MODELADO PREDICTIVO DEL SCORE CREDITICIO PARA CLIENTES NO BANCARIZADOS: UNA APLICACIÓN DE CATBOOST

https://doi.org/10.22533/at.ed.128112517036

Data de aceite: 04/07/2025

Carlos Alberto Peña Miranda
Universidad Nacional Mayor de San Marcos
https://orcid.org/0000-0002-4339-4615

Jesús Adalberto Zelaya Contreras
Universidad Nacional Mayor de San Marcos
https://orcid.org/0009-0003-4767-6366

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

RESUMEN: Este trabajo presenta el diseño y la evaluación de un modelo de Score de admisión para tarjetas de crédito dirigido a clientes No Hit, es decir, personas sin historial en el sistema financiero formal. El modelo busca predecir el riesgo crediticio clasificando a los solicitantes como "buenos" o "malos", según su probabilidad de presentar un incumplimiento mayor a 60 días en los 12 meses posteriores a la emisión del crédito. Para su desarrollo, se empleó el algoritmo CatBoost, con implementación en Python. El proceso incluyó un preprocesamiento exhaustivo de datos, que abarcó la imputación de valores faltantes y el tratamiento de variables categóricas. El modelo fue entrenado con 54 337 registros y validado con 17 940. Los resultados evidencian un desempeño competitivo, con métricas como el coeficiente GINI (45.14 % en entrenamiento y 42.84 % en validación), AUC y KS. Además, la tabla de eficiencia muestra una relación inversa entre el score y la tasa de incumplimiento. La interpretación del modelo se realizó mediante valores SHAP, lo que permitió identificar las variables más influyentes de forma transparente. En conjunto, esta propuesta contribuye a la inclusión financiera responsable, facilitando la evaluación del riesgo en poblaciones tradicionalmente desatendidas.

PALABRAS CLAVE: Score crediticio, CatBoost, variables categóricas, riesgo de default, SHAP values.

PREDICTIVE CREDIT SCORE MODEL FOR NON-BANK CUSTOMERS: A CATBOOST APPLICATION

ABSTRACT: This work presents the design and evaluation of a credit admission scoring model for "No Hit" clients, that is, individuals without a formal credit history in the financial

system. The model aims to predict credit risk by classifying applicants as “good” or “bad” based on the probability of a payment default exceeding 60 days within 12 months after credit issuance. The model was developed using the CatBoost algorithm, implemented in Python. The process involved extensive data preprocessing, including the imputation of missing values and the treatment of categorical variables. It was trained on 54,337 records and validated with 17,940. The results demonstrate competitive performance, with metrics such as the GINI coefficient (45.14% in training and 42.84% in validation), AUC, and KS. Additionally, the efficiency table shows an inverse relationship between the score and the default rate. Model interpretability was achieved using SHAP values, allowing for transparent identification of the most influential variables. Overall, this approach contributes to responsible financial inclusion by enabling credit risk assessment in traditionally underserved populations.
KEYWORDS: Credit score, CatBoost, categorical variables, default risk, SHAP values.

INTRODUCCIÓN

En el sector financiero, la evaluación del riesgo crediticio es fundamental para la toma de decisiones en la aprobación de tarjetas de crédito. Para los solicitantes sin historial crediticio en el sistema financiero formal, conocidos como segmento No Hit o no bancarizados esta tarea representa un desafío significativo debido a la ausencia de información crediticia previa. El Score de Admisión de Tarjetas de Crédito No Hit es un modelo predictivo diseñado para abordar esta problemática, estimando la probabilidad de que un solicitante incurra en un incumplimiento de pago (default) superior a 60 días dentro de los 12 meses posteriores a la emisión del crédito.

Este modelo clasifica a los solicitantes en dos categorías: bueno o malo, basándose en variables sociodemográficas, censales y transaccionales, en lugar de un historial crediticio tradicional. Su implementación permite a las entidades financieras optimizar la inclusión financiera, ampliando el acceso al crédito de manera responsable y sostenible para un segmento de la población históricamente desatendido.

Este trabajo presenta el diseño, desarrollo y evaluación de dicho modelo, destacando su capacidad predictiva y su relevancia en la toma de decisiones crediticias.

DISEÑO DEL MODELO

Población objetivo

La población objetivo del modelo son las personas con Documento Nacional de Identidad (DNI) que solicitan una tarjeta de crédito (TC) en el sistema financiero. Se considera que un solicitante pertenece al segmento no bancarizado (No Hit) si en los últimos 12 meses disponibles en el Registro de Créditos y Cobranzas (RCC) no ha registrado saldo, línea de crédito ni contingente en ninguna entidad financiera. Para la construcción

del modelo, se trabajó con datos de cuatro meses para el entrenamiento y un mes adicional para la validación fuera de tiempo (Out Of Time).

El principal objetivo del modelo predictivo No Hit es estimar la probabilidad de default del solicitante, es decir, la probabilidad de que la persona incurra en un atraso en el pago de su tarjeta de crédito por más de 60 días dentro de los 12 meses siguientes a la emisión del crédito. Esta estimación se basa en variables del perfil del solicitante.

Adicionalmente, se han aplicado exclusiones a nivel de la población analizada por consideraciones del negocio. Por ejemplo, se han retirado aquellos registros de personas con antecedentes de transacciones fraudulentas o posibles casos de lavado de activos, también aquellas personas que ya presentan morosidad al inicio del periodo de observación, dado que el objetivo es aprender sobre los clientes que se atrasan después del otorgamiento del crédito. Para optimizar la capacidad predictiva, se empleó una base de datos integral que incluye características sociodemográficas, censales y transaccionales de los clientes.

Variable objetivo

En este modelo, la variable objetivo (target) es la clasificación del solicitante en bueno o malo, en función de su comportamiento crediticio posterior a la obtención de la tarjeta. Una persona es considerada mala si presenta un incumplimiento de pago superior a 60 días en una ventana de observación de 12 meses posterior al crédito.

Preprocesamiento de datos

Ahora pasamos al entrenamiento del modelo. Para ello, vamos a preparar las variables que servirán para el aprendizaje del modelo, es decir, las variables independientes, también conocidas como variables de entrada y denotadas con X.

