MACHINE LEARNING APLICADO A LA GESTIÓN DE RIESGOS. CASO DE UNA CARTERA CASTIGADA DE CRÉDITOS POR CONVENIOS DE UNA ENTIDAD FINANCIERA EN PERÚ

Lic. Rayza Gomez
Universidad de Buenos Aires. Facultad de Ciencias Económicas
rayza.gomez@hotmail.com

Recibido el 4 de agosto de 2023. Aceptado el 27 de septiembre de 2023

Resumen

Hoy en día, las áreas encargadas de brindar soporte analítico dentro de las entidades financieras se enfrentan a un desafío constante. Parte de este desafío consiste en replantear sus procesos internos, cuyo rol principal es dar soporte a áreas importantes como la de gestión de riesgos.

Esta situación, ha dado lugar a que el uso de modelos de aprendizaje automático o *machine learning* surjan como alternativa de solución para las organizaciones y el sector bancario no es la excepción. Por lo que, este trabajo busca proponer un modelo de árbol de decisión como alternativa frente a un modelo tradicional *logit*, aplicado al análisis de créditos castigados del producto Convenios, durante el 2020 en el área de gestión de riesgos de una entidad financiera situada en Perú.

Los modelos de *machine learning* destacan por tener la capacidad de detectar patrones significativos a partir de los datos y realizar predicciones a partir de los mismos. Por otro lado, el modelo *logit* o de regresión logística estima la probabilidad de ocurrencia de un evento en base a un conjunto de variables explicativas.

Para contextualizar, en primer lugar, se va a describir la problemática actual que atraviesa el área de gestión de riesgos de esta entidad. Luego, se explica la metodología utilizada y finalmente se aplicarán dos modelos: uno de árbol de decisión y otro de regresión logística, que serán comparados con métricas como el *accuracy* o precisión a fin de comparar su capacidad predictiva.

Palabras Clave: *machine learning,* modelo logístico*,* crédito castigado.

Código JEL
G32

MACHINE LEARNING APPLIED TO RISK MANAGEMENT. CASE OF A CREDIT PORTFOLIO WRITE-OFF DUE TO AGREEMENTS OF A FINANCIAL INSTITUTION IN PERU

Abstract

Today, the areas in charge of providing analytical support within financial institutions face a constant challenge. Part of this challenge consists of rethinking your internal processes, whose main role is to support important areas such as risk management.

This situation has led to the use of machine learning models emerging as an alternative solution for organizations and the banking sector is no exception. Therefore, this work seeks to propose a decision tree model as an alternative to a traditional logit model, applied to the analysis of written-off loans of the Agreements product, during 2020 in the risk management area of a financial entity located in Peru.

Machine learning models stand out for having the ability to detect significant patterns from data and make predictions from them. On the other hand, the logit or logistic regression model estimates the probability of occurrence of an event based on a set of explanatory variables.

To contextualize, first of all, the current problems that the risk management area of this entity is going through will be described. Then, the methodology used is explained and finally two models will be applied: one of decision tree and another of logistic regression, which will be compared with metrics such as accuracy or precision in order to compare their predictive capacity.

Keywords: machine learning, logistic regression, punished credit.

JEL Code

G32

1. Situación actual

La presente investigación se desarrolla en el contexto de una entidad financiera ubicada en Perú con casa matriz en Colombia. La entidad se caracteriza por tener un portafolio importante de créditos por Convenios. Esta modalidad consiste en otorgar créditos dinerarios de consumo a solicitud del empleado de empresas públicas o privadas, bajo la condición de que las cuotas sean descontadas por el empleador dentro de la planilla de pagos (Antúnez de Mayolo et al, 2020). La problemática tiene lugar en el área de gestión de riesgos, la cual tiene como finalidad velar por que la cartera se mantenga en niveles adecuados de riesgos e identificar motivos que generen incumplimiento en el pago puntual de los préstamos.

