \in R_1) + c_2 \cdot \mathbb{I}(x_i \in R_2) + c_3 \cdot \mathbb{I}(x_i \in R_3)$$

Ahora, evaluemos un cliente con un ingreso mensual de $x_1 = 4$ mil dólares y un mal historial de pagos ($x_2 = 0$):

$$f(4,0) = 0 \cdot \mathbb{I}((4,0) \in R_1) + 1 \cdot \mathbb{I}((4,0) \in R_2) + 0 \cdot \mathbb{I}(4,0 \in R_3)$$

Dado que (4,0) pertenece a R_3, donde la etiqueta es $c_3 = 0$, se obtiene:

$$f(4,0) = 0.$$

Por lo tanto, el préstamo es *rechazado*.

c. Random forest. El *random forest* es un método de ensamblaje que combina múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste.

Un método de ensamblaje es una técnica que combina las predicciones de varios modelos base (en este caso, árboles de decisión) para obtener un modelo más robusto y preciso. Al fusionar múltiples modelos, se reduce la variabilidad y se mejora la generalización en nuevos datos.

El sobreajuste ocurre cuando un modelo aprende demasiado bien los detalles y ruido del conjunto de entrenamiento, lo que le impide generalizar correctamente a nuevos datos. En el caso de los árboles de decisión, esto sucede cuando el árbol es demasiado profundo y se ajusta excesivamente a las observaciones de entrenamiento. Random forest ayuda a mitigar este problema al promediar múltiples árboles, reduciendo la dependencia de un solo modelo y mejorando la estabilidad de las predicciones.

Cada árbol en el bosque se entrena con una muestra aleatoria del conjunto de datos (muestreo con reemplazo) y un subconjunto aleatorio de características. La predicción final es el promedio (en regresión) o la moda (en clasificación) de las predicciones de todos los árboles. Formalmente, la predicción de un random forest puede expresarse como:

$$f(x_i) = \frac{1}{T}\sum_{i=1}^{T} f_t(x_i),$$

donde T es el número de árboles y f_t es la predicción del árbol t-ésimo. random forest es robusto frente al sobreajuste y puede manejar grandes conjuntos de datos, pero su complejidad computacional puede ser un desafío.

Ejemplo 10. Supongamos que un banco utiliza un modelo de random forest para predecir si un cliente será aprobado para un préstamo (y = 1) o no (y = 0). Se han entrenado T = 5 árboles de decisión, los cuales producen las siguientes predicciones para un cliente con ingreso mensual de $4 000 y un buen historial de pagos:

$$f_1(x_i) = 1,\ f_2(x_i) = 0,\ f_3(x_i) = 1,\ f_4(x_i) = 1,\ f_5(x_i) = 0,$$

Para clasificación, la predicción final se obtiene mediante la *moda* de las predicciones individuales:

$$f(x_i) = \text{moda}([1, 0, 1, 1, 0]) = 1.$$

Por lo tanto, el modelo predice que el cliente será aprobado para el préstamo.

Si en lugar de clasificación fuera un problema de regresión, por ejemplo, estimar el monto del préstamo aprobado en miles de dólares, y los cinco árboles produjeran las siguientes predicciones:

$$f_1(x_i) = 20,\ f_2(x_i) = 25,\ f_3(x_i) = 22,\ f_4(x_i) = 23,\ f_5(x_i) = 21,$$

En este caso, la predicción final sería el *promedio* de los valores predichos:

$$f(x_i) = \frac{1}{5}(20 + 25 + 22 + 23 + 21) = \frac{111}{5} = 22.2$$

Por lo tanto, el modelo estima que el cliente recibiría un préstamo de aproximadamente $22 200.

LIMITACIONES DE LOS MODELOS CLÁSICOS

A pesar de su popularidad y amplio uso en la predicción de riesgo crediticio, los modelos clásicos presentan varias limitaciones que pueden afectar su rendimiento en ciertos escenarios. A continuación, se describen las principales limitaciones de los modelos clásicos, como la regresión logística, los árboles de decisión, random forest.

a. Limitaciones de la regresión logística.