Contamos con una base de datos de un total de 54 337 registros para los períodos de entrenamiento, que abarcan de junio de 2021 a septiembre de 2021. Por otro lado, el período fuera de tiempo (OOT, por sus siglas en inglés Out Of Time), que se utilizará para evaluar el rendimiento del modelo en un período no visto durante el entrenamiento, contiene un total de 17 940 registros y corresponde a octubre de 2021.

A continuación, se presenta la descripción de las 24 variables utilizadas en el modelo, distribuidas en las Tablas 1, 2 y 3.

Descripción de la variable	Nombre de variable	Uso en el modelo	Tipo de variable
Nivel educativo	lvl_edu_new	Entrada	Categórica
Flag de situación laboral dependiente	flag_dependiente	Entrada	Numérica
Número de hijos en el hogar	nro_hijoshog	Entrada	Numérica
Clasificación por código de ubigeo	ubigeo_cat	Entrada	Categórica
Proporción de hogares con refrigeradoras o congeladoras (2017)	rat_reften_2017	Entrada	Numérica
Monto de deuda en otras empresas en los últimos 12 meses	mnt_ddaotremp12m	Entrada	Numérica
Proporción de viviendas tipo casa o departamento (2017)	propviv_tipviv_2017	Entrada	Numérica
Proporción de viviendas con paredes de ladrillo/bloque de cemento	propviv_matnoble	Entrada	Numérica
Promedio ponderado de gastos mensuales por manzana (2022)	promgast_xhog22	Entrada	Numérica

Tabla 1 *Descripción de variables sociodemográficas y del hogar*

Descripción de la variable	Nombre de variable	Uso en el modelo	Tipo de variable
Promedio de saldo tipo pasivo en los últimos 6 meses	prm_sldtippas06m	Entrada	Numérica
Promedio de deuda de telecomunicaciones en los últimos 12 meses	prm_ddatelc12m	Entrada	Numérica
Promedio de transacciones monetarias y no monetarias en ATM en 12 meses	atm_trx_prom	Entrada	Numérica
Transacción promedio en banca por internet en los últimos 12 meses	bpi_trx_prom	Entrada	Numérica
Transacción promedio en agentes financieros en los últimos 12 meses	age_trx_prom	Entrada	Numérica
Transacción promedio en tiendas de la entidad en los últimos 12 meses	tie_trx_prom	Entrada	Numérica
Promedio de saldo en cuenta sueldo vista en 12 meses	prm_sldctasld12m	Entrada	Numérica
Número de incrementos en saldo tipo pasivo en los últimos 12 meses	nro_incsldtippas12m	Entrada	Numérica

Tabla 2 *Descripción de variables financieras y transaccionales*

Descripción de la variable	Nombre de variable	Uso en el modelo	Tipo de variable
Número promedio de entidades de telecomunicaciones reportantes (12 meses)	prm_telcorep12m	Entrada	Numérica
Número de correos registrados	nro_email	Entrada	Numérica
Cantidad de transacciones por app o banca por internet en el último mes	ctd_uso_app_bpi_mon	Entrada	Numérica
Agrupación por situación laboral y edad	segmento_pea_new	Entrada	Categórica
Agrupación de variables del rubro farmacia	comp1	Entrada	Numérica
Agrupación de variables del rubro supermercados	comp2	Entrada	Numérica
Agrupación de variables del rubro tiendas por departamento	comp3	Entrada	Numérica

Tabla 3 *Descripción de variables por canales digitales y comerciales*

Como primer paso, revisamos el porcentaje de datos faltantes en las variables independientes. De las 24 variables, 17 presentan valores faltantes en la siguiente proporción, como se muestra en la siguiente Figura 1:

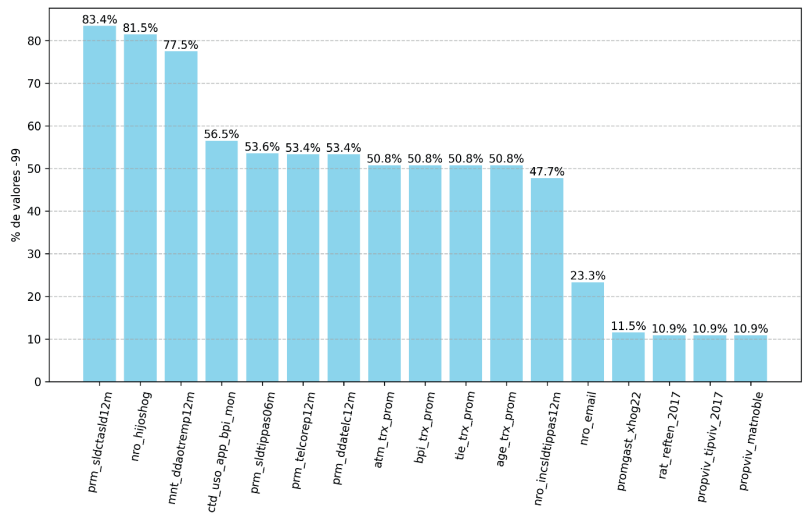


Figura 1 *Porcentaje de valores faltantes en las variables independientes*

Tratamiento de datos faltantes

Para tratar estos datos faltantes, se ha decidido, por criterio experto, reemplazar los valores faltantes por -99.

Esta decisión se basa en la correlación que tienen las variables con respecto al puntaje final predictivo asignado por el modelo (el score).

En un proceso estándar, este paso suele realizarse después del entrenamiento, ya que requiere evaluar cómo el score del modelo se comporta con respecto a la variable, es decir, si un mayor valor de la variable impacta positiva o negativamente en la predicción. Sin embargo, para efectos de comprensión de la imputación, lo incluimos en esta sección.

- prm_sldtippas06m: Esta variable es el promedio de saldo que tiene la persona en sus cuentas de tipo pasivo (por ejemplo, cuentas de ahorros) en los últimos 6 meses. Esta variable presenta una correlación positiva con el score, como se puede observar en la Figura 2.