Sin embargo, existen inconvenientes para llevar a cabo estos fines de manera óptima. Para ser más preciso, el incremento de clientes impulsó a que la cartera de clientes -sobre todo la cartera de convenios- sea cada vez más diversa. Las metodologías vigentes se basan en estadísticas descriptivas y modelos estadísticos tradicionales como los de regresión logística y estos resultan insuficientes para identificar aquellos motivos que hacen que el cliente caiga en *default* o llegue a la instancia de ser castigado[1]. Frente a esta problemática, surge la iniciativa de probar con un modelo de *machine learning* de árbol de decisión como alternativa que supere a los modelos actuales en términos de precisión.

En esta oportunidad, se va a poner foco en el portafolio de créditos por convenios que han sido castigados durante el año 2020 así como sus principales características tales como la edad, sueldo, género y si el cliente trabaja o no para las fuerzas armadas (FFAA); debido a que la gerencia ha detectado que muchos de estos clientes tienen la particularidad de haber sido cesados de su centro de trabajo; lo cual resulta interesante, ya que la esencia del crédito por convenio radica en el descuento directo de las cuotas por parte de la empresa donde labora el cliente.

2. Metodología

Los bancos pueden mejorar la gestión de su portafolio con ayuda de modelos basados en matemáticas y datos. La digitalización y el incremento del número de

[1] Cartera castigada: créditos clasificados como pérdida, íntegramente provisionados, que han sido retirados de los balances de las empresas. Para castigar un crédito, debe existir evidencia real de su irrecuperabilidad o debe ser por un monto que no justifique iniciar acción judicial o arbitral (SBS, 2015).

clientes han generado la necesidad de desarrollar metodologías capaces de clasificar o predecir con la mayor precisión posible. Dentro del rubro financiero se están aplicando algoritmos de *machine learning*, los cuales se han adaptado con éxito al número de datos con información financiera de clientes. Este tipo de algoritmos tienen la capacidad de ser aplicables a diferentes áreas del negocio. Desde la detección del riesgo hasta la identificación de identidad con apoyo de biometría e imágenes (Leo, Sharma y Maddulety, 2019). Dada esta versatilidad, surge la inquietud de aplicar este tipo de modelos en el marco de la gestión de riesgo de crédito de la entidad que se está tomando como referencia, con la finalidad de resolver la problemática descrita en el punto anterior.

2.1. Árbol de decisión (Decision tree)

El árbol de decisión es un método no paramétrico que no requiere supuestos distribucionales, permite detectar interacciones, modela relaciones no lineales y no es sensible a la presencia de datos faltantes y *outliers* (Breiman, Friedman, Olshen & Stone, 1984). Operan con diferentes metodologías, entre las que están los Cart, Chaid y Chaid exhaustivo. Estas difieren en la forma de asignación, reglas de partición y criterios de parada. Cualquiera sea el método, se generan n nodos terminales y una escala de probabilidades con n posibles valores que es el resultado final (Hernandez C.,2004).

Esta investigación se va a centrar en los árboles de decisión de clasificación, los cuales funcionan como diagramas de flujo. Cada nodo de un árbol de decisión representa un punto de decisión que se divide en dos nodos de hoja. Cada uno de estos nodos representa el resultado de la decisión y cada una de las decisiones también puede convertirse en nodos de decisión. Eventualmente, las diferentes decisiones conducirán a una clasificación final. El siguiente diagrama ilustra como es que funcionan los árboles de decisión para llegar a un resultado final:

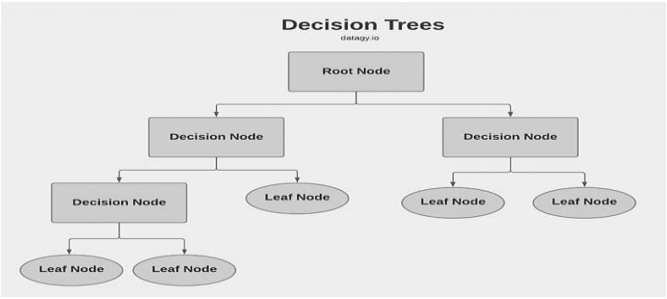

Figura 1: Diagrama de árbol de árbol de decisión. Fuente: Datagy

Supongamos que se tienen 10 variables predictoras con dos categorías o clases cada una. Las posibles combinaciones serían más de mil. Es ahí donde cobra importancia el árbol de decisión, ya que el algoritmo devolverá el mejor árbol que tome la decisión más acertada haciendo uso de las probabilidades. Para este caso en específico, las variables decisoras serán el rango de edad, rango de ingresos, género y si el cliente trabaja o no para las fuerzas armadas (FFAA) y la variable dependiente es si el cliente llegó a ser castigado por haber sido cesado de su empleo.