La regresión logística, aunque es un modelo simple y eficiente, tiene varias limitaciones:

- Linealidad: La regresión logística asume una relación lineal entre las características y el logaritmo de la probabilidad de default. Esto limita su capacidad para capturar relaciones no lineales entre las variables, lo que puede resultar en un modelo subóptimo cuando las relaciones en los datos son complejas (Hosmer, Lemeshow & Sturdivant, 2013).
- Manejo de variables categóricas: Aunque la regresión logística puede manejar variables categóricas mediante codificación one-hot, este enfoque puede aumentar significativamente la dimensionalidad del problema, especialmente cuando hay muchas categorías. Esto puede llevar a un sobreajuste y a un mayor costo computacional (Menard, 2002).
- Sensibilidad a datos desbalanceados: En problemas de riesgo crediticio, donde la mayoría de los clientes no incurren en default, la regresión logística puede tener dificultades para predecir correctamente la clase minoritaria (default), ya que tiende a favorecer la clase mayoritaria (King & Zeng, 2001).

b. Limitaciones de los árboles de decisión

Los árboles de decisión, aunque son intuitivos y fáciles de interpretar, presentan las siguientes limitaciones:

- Sobreajuste}: Los árboles de decisión tienden a sobreajustarse a los datos de entrenamiento, especialmente cuando son muy profundos. Esto ocurre porque el modelo puede memorizar los datos en lugar de generalizar patrones subyacentes (Breiman, Friedman, Olshen & Stone, 1984).
- Inestabilidad: Pequeños cambios en los datos de entrenamiento pueden resultar en árboles de decisión completamente diferentes. Esto hace que el modelo sea menos robusto frente a variaciones en los datos (Breiman, Friedman, Olshen & Stone, 1984).
- Manejo de variables categóricas: Aunque los árboles de decisión pueden manejar variables categóricas de manera nativa, su rendimiento puede verse afectado cuando hay muchas categorías o cuando las categorías tienen una distribución desigual (Breiman, Friedman, Olshen & Stone, 1984).

c. Limitaciones de random forest

Aunque random forest es más robusto que un solo árbol de decisión, todavía tiene algunas limitaciones:

- Complejidad computacional: random forest requiere entrenar múltiples árboles de decisión, lo que puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes o con muchas características (Breiman, 2001).
- Interpretabilidad: Aunque random forest mejora la precisión, pierde la interpretabilidad de un solo árbol de decisión, ya que combina las predicciones de muchos árboles (Breiman, 2001).
- Manejo de datos desbalanceados: Al igual que los árboles de decisión, random forest puede tener dificultades para predecir correctamente la clase minoritaria en problemas desbalanceados, a menos que se utilicen técnicas de balanceo (Breiman, 2001).

CONCLUSIÓN

El artículo destaca la importancia de los modelos predictivos en el riesgo crediticio, ya que permiten predecir eventos como el incumplimiento de pagos mediante técnicas supervisadas (regresión, clasificación) y no supervisadas (clustering). Estos modelos mejoran la toma de decisiones y reducen riesgos financieros. Sin embargo, presentan limitaciones: la regresión logística asume linealidad y sufre con datos desbalanceados; los árboles de decisión son inestables y propensos a sobreajuste; y el random forest, aunque más robusto, es menos interpretable y computacionalmente costoso. Como solución, se sugiere explorar modelos más avanzados como CatBoost, un algoritmo de gradient

boosting especialmente diseñado para manejar variables categóricas de manera nativa y eficiente, evitando los problemas de codificación que afectan a otros modelos. CatBoost también incorpora técnicas innovadoras para reducir el sobreajuste y mejorar el rendimiento con datos desbalanceados, como el uso de permutaciones ordenadas y un esquema de crecimiento de árboles equilibrado. Además, su capacidad para proporcionar importancia de características lo hace más interpretable que otros enfoques complejos.

REFERENCIA

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). Classification and regression trees (1st ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Hosmer, D. W. J., Lemeshow, S., y Sturdivant, R. X. (2013). Applied logistic regression. Wiley. doi: 10.1002/9781118548387

King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Thomas, L; Crook, J. y Edelman, D. (2017). Credit Scoring and Its Applications. Mathematics In Industry

Menard, S. (2002). Applied logistic regression analysis. Thousand Oaks: Sage Publications.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). Catboost: Unbiased boosting with categorical features. arXiv preprint, v5. https://doi.org/10.48550/arXiv.1706.09516