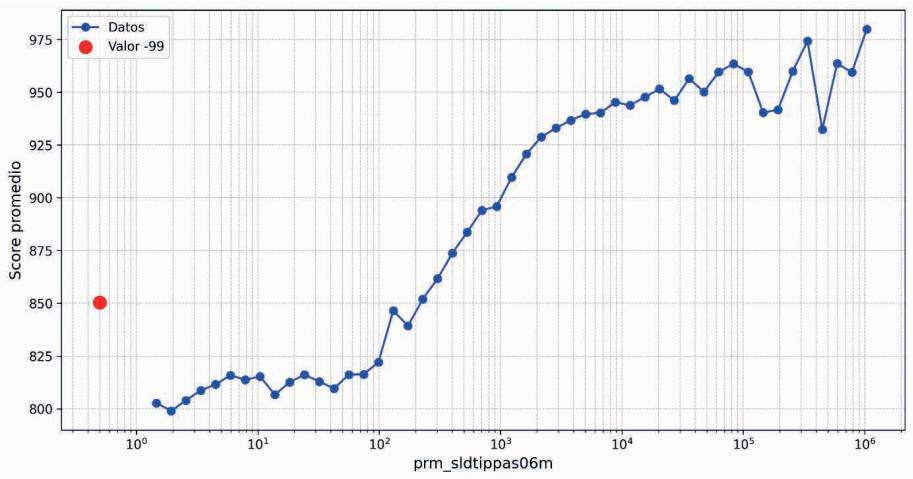


Figura 2 *Relación entre la variable promedio de saldo en cuenta de tipo pasivo en los últimos 6 meses y el score del modelo.*

Es decir, a mayor promedio de saldo, el cliente dispone de mejores fondos, lo que se traduce en un mejor perfil para la asignación de una tarjeta de crédito, razón por la cual se le otorga un score más alto. En el caso de los valores faltantes (missing), se les asigna el valor de -99, lo que implica en una predicción de un score menor. Esto se debe a que no se tiene información sobre sus fondos, lo que podría indicar un perfil de mayor riesgo.

- prm_ddatelc12m: Esta variable es el promedio de deuda que tiene la persona en empresas de telecomunicaciones en los últimos 12 meses (por ejemplo, si tiene deudas en el pago de su línea telefónica). Esta variable presenta una correlación negativa con el score como se muestra en la Figura 3.

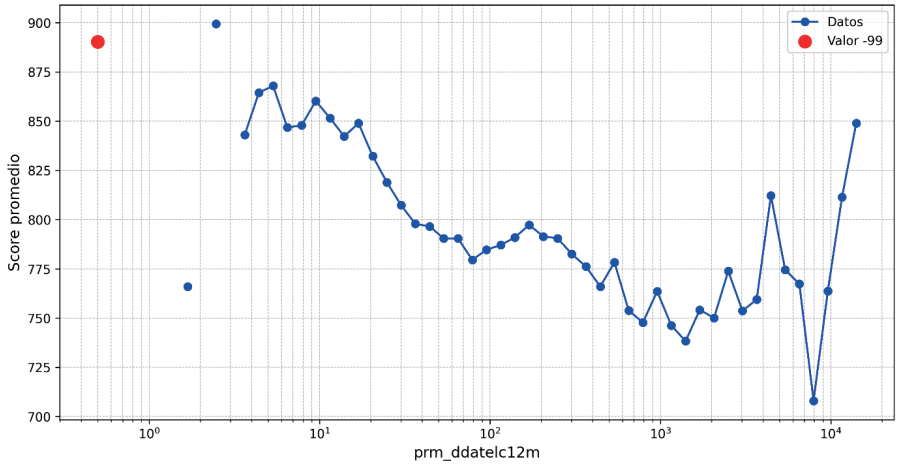


Figura 3 *Relación entre la variable promedio de deuda en telecomunicaciones en los últimos 12 meses y el score del modelo*

Es decir, a mayor deuda, el modelo asigna un score menor, ya que se trata de un perfil más riesgoso. En este caso, el dato faltante imputado con -99 indica- ría, posiblemente, que la persona no tiene deuda en telecomunicaciones. Por ello, se le asigna un score más alto, al considerarse un perfil con menor riesgo.

- prm_telcorep12m: Esta variable es el número promedio de entidades de telecomunicaciones reportantes en los últimos 12 meses (por ejemplo, si la persona debe a más de una entidad telefónica). Esta variable presenta una correlación negativa con el score, como se muestra en la Figura 4.

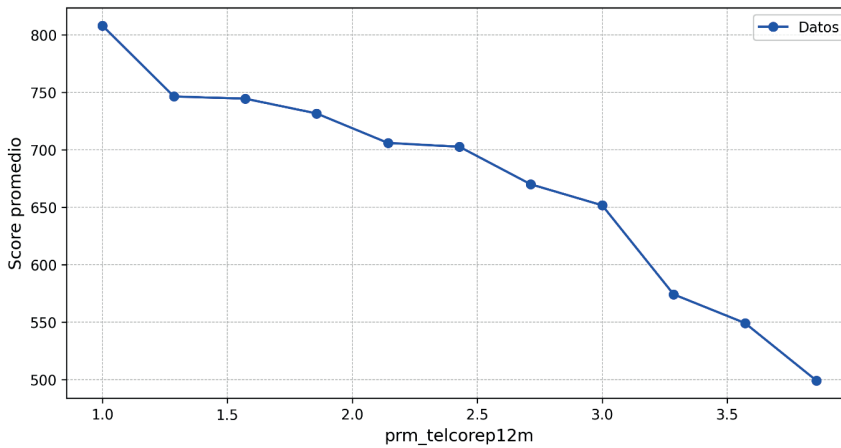


Figura 4 *Relación entre la variable número promedio de entidades de telecomunicaciones reportantes en los últimos 12 meses y el score del modelo sin considerar el dato faltante*

Es decir, a mayor número de empresas a las que se deba, el modelo asigna un score menor, ya que se trata de un perfil más riesgoso. En este caso, el dato faltante imputado con -99 indicaría, posiblemente, que la persona no tiene deudas en empresas de telecomunicaciones. Por ello, se le asigna un score más alto, como se muestra en la Figura 5.