2.2. Modelo de regresión logística

Es una técnica estadística que permite estimar la relación entre una variable dependiente dicotómica no métrica y un conjunto de variables que pueden ser métricas y no métricas.

La mayoría de los problemas que se encuentran en el mundo real derivan del intento de expresar una probabilidad comprendida entre 0 y 1 a través de una forma lineal que en principio no está acotada, y el fin de este modelo es acotar esta forma funcional (Álvarez Garcia, 2007). Es decir, en lugar de una relación del tipo:

$$y_i = x'_i\beta + \mu_i \qquad (1)$$

Se formula la siguiente relación:

$$y_i = F(x'_i\beta) + \mu_i \qquad (2)$$

Donde $F(.)$ es una función de R en R, β un vector de parámetros desconocidos k-dimensional y μ_i un término de error con $E(\mu_i) = 0$, lo cual implica

$$E(y_i) = F(x'_i\beta) \qquad (3)$$

$$\rightarrow Pr(y_i = 1) = F(x'_i\beta) \qquad (4)$$

Las igualdades (3) y (4) son totalmente equivalentes por ser y_i una variable binaria. En la práctica, es más habitual formular los modelos para variables binarias utilizando la forma de la ecuación (4). La distribución implica suponer que la variable dependiente sigue una distribución binomial tal que:

$$\begin{cases} y_i = 1 & \Pr(y_i = 1) = F(x'_i\beta) \\ y_i = 0 & \Pr(y_i = 0) = 1 - F(x'_i\beta) \end{cases}$$

La especificación para la distribución logística es la siguiente:

$$F(x'_i\beta) = \frac{exp(x'_i\beta)}{1 + exp(x'_i\beta)} = \frac{1}{1 + exp(-x'_i\beta)}$$

En la actualidad, se puede encontrar en la literatura distribuciones diferentes a la logística o a la normal, sin embargo, hasta ahora estas han sido las más utilizadas.

Es importante mencionar que bajo la especificación *logit*, la función de regresión es no lineal en los parámetros, por lo cual el modelo no puede estimarse con mínimos cuadrados ordinarios, y como medida alternativa, se debe optar por la estimación de máxima verosimilitud. No se va a desarrollar este método de estimación, ya que escapa de los objetivos de esta investigación.

Para resolver el problema planteado, la especificación del modelo logístico sería la siguiente:

$$Prob(Y_i = Castigado\ por\ cese) = \frac{1}{1 + e^{-\alpha - \beta_1 * edad - \beta_2 * ingresos - \beta_3 * género - \beta_4 * FFAA}}$$

2.3. Criterios de evaluación de modelos

Los modelos se evalúan en función de su capacidad predictiva cuando se tiene que discriminar entre una u otra clase en cuestión. La evaluación se da comparando las clases predichas por el modelo con respecto a la clase real.

Una vez seleccionadas las variables que se utilizarán para modelar, es necesario dividir el conjunto de datos en entrenamiento y prueba. Es importante realizar este proceso a fin de aislar variables que el modelo no haya "visto", para saber si realmente aprendió a desarrollar la tarea que buscaba aprender o simplemente memorizó los datos que se usaron en el entrenamiento (Martinez, 2022).

2.3.1. Matriz de confusión

Una manera de comparar las predicciones del modelo con la clase real a la que pertenece cada individuo es la matriz de confusión:

		PREDICCIÓN	
		NEGATIVO	POSITIVO
OBSERVACIÓN	NEGATIVO	Verdaderos negativos (VN)	Falsos Negativos (FN)
	POSITIVO	Falsos Positivos (FP)	Verdaderos Positivos (VP)

Figura 2: Matriz de confusión. Fuente: Elaboración propia

Verdaderos Positivos (VP): Número de observaciones que se clasificaron correctamente como "positivos".

Verdaderos Negativos (VN): Número de observaciones que se clasificaron correctamente como "negativos".

Falsos Positivos (FP): También conocido como error tipo I, es el número de observaciones que se clasificaron incorrectamente como "positivos".

Falsos Negativos (FN): También conocido como error tipo II, es el número de observaciones que se clasificación incorrectamente como "negativos".

2.3.2. Métricas de evaluación

Las métricas más usadas para la evaluación de modelos de respuesta binaria son:

Exactitud (*Accuracy*): Proporción de predicciones correctas.

$$Exactitud = \frac{VP + VN}{Total\ de\ obs.} = \frac{VP + VN}{VP + FP + FN + VN}$$

Tasa de Error: Proporción de observaciones clasificadas incorrectamente.

$$Tasa\ de\ error = 1 - Exactitud = \frac{FP + FN}{Total\ de\ obs.}$$

Sensibilidad (Precisión): También conocido como tasa de verdaderos positivos, es la proporción de casos positivos que fueron correctamente identificados.

$$Sensibilidad = \frac{VP}{Total\ positivos} = \frac{VP}{VP + FN}$$

Especificidad (*Recall*): También conocido como tasa de verdaderos negativos, es la proporción de casos negativos correctamente identificados.

$$Especificidad = \frac{VN}{Total\ negativos} = \frac{VN}{VN + FP}$$

Tasa de Falsos positivos: También conocido como Error tipo I, es la probabilidad de que se dé un resultado positivo cuando el valor verdadero es negativo.

$$TFP = 1 - Especificidad = \frac{FP}{VN + FP}$$

Tasa de Falsos negativos: También conocido como Error tipo II, es la probabilidad de que la prueba pase por alto un verdadero positivo, es decir, que se dé un resultado negativo cuando el verdadero valor es positivo.

$$TFN = 1 - Sencibilidad = \frac{FN}{VP + FN}$$

2.3.3. Curva ROC (Receirver Operating Characteristic)

La curva ROC es una representación gráfica de la fracción de falsos positivos (abscisas) frente a la fracción de verdaderos positivos (ordenadas). Son de amplio uso para la evaluación de métodos clasificatorios. Estos tratan de identificar a que tipo de eventos pertenecen las observaciones (Bouza, 2021), en particular, para este caso de estudio, la problemática planteada requiere identificar clientes que han sido castigados por que fueron cesados en base a características como el género, ingresos y edad.

$$ROC = \begin{cases} y = FP \\ x = VP \end{cases}$$

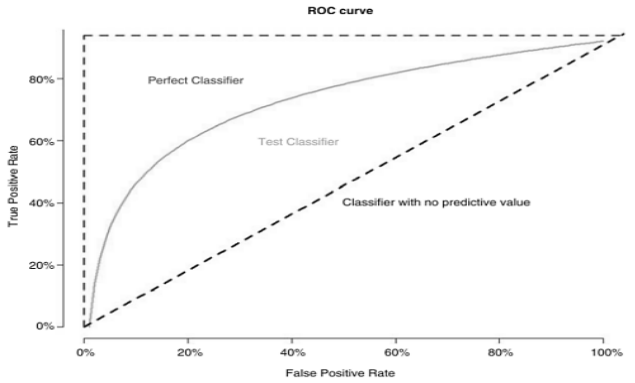

Figura 3: Curva ROC. Fuente: Ivo Dinov, 2018

Las líneas punteadas laterales del plano representan una clasificación perfecta donde se tiene 0 % de falsos positivos y 100 % de verdaderos positivos, la curva de color gris representa un ejemplo de curva de ROC para un modelo de clasificación, y la línea diagonal color negro, corresponde a un modelo incapaz de discriminar entre las clases positivas y negativas. Una manera de cuantificar el rendimiento de este tipo de modelos es el área bajo la curva (AUC_2). Esta medida determina que tan bueno es el modelo para discriminar entre una clase u otra. Según Ivo D. Dinov (2018), la siguiente tabla muestra la valoración del modelo en función al valor del AUC bajo la curva ROC:

AUC	Desempeño
0.5-0.6	Sin discriminación
0.6-0.7	Malo
0.7-0.8	Regular
0.8-0.9	Bueno
0.9-1.0	Excelente

Tabla 1: Valoración AUC. Fuente: Ivo Dinov, 2018

[2] Área Under the Curve

3. Modelo de árbol de decisión para créditos castigados

La fuente principal de datos para este estudio será la cartera de créditos por convenio que fue castigada por la entidad financiera durante el año 2020.