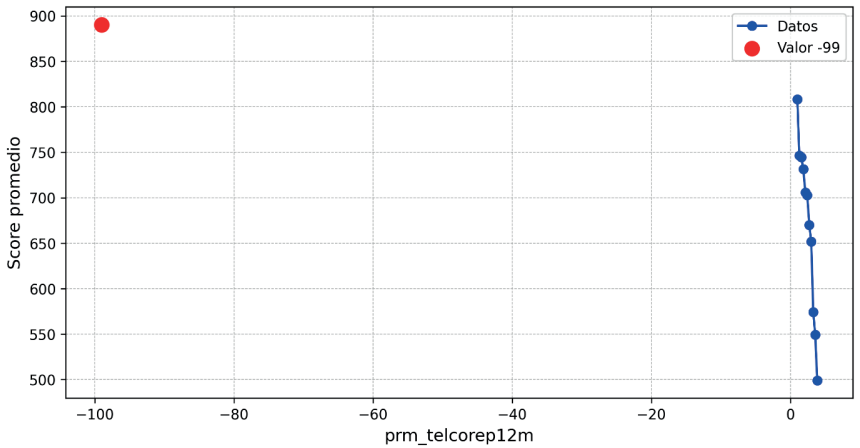


Figura 5 *Relación entre la variable número promedio de entidades de telecomunicaciones reportantes en los últimos 12 meses y el score del modelo*

Se aplica un proceso similar al resto de las variables con datos faltantes. Además de las 24 variables totales, 3 de ellas, las variables lvl_edu_new, segmento_pea_new, ubigeo_cat son de tipo categórico. Cada una de estas variables tiene las siguientes categorías, cuyo significado se detalla a continuación. Se ordenaron según su correlación significativa con el valor predicho, como se muestra a continuación:

lvl_edu_new: Esta variable representa el nivel educativo de la persona. La categoría 1 representa haber culminado los estudios de secundaria completa, la categoría 2 representa haber llegado hasta el nivel técnico o universitario, y la categoría 3 representa tener títulos de posgrado. Para nuestra población de entrenamiento representado en la Tabla 4.

lvl_edu_new	cantidad	proporción
1	26 819	49.4 %
2	23 642	43.5 %
3	3 876	7.1 %

Tabla 4 *Distribución de la variable nivel educativo*

La mayor parte se concentra en la categoría 1 con un 49.4%, seguido de la categoría 2 con 43.5% y, finalmente, la categoría 3 con 7.1%. También notamos que sigue una correlación positiva con el score, como se muestra en la Figura 6.

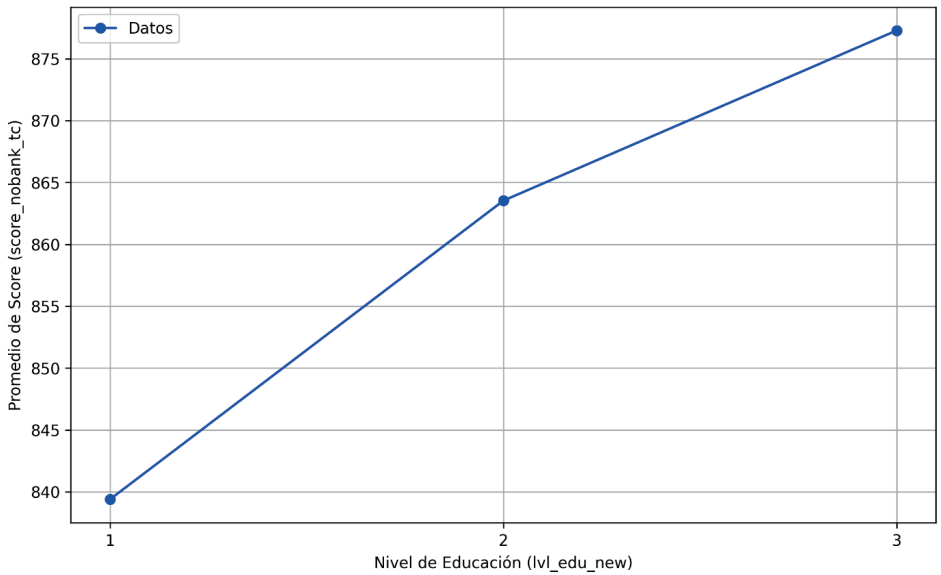


Figura 6 Relación entre la variable nivel educativo de la persona y el score del modelo

Esto debido a que un mayor nivel de estudios también se relaciona con un mayor poder adquisitivo, por lo que representa un mejor perfil.

- segmento_pea_new: Esta variable representa una agrupación por situación laboral y edad de la persona, su distribución en los periodos de entrenamiento está representado en la Tabla 5.

segmento_pea_new	cantidad	proporción
1	16 622	30.6 %
2	20 227	37.2 %
3	17 488	32.2 %

Tabla 5 *Distribución de la variable agrupación por situación laboral y edad de la persona*

La categoría 1 agrupa a jóvenes o personas en etapas tempranas de su carrera laboral, con menor estabilidad laboral o ingresos más bajos. La categoría 2 agrupa a personas con mayor estabilidad laboral y experiencia, pero que aún no alcanzan un nivel de ingresos o antigüedad significativo. Por último, la categoría 3 corresponde a personas con mayor experiencia laboral, ingresos altos y estabilidad financiera. Por ello, mantiene una correlación positiva con el score, como se muestra en la Figura 7.

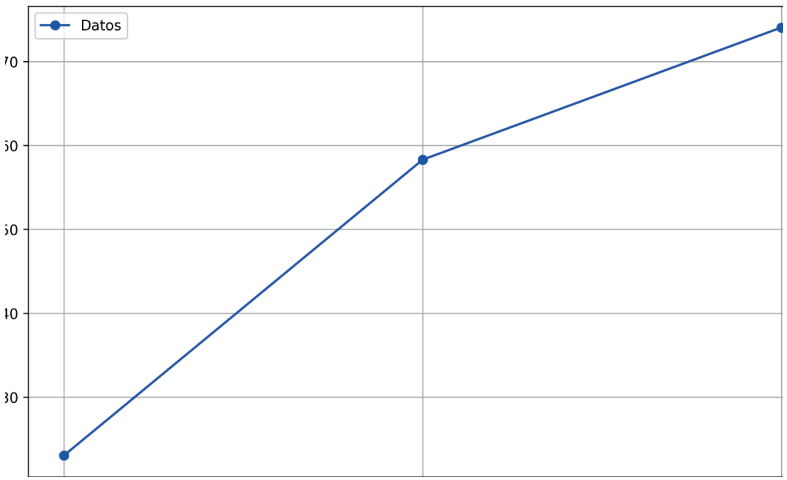


Figura 7 *Relación entre la variable agrupación por situación laboral y edad de la persona y el score del modelo*

Debido a que las personas en la categoría 1, al encontrarse en una etapa temprana de su carrera profesional, podrían representar un perfil más riesgoso. En contraste, aquellas en la categoría 3, con empleo más estable y mayor experiencia, tienen un mayor nivel adquisitivo.