La variable de estudio es el motivo de castigo "cese" que toma valores 0 y 1, donde 1 representa al cliente que dejó de pagar por haber sido cesado en su centro de labores y 0 caso contrario. El análisis tiene una parte exploratoria de las variables predictoras, seguido del entrenamiento del modelo de árbol de decisión y la aplicación de un modelo de regresión logística. Para verificar la eficiencia que tiene el árbol de decisión, se va a contrastar con los resultados de un modelo logístico tradicional; ambos modelos serán aplicados a la cartera castigada.

3.1. Análisis exploratorio

El set de datos está compuesto por 1386 clientes, no registra valores perdidos ni filas duplicadas.

Number of observations	1386
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Figura 4: Descripción del set de datos. Fuente: Elaboración propia

Variables explicativas no numéricas:

Se tienen cuatro variables no numéricas: rango de edad, rango de ingresos, género y fuerzas armadas. Todas son del tipo cualitativo que describen características personales del cliente.

Para la edad se tienen cuatro rangos que agrupan a clientes menores de 21 años hasta los que superan los 51. Se observa que, en esta cartera, la mayor parte supera los 51 años, seguido de los que tienen entre 28 y 38 (Gráfico 1).

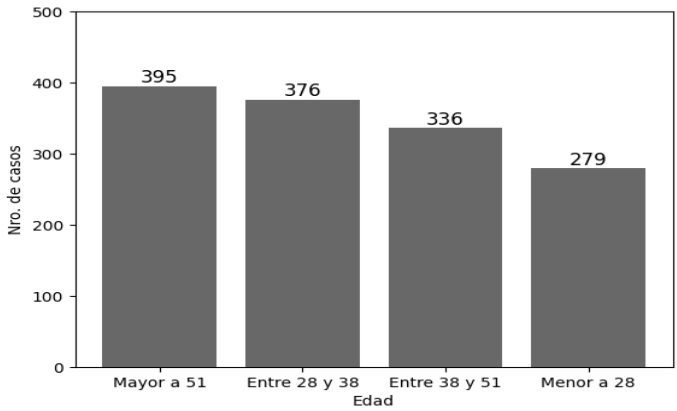

Gráfico 1: Rango de edad en años. Fuente: Elaboración propia

Los ingresos también están representados en cuatro clases diferentes. No existe concentración de una en específico. La clase con más individuos es la que posee ingresos netos mensuales entre S/2100 y S/2800, seguido de los que ganan menos de S/1400. Predominando así clientes con niveles de ingresos que están cerca al sueldo mínimo (S/1025).

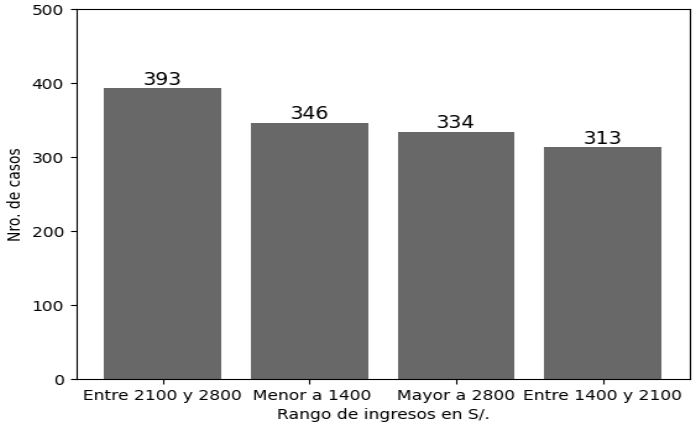

Gráfico 2: Rango de ingresos en nuevos soles (S/). Fuente: Elaboración propia

Con relación al género, predominan los varones, representando el 80.7% del total.

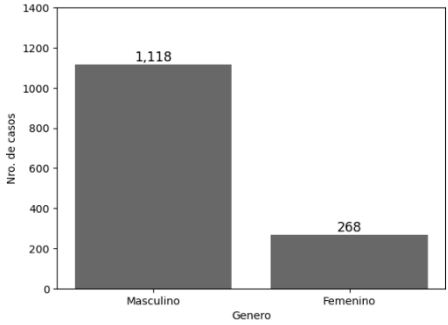

Gráfico 3: Género. Fuente: Elaboración propia

La variable Fuerzas Armadas, es un indicador que marca si el cliente trabajaba en una entidad correspondiente a instituciones gubernamentales de las fuerzas armadas del Perú. Se consideró esta variable dado que el sector público, en específico las fuerzas armadas forman parte del mercado objetivo de créditos por convenios de la entidad financiera. El 63.8% de la muestra en estudio trabajó en las fuerzas armadas (FFAA).

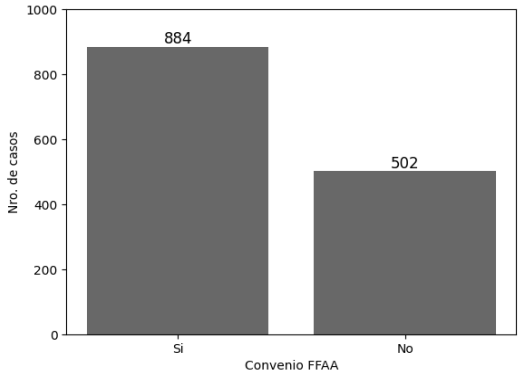

Gráfico 4: Convenio Fuerzas Armadas. Fuente: Elaboración propia

Variable respuesta:

La variable en estudio es el motivo de castigo "cese". Este motivo, agrupa aquellos clientes fueron despedidos y como consecuencia, tuvieron falta de ingresos para cumplir con sus obligaciones. En esta cartera, se observa que, de los castigados durante el 2020, 626 clientes (45%) fueron despedidos de su centro de labores.

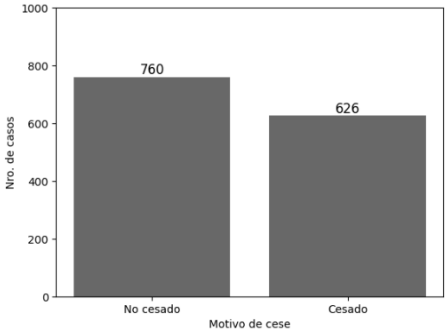

Gráfico 5: Variable respuesta. Fuente: Elaboración propia

3.2. Modelo de árbol de decisión

Este modelo tiene una precisión del 65%. Si se pone atención a la clase de interés, que es el grupo de clientes castigados porque dejaron de pagar dada su situación de despido (cese), la sensibilidad o tasa de acierto para este grupo es del 74%. Por otro lado, la clase correspondiente a otros motivos tiene una precisión del 53%. Esto pondera a la baja la precisión total del modelo. La curva ROC (Gráfico 7), tiene un AUC de 73%. De acuerdo con la tabla 1 (Punto 3.3.3), este modelo tiene un rendimiento regular, ya que no supera el 80%.

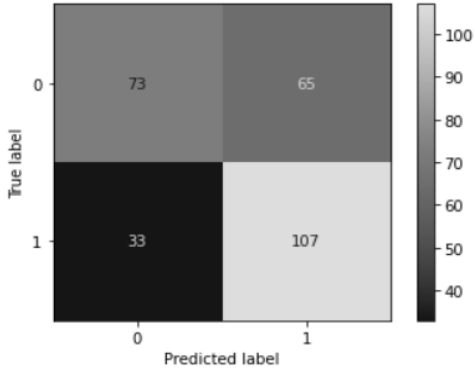

Gráfico 6: Matriz de confusión -Modelo árbol de decisión.