- ubigeo_cat: Esta variable representa la clasificación por código de ubigeo, el cual indica la ubicación geográfica del lugar en el que nació la persona, su distribución en los periodos de entrenamiento está representada en la Tabla 6.

segmento_pea_new	cantidad	proporción
1	4 603	8.5 %
2	3 822	7.0 %
3	8 250	15.2 %
4	12 367	22.8 %
5	18 719	34.4 %
6	6 576	12.1 %

Tabla 6 *Distribución de la variable clasificación por código de ubigeo*

El valor 1 representa zonas rurales o de menor desarrollo económico. El valor 2 representa zonas semi-rurales o en proceso de desarrollo. El valor 3 representa ciudades pequeñas o intermedias. El valor 4 representa ciudades medianas con economías en crecimiento. El valor 5 representa ciudades grandes con alto desarrollo económico. El valor 6 corresponde a zonas urbanas con mayor desarrollo y acceso a recursos. Por ello, mantiene una correlación positiva con el score, como se muestra en la Figura 8.

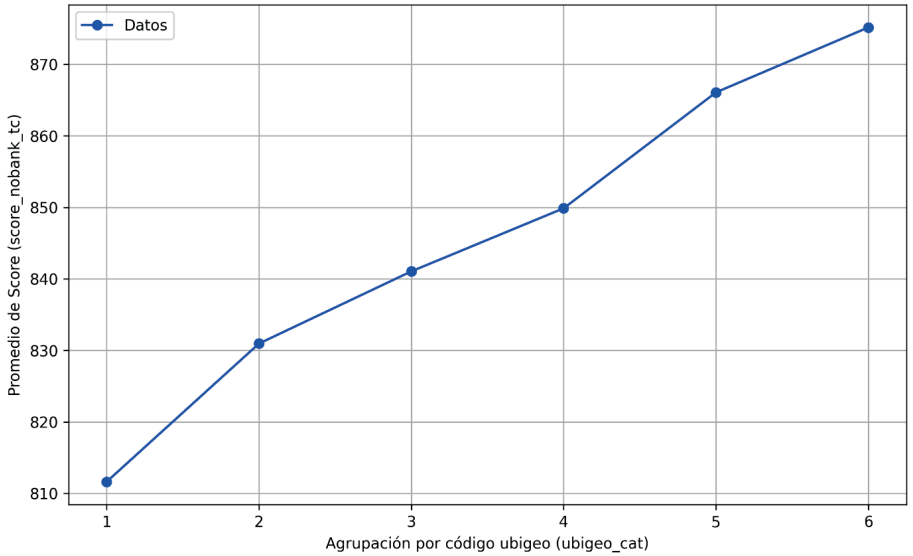


Figura 8 *Relación entre la variable clasificación por código de ubigeo y el score del modelo*

Las zonas con mayor desarrollo económico (valores más altos) tienden a tener mayor acceso a servicios financieros, mejores oportunidades laborales y mayores ingresos. Por el contrario, las zonas con menor desarrollo (valores más bajos) pueden presentar menor acceso a servicios financieros, economías menos diversificadas y menores ingresos.

A modo de resumen, las Tablas 7, 8 y 9 describen las variables considerando la imputación de valores faltantes. Estas tablas presentan el conteo total de registros de cada variable (count), el valor promedio (mean), la desviación estándar (std), el valor mínimo (min), el primer cuartil (25%), el segundo cuartil o mediana (50%), el tercer cuartil (75%) y, por último, el valor máximo (max).

Variable	count	mean	std	min	25%	50%	75%	max
prm_ sldtippas06m	54 337	1 412.1	14 441.3	-99.0	-99.0	-99.0	115.8	1 184 547.4
prm_ddatelc12m	54 337	44.9	362.6	-99.0	-99.0	-99.0	32.5	15 451.5
bpi_trx_prom	54 337	-27.8	93.5	-99.0	-99.0	-99.0	6.8	2 434.2
atm_trx_prom	54 337	-49.7	50.1	-99.0	-99.0	-99.0	0.2	50.6
age_trx_prom	54 337	-50.1	49.7	-99.0	-99.0	-99.0	0.0	56.2
tie_trx_prom	54 337	-50.2	49.6	-99.0	-99.0	-99.0	0.0	14.7
prm_ sldctasld12m	54 337	70.2	1 686.4	-99.0	-99.0	-99.0	-99.0	166 222.0
mnt_ ddaotremp12m	54 337	90.3	3 690.0	-99.0	-99.0	-99.0	-99.0	514 961.5
promgast_ xhog22	54 337	2 045.4	1 034.3	-99.0	1 825.9	2 046.3	2 417.9	6 146.0

Nota. Se observa la presencia de valores faltantes codificados como -99.0.

Tabla 7 Descripción estadística de las variables financieras utilizadas en el modelo

Variable	count	mean	std	min	25%	50%	75%	max
lvl_edu_new	54 337	1.6	0.6	1.0	1.0	2.0	2.0	3.0
ubigeo_cat	54 337	4.0	1.4	1.0	3.0	4.0	5.0	6.0
nro_hijoshog	54 337	-80.4	38.9	-99.0	-99.0	-99.0	-99.0	5.0
segmento_pea_new	54 337	2.0	0.8	1.0	1.0	2.0	3.0	3.0
rat_reften_2017	54 337	-10.1	31.1	-99.0	0.6	0.8	0.9	1.0
propviv_tipviv_2017	54 337	-10.0	31.1	-99.0	0.8	1.0	1.0	2.0
propviv_matnoble	54 337	-10.1	31.1	-99.0	0.5	0.8	0.9	1.0
flag_dependiente	54 337	0.6	0.5	0.0	0.0	1.0	1.0	1.0

Nota. Se observa la presencia de valores faltantes codificados como -99.0.