Fuente: Elaboración propia

	precision	recall	f1-score	support
0	0.69	0.53	0.60	138
1	0.62	0.76	0.69	140
accuracy			0.65	278
macro avg	0.66	0.65	0.64	278
weighted avg	0.66	0.65	0.64	278

Figura 5: Reporte de clasificación.

Fuente: Elaboración propia

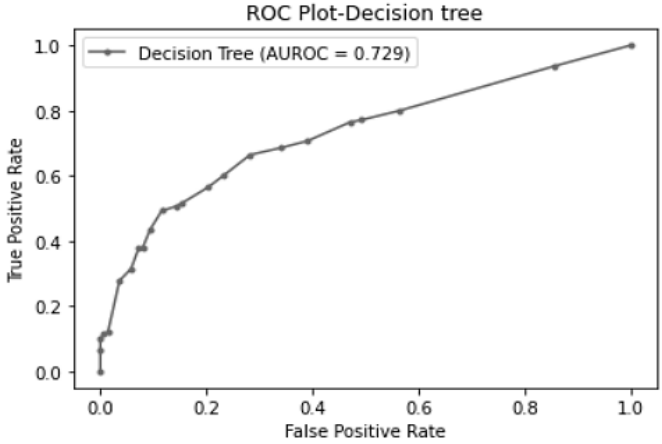

Gráfico 7: Curva ROC -Modelo árbol de decisión.

Fuente: Elaboración propia con librería *sklearn*

3.3. Modelo de regresión logística

Este modelo predice probabilidades de ocurrencia para un evento en específico. Las mismas se calculan en función a un grupo de variables predictoras. El contexto de esta investigación plantea que, a través de este modelo, se calcule la probabilidad de que un cliente haya sido castigado por el motivo cese. El resultado indica que, a nivel global este modelo es significativo, considerando los ingresos, edad, género y si el cliente cumple o no con la condición de haber trabajado en una entidad de fuerzas armadas. Sin embargo, las métricas de evaluación muestran que el modelo tiene baja capacidad predictiva. Solo el *accuracy* es del 56%; a nivel de grupo, el *recall* de la clase de interés (1) es del 37%, teniendo mayor acierto para los otros casos donde el cliente no fue castigado por despido, es decir dejó de pagar por otros motivos. El área bajo la curva es cercana al modelo base (sin variables predictoras), tomando un valor de 59.7%.

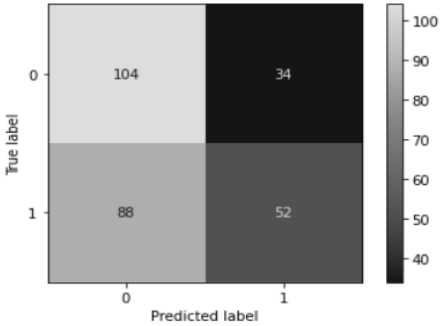

Gráfico 8: Matriz de confusión -Modelo regresión logística

Fuente: Elaboración propia

	precision	recall	f1-score	support
0	0.54	0.75	0.63	138
1	0.60	0.37	0.46	140
accuracy			0.56	278
macro avg	0.57	0.56	0.55	278
weighted avg	0.57	0.56	0.54	278

Figura 6: Reporte de clasificación. Fuente: Elaboración propia

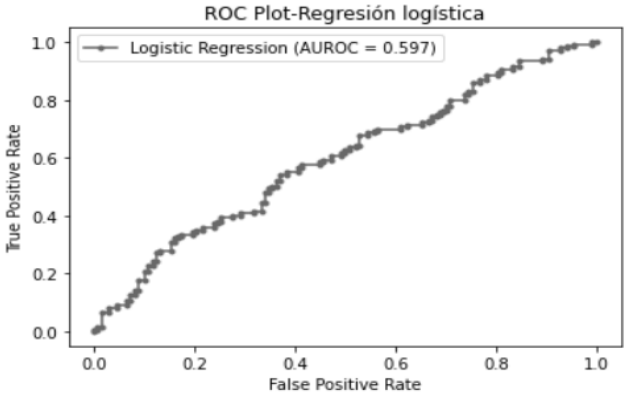

Gráfico 9: Curva ROC - Modelo de reg. logística

Fuente: Elaboración propia

3.3.1. Comparación de modelos

Se puede tomar la tabla 6 para evaluar la capacidad predictiva de cada modelo. Se observa que el mejor desempeño del modelo de árbol de decisión o *decisión tree*, en tanto que el modelo de regresión logística tiene tasas bajas de precisión.