Tabla 8 *Descripción estadística de las variables sociodemográficas y del hogar*

Variable	count	mean	std	min	25%	50%	75%	max
comp1	54 337	-0.2	1.5	-29.2	-1.1	-0.6	1.1	2.3
comp2	54 337	-1.3	1.7	-44.1	-2.7	-2.0	0.9	0.9
comp3	54 337	-0.9	1.1	-33.3	-1.6	-0.5	-0.5	1.0
ctd_uso_app_bpi_mon	54 337	-53.4	52.8	-99.0	-99.0	-99.0	0.0	675.0
prm_telcorep12m	54 337	-52.5	49.8	-99.0	-99.0	-99.0	1.0	4.7
nro_email	54 337	-22.0	42.5	-99.0	-99.0	1.0	1.0	3.0

Nota. Se observa la presencia de valores faltantes codificados como -99.0.

Tabla 9 Descripción estadística de variables agrupadas por componentes y uso digital

DIVISIÓN DE DATOS

Ya con el conjunto dataset de 24 variables definido, procedemos a realizar el entrenamiento del modelo. Para ello, debemos dividir la muestra inicial en dos partes: una muestra para desarrollo (también conocida como de entrenamiento) representado en la Tabla 10.

Target	Cantidad	Proporción
Bueno	46 442	85.5 %
Malo	7 895	14.5 %
Total	54 337	100 %

Tabla 10 *Distribución de la variable objetivo (target) en la población de desarrollo*

Y para validación, representada en la Tabla 11.

Target	Cantidad	Proporción
Bueno	14 931	83.2 %
Malo	3 009	16.8 %
Total	17 940	100 %

Tabla 11 *Distribución de la variable objetivo (target) en la población de validación*

La población de desarrollo se empleará para ajustar el modelo, mientras que la de validación servirá para verificar la precisión de los resultados obtenidos.

El objetivo de esta división es crear modelos que puedan aplicarse de manera efectiva a casos nuevos, distintos de los utilizados para su desarrollo. Si se usara toda la muestra de análisis para ajustar los modelos, estos podrían diferenciar adecuadamente entre clientes buenos y malos, pero no se garantizaría que los resultados sean similares al aplicar dichos modelos en situaciones nuevas.

La base de desarrollo abarca los períodos de junio 2021 hasta septiembre 2021 (4 stocks), como se muestra en la Tabla 12.

Mes	Bueno	Malo	Total	Tasa Malos
Junio 2021	13 868	2 310	16 178	14.28 %
Julio 2021	10 286	1 698	11 984	14.17 %
Agosto 2021	10 433	1 896	12 329	15.38 %
Septiembre 2021	11 855	1 991	13 846	14.38 %

Tabla 12 *Distribución mensual de la variable objetivo (target) en la población de desarrollo*

Por su parte, la base fuera de tiempo comprende únicamente octubre de 2021 (1 stock), como se muestra en la Tabla 13.

Mes	Bueno	Malo	Total	Tasa Malos
Octubre 2021	14 931	3 009	17 940	16.77 %

Tabla 13 *Distribución mensual de la variable objetivo (target) en la población de validación*

IMPLEMENTACIÓN CON CATBOOST

Ahora pasamos al entrenamiento del modelo, utilizando el algoritmo CatBoost (Prokhorenkova, 2019). Para ello, empleamos la librería catboost, específicamente la función llamada CatBoostClassifier. A continuación, se especifican en detalle los distintos hiperparámetros utilizados para ajustar el entrenamiento del modelo.

- Número de iteraciones (iterations): Cantidad máxima de árboles de decisión que se construirán durante el entrenamiento del modelo. Un mayor número de iteraciones puede mejorar la precisión, pero también aumentar el tiempo de entrenamiento.
- Profundidad (depth): Número máximo de niveles que puede tener cada árbol de decisión. Una mayor profundidad permite que el modelo capture relaciones más complejas, pero también puede aumentar el riesgo de sobreajuste.
- Ratio de aprendizaje (learning_rate): Factor que controla la magnitud de los ajustes en los pesos del modelo después de cada actualización. Valores pequeños hacen que el aprendizaje sea más lento pero estable, mientras que valores grandes pueden acelerar el entrenamiento, pero aumentar el riesgo de sobreajuste.
- Factor de regulación (l2_leaf_reg): Parámetro que controla la penalización L2 en las hojas del árbol para evitar que el modelo se ajuste demasiado a los datos de entrenamiento. Ayuda a mejorar la generalización.
- Coeficiente de regularización del tamaño del modelo (model_size_reg): Influye en el tamaño del modelo cuando se trabajan con variables categóricas. Puede ayudar a reducir el consumo de memoria y mejorar la eficiencia del modelo.
- Método del subespacio aleatorio (rsm): Proporción de características del conjunto de datos que se utilizarán en cada división del árbol. Un valor bajo reduce el número de variables por árbol, introduciendo aleatoriedad y mejorando la generalización.
- Iteraciones de estimaciones de hojas (leaf_estimation_iterations): Número de pasos utilizados para calcular los valores de las hojas en cada árbol. En lugar de hacer una sola actualización, CatBoost puede usar múltiples pasos basados en los métodos de gradiente o Newton para mejorar la precisión de los valores de las hojas.

- langevin: Activa el método Stochastic Gradient Langevin Boosting, que introduce ruido en el proceso de optimización para mejorar la generalización del modelo y reducir el sobreajuste.
- bagging_temperature: Controla la aleatoriedad en la selección de muestras para entrenar cada árbol del modelo. Un valor bajo (cercano a 0) hace que la selección de muestras sea más uniforme, lo que da como resultado un modelo más estable, pero con menor diversidad entre los árboles. Un valor alto aumenta la aleatoriedad en la selección de datos, generando árboles más diversos, lo que puede mejorar la capacidad del modelo para capturar patrones complejos, pero también podría aumentar la varianza.
- Semilla (random_seed): Número que establece un punto de partida fijo para la generación de valores aleatorios durante el entrenamiento. Esto permite obtener resultados reproducibles al entrenar el modelo varias veces.
- Número máximo de hojas (max_leaves): Cantidad máxima de hojas que puede tener un árbol. Un número mayor permite modelos más complejos, pero también puede aumentar el riesgo de sobreajuste.
- Mínimo de datos en una hoja (min_data_in_leaf): Número mínimo de muestras de entrenamiento que debe contener cada hoja del árbol. Un valor más alto ayuda a evitar que el modelo aprenda patrones demasiado específicos del conjunto de datos de entrenamiento.
- Función de perdida (loss_function): Métrica utilizada para evaluar el desempeño del modelo durante el entrenamiento. La función elegida define el tipo de problema que se está resolviendo, como clasificación o regresión.
- verbose: Controla la cantidad de información que se muestra en la consola durante el entrenamiento. Valores más altos generan mensajes más detallados sobre el progreso del modelo.

Considerando estos hiperparámetros, se entrenó el modelo con los siguientes valores que se muestran en la Tabla 14

Hiperparámetro	Valor
iterations	230
depth	5
learning_rate	0.08
l2_leaf_reg	15
model_size_reg	13
rsm	0.5
leaf_estimation_iterations	10
langevin	True
bagging_temperature	0.5
random_seed	1234
max_leaves	32
min_data_in_leaf	270
loss_function	Logloss
verbose	True

Tabla 14 *Valores numéricos de los hiperparámetros del modelo*

RESULTADOS

Métricas de evaluación

Ahora presentamos los resultados de nuestro modelo. Primero, mostramos la gráfica de la curva ROC de nuestro modelo entrenado, que mide la capacidad del modelo para distinguir entre clases según Hand & Till (2001), y se encuentra representada en la Figura 9.

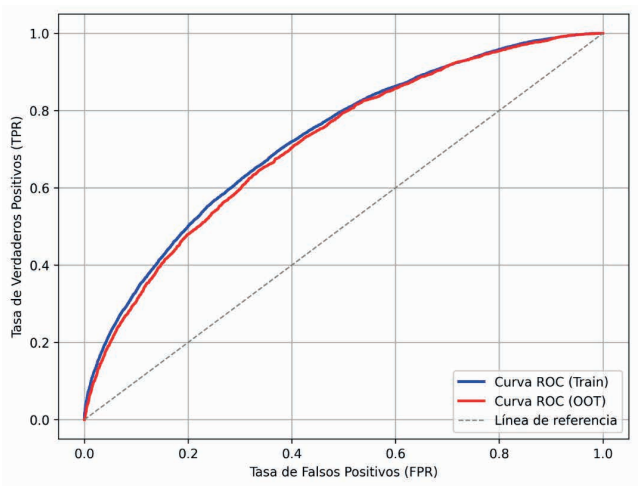


Figura 9 *Curva ROC del modelo de clasificación*

La Figura 9 muestra la curva ROC del modelo. La línea azul representa la muestra de desarrollo y la roja, la muestra de validación. Ambas curvas están por encima de la línea de referencia, lo que indica una buena capacidad discriminatoria del modelo.

A continuación, mostramos los indicadores GINI, AUC y KS sobre la muestra de entrenamiento, también llamada desarrollo, y sobre la muestra de validación, también llamada OOT, representados en la Tabla 15.

Base	GINI	AUC	KS
Desarrollo (Train)	45.14 %	72.57 %	32.28 %
Validación (OOT)	42.84 %	71.42 %	30.84 %

Nota. Se observa una leve disminución de las métricas en la muestra de validación, lo que indica una buena capacidad de generalización del modelo.

Tabla 15 *Métricas de desempeño del modelo en las muestras de desarrollo y validación*

Estos indicadores fueron obtenidos haciendo uso de la librería scipy.stats, específicamente la función ks_2samp para el indicador KS, y la librería sklearn.metrics usando la función roc_auc_score tanto para GINI como AUC.

Tabla de eficiencia

También agregamos, a continuación, la tabla de eficiencia. Esta muestra el poder de discriminación del modelo en el conjunto de datos de validación, organizando los puntajes en quintiles y comparándolos con la proporción de la tasa de malos. Representada en la siguiente Tabla 16.

Rango de score	Promedio	Mínimo	Máximo	Total	Cantidad de malos	Tasa de malos
(0.0, 784]	700.74	181	784	3,624	1,257	34.69 %
(784, 848]	819.83	785	848	3,580	716	20.00 %
(848, 890]	870.44	849	890	3,610	531	14.71 %
(890, 928]	909.48	891	928	3,589	342	9.53 %
(928, 1000]	952.50	929	995	3,537	163	4.61 %

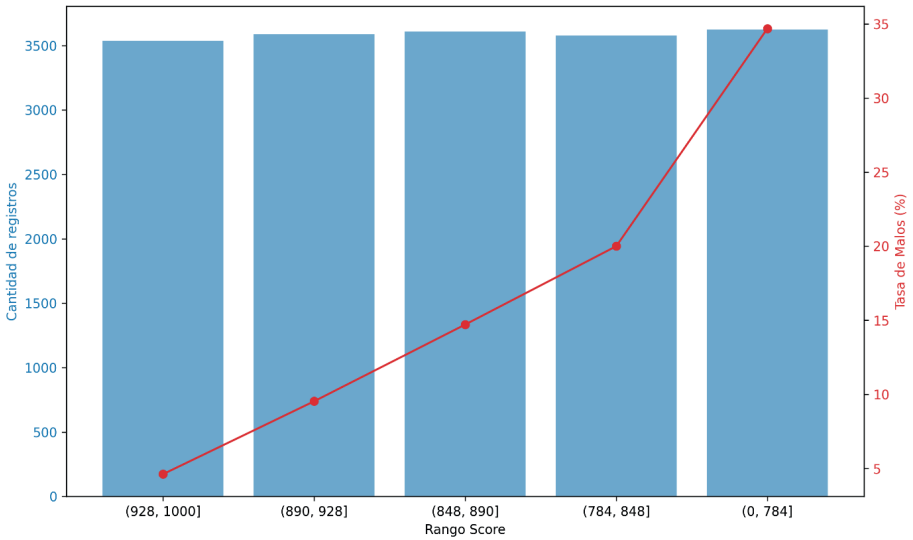
Nota. Se observa una disminución de la tasa de malos conforme aumenta el score, lo que refleja la eficiencia del modelo para discriminar entre buenos y malos.

Tabla 16 *Eficiencia del modelo: distribución de malos según rangos de score*

La Tabla 16 demuestra la capacidad discriminatoria del modelo: en el rango de score más bajo (0-784), la tasa de malos alcanza el 34.69%, mientras que en el rango superior (928-1000) se reduce al 4.61%. Esto confirma la correlación negativa entre el score y el riesgo de default.