MODELO	Accuracy	Sensibilidad	Especificidad	TFP	TFN
Decisión Tree	64.7%	76.4%	68.9%	31.1%	23.6%
Regresión logística	56.1%	37.1%	54.2%	45.8%	62.9%

Tabla 2: Métricas de evaluación. Fuente: Elaboración propia

El gráfico 10, compara la curva ROC para ambos modelos, en este se evidencia de forma conjunta que el modelo de árbol de decisión tiene un AUC superior al de la regresión logística.

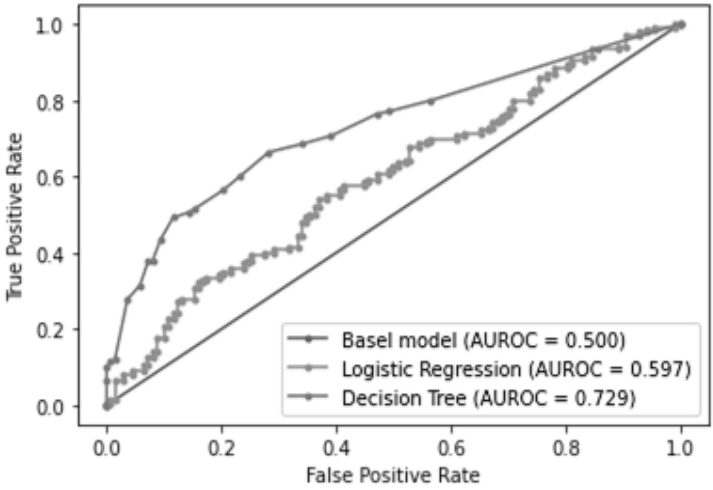

Gráfico 10: Curva ROC. Fuente: Elaboración propia

4. Conclusiones

Habiendo aplicado las metodologías desarrolladas; se concluye que el modelo de árbol de decisión planteado mostró mayor precisión que el modelo logístico para identificar clientes que fueron castigados por cese. De acuerdo con las métricas de evaluación, este predice mejor en base a la edad, ingresos mensuales, género y si trabajó o no para las fuerzas armadas (FFAA).

Emplear modelos de machine learning como el árbol de decisión, representa una alternativa que permite aprovechar al máximo la información de los clientes. Esto en el sentido de que este tipo de algoritmos tienen la capacidad computacional de hacer múltiples combinaciones entre las variables predictoras. Cosa que el modelo de regresión logística simple no hace.

Los resultados obtenidos en el modelo de árbol de decisión permite caracterizar e identificar patrones entre los clientes que fueron castigados por motivo de

cese. Conocer características en específico, permite hacer un *feedback* de cara a mejorar políticas de aceptación de riesgo y poner énfasis a determinados clientes a la hora de hacer seguimiento de cartera. Así mismo, se puede adoptar alguna estrategia que impida que el crédito llegue a la instancia de ser castigado.

Referencias bibliográficas

Álvarez Garcia, B. (2007). Modelos de elección binaria.

Bouza, C. (2021). Las curvas ROC teoría y herramientas para su uso.

Breiman, Friedman, Olshen, & Stone. (1984). *Classification and regresion trees.*

Hernández, C. (2004). Aplicación de árboles de decisión en modelos de riesgo de crédito. *Revista colombiana de estadística*.

Ivo, D. (2018). Data Science and predictive analytics.

Leo, Sharma, & Maddulety. (2019). Machine Learning in Banking Risk Management:A literature review. *MDPI*.

Martínez Fernandez, T. C. (2022). Comparación de modelos machine learning aplicados al riesgo de crédito.

Mayolo, A. d., Falcón, P., & Machaca, Q. (2020). *Plan de Negocio para la introducción del producto Crédito por Convenio en las pequeñas y medianas empresas de Lima.* Lima.