En el rango superior (score de 928 a 1000), la tasa de malos desciende al 4.61%. Esto indica que, para registros con puntajes elevados, el modelo predice que la mayoría tendrá un buen comportamiento financiero, como era de esperarse.

A continuación, podemos observar esta correlación negativa entre puntaje y tasa de malos, evidenciando que la tasa de malos aumenta a medida que el score disminuye, como se muestra en la Figura 10.



Nota. La barra representa la cantidad de registros en cada rango de score, mientras que la línea roja muestra la tasa de malos

Figura 10 Distribución de registros por rango de score y tasa de malos

En resumen, nuestro modelo tiene un buen poder predictivo, ya que asigna un score bajo a quienes presentan un comportamiento moroso y un score alto a quienes no.

Importancia de variables

En este apartado, mostramos cuáles de las 24 variables utilizadas en el modelo son las más influyentes en la predicción. Para ello, nos basamos en el concepto de los valores SHAP, que nos permiten medir la contribución de cada variable en la toma de decisiones del modelo.

Además, presentamos una tabla que organiza las variables en un ranking de importancia, es decir, en orden de relevancia según su impacto en las predicciones. Esto nos permitirá identificar cuáles son los factores clave que el modelo considera al clasificar a los solicitantes.

- **Importancia de Variables por SHAP value mean**

SHAP (SHapley Additive exPlanations) es un método basado en la teoría de valores de Shapley, que se usa para explicar el impacto de cada variable en la predicción de un modelo. Su objetivo es distribuir de manera justa la contribución de cada variable en la decisión final del modelo, similar a cómo en un juego cooperativo se reparte el valor generado entre los jugadores según su participación.

El concepto subyacente a la importancia de las características SHAP es sencillo: aquellas características con valores de Shapley de magnitud elevada son más relevantes. Para obtener la importancia a nivel global, calculamos el promedio de los valores absolutos de Shapley de cada característica en el conjunto de datos (Lundberg & Lee, 2017).

La importancia de las características SHAP es una alternativa a la importancia de las características de permutación, pero funcionan de manera diferente. La importancia de las características de permutación mide cuánto empeora el desempeño del modelo cuando una variable se altera al azar. Es decir, si quitamos o desordenamos una variable importante, el modelo hará peores predicciones, lo que indica que dicha variable era relevante.

Por otro lado, SHAP mide la contribución de cada variable en cada predicción individual y luego promedia estas contribuciones. En otras palabras, en lugar de ver cuánto empeora el modelo al quitar una variable, SHAP nos dice cuánto aporta cada variable en cada caso.

Ejemplo 1. Imaginemos que tenemos un modelo que predice si una persona aprobará un préstamo. Supongamos que una de las variables es el historial de pagos. Si al eliminar o alterar esta variable el modelo empieza a equivocarse más, significa que era una variable importante (esto sería la importancia por permutación). En cambio, SHAP analizaría cuánto peso tiene esta variable en cada decisión del modelo y sacaría un promedio, lo que nos diría qué tan relevante es en general.

A continuación, mostramos la Figura 11 de importancia de variables.

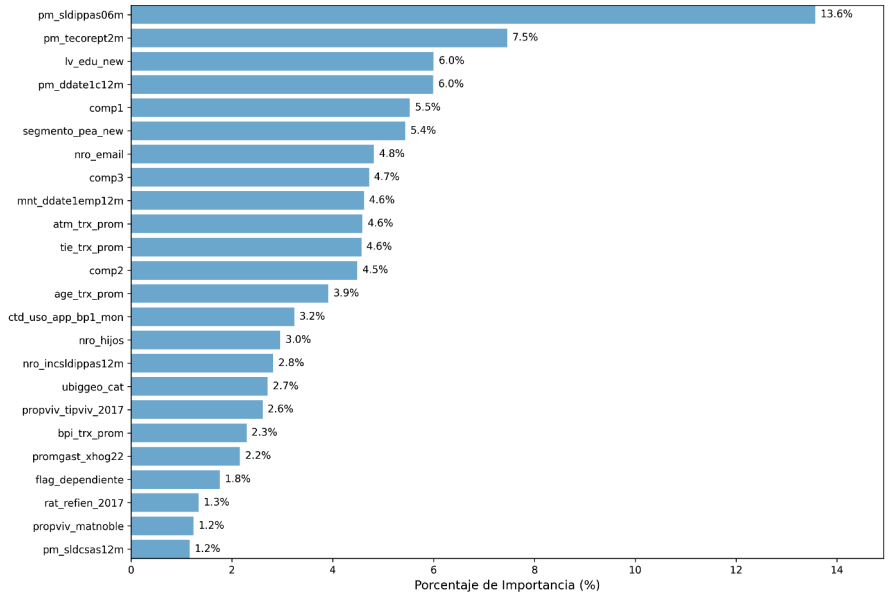


Figura 11

Ranking de importancia de variables por SHAP value

Nota. El porcentaje de importancia indica la contribución relativa de cada variable en el modelo predictivo

CONCLUSIÓN

Este trabajo desarrolló y evaluó un modelo predictivo de riesgo crediticio utilizando técnicas de aprendizaje automático, con un enfoque en su capacidad de discriminación.

Los principales hallazgos de este estudio se resumen en los siguientes puntos:

- Se desarrolló un modelo basado en el algoritmo CatBoost, que mostró un desempeño competitivo según métricas estándar de clasificación, como el coeficiente GINI, el área bajo la curva ROC (AUC) y el estadístico de Kolmogorov-Smirnov (KS) (ver Tabla 15).
- Se identificaron las variables más influyentes en la predicción del riesgo crediticio mediante valores SHAP (ver Figura 11), lo que permitió una interpretación transparente del impacto de cada factor en la clasificación de los solicitantes.
- Los resultados demostraron que el modelo diferencia eficazmente entre clientes “buenos” y “malos” (ver Tabla 16), asignando mayores probabilidades de incumplimiento a aquellos con mayor riesgo financiero.
- Se comprobó que la tasa de clientes con incumplimiento es inversamente proporcional al puntaje asignado por el modelo (ver Figura 10), lo que confirma su capacidad predictiva.

REFERENCIA

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2), 171–186. https://doi.org/10.1023/A:1010920819831

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.