ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

TÍTULO DE LA INVESTIGACIÓN

PREDICCIÓN DE INCUMPLIMIENTO DE PRÉSTAMOS PARA CLIENTES DE ENTIDADES BANCARIAS

Profesor

Ing. Manuel Eugenio Morocho Cayamcela. MSc, PhD

Autores

Ing. Ronald Stalin Cedeño Mera

Ing. Santiago Fernando Pacheco Estrella

2024

RESUMEN

La administración de riesgos crediticios es crucial para el funcionamiento de las instituciones financieras. Los avances en este campo han permitido desarrollar estrategias cada vez más efectivas para evaluar el riesgo. Los modelos de scoring y la implementación de sistemas seguros para procesar pagos son fundamentales. Las variables cuantitativas (historia crediticia, ingreso, patrimonio) y cualitativas (experiencia laboral, estabilidad residencial) son importantes para la toma de decisiones. Nuestro análisis utilizando SMOTE presenta mejoras en la precisión del modelo base e identifica las variables claves a evaluar. La combinación de estas por su naturaleza (cuantitativas y cualitativas) son esenciales para unos resultados eficaces.

ABSTRACT

Credit risk management is crucial for the functioning of financial institutions. Advances in this field have allowed the development of increasingly effective strategies for assessing risk. Scoring models and the implementation of secure systems to process payments are essential. Quantitative variables (credit history, income, assets) and qualitative variables (work experience, residential stability) are important for decision making. Our analysis using SMOTE presents improvements in the accuracy of the base model and identifies the key variables to evaluate. The combination of these, due to their nature (quantitative and qualitative), is essential for effective results.

Proyecto y Código Fuente:

Cedeño & Pacheco. (2024, July 21). *Sanfer1ec/ProyectoUdla: LoanDefault*. GitHub. https://github.com/sanfer1ec/ProyectoUdla

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Los líderes bancarios en el Ecuador inherentemente se ven en la obligación de mejorar las prácticas de administración de riesgos.

La administración de riesgos crediticios es un aspecto crucial en la gestión de instituciones financieras, ya que permite evaluar y mitigar los riesgos asociados con la concesión de créditos. En este contexto, los avances en la administración de estos han permitido desarrollar estrategias más efectivas para evaluar y tener las claves para tratar con los diferentes clientes.

Los modelos analíticos de scoring, son modelos de puntuación crediticia que juegan un papel fundamental para proporcionar una evaluación objetiva y precisa del riesgo. La implementación de sistemas seguros y eficientes para procesar pagos son importantes, pero más importante aún es tener una analítica que permita anticipar el futuro, con herramientas que minimicen las pérdidas y maximicen la rentabilidad.

En la toma de decisiones, las variables cuantitativas, como la historia crediticia, el ingreso y el patrimonio, son fundamentales. Sin embargo, las variables cualitativas, como la experiencia laboral y la estabilidad residencial, también tienen incidencia y peso a la hora de tomar una resolución.

En nuestro análisis exploratorio de datos, tomamos una base de datos de una institución financiera siguiendo los parámetros de preprocesamiento y tratamiento de datos, como plus de nuestra estrategia utilizamos SMOTE (Synthetic Minority Over-sampling Technique) para equilibrar la clase minoritaria y mejorar la precisión. En el estudio utilizamos como eje la correlación con la variable status como objetivo lo que permitió identificar los hallazgos más importantes para la toma de decisiones.

Algunos procesos de optimización descubren en el conjunto de datos la correlación que aporta a mejorar la precisión de los modelos ya estudiados.

I. REVISIÓN DE LITERATURA

Avances en la administración de riesgos crediticios

Los avances en las tecnologías de información en la administración del riesgo crediticio permiten que las entidades financieras crediticias automaticen las decisiones sobre aceptación o rechazo de una solicitud de crédito. y la administración de una cartera crediticia (clientes consolidados) en un *cross-sell* (venta cruzada). Hace unos años, dicha administración crediticia se realizaba solo con la experiencia o percepción del ejecutivo. Ahora, uno de los modelos más usados para la evaluación de créditos es el *scoring model* (modelo de calificación), el cual determina un *score* (puntaje) para clientes que solicitan un crédito identificando a aquellos que tienen la posibilidad de incumplir con sus pagos. La literatura, con respecto de la calificación crediticia, es amplia, y basta mencionar los modelos de Rosenberg y Gleit (1994), Merton (1974), Hand y Jacka (1998), Thomas, Crook y Edelman (1992) y Lewis (1992), mientras que para calificaciones de comportamiento o *behavioral scoring* se cuenta con el trabajo de Mays (1998) (Trejo García et al., 2017).

Modelos analíticos de scoring

Existen en la literatura especializada un conjunto amplio de métodos y técnicas cuantitativas para predecir la probabilidad de que un cliente falle o incumpla y, por consiguiente, no se recupere el crédito otorgado por alguna institución financiera. Los modelos *scoring* son herramientas que utilizan la clasificación de los solicitantes o clientes consolidados por nivel de riesgo con base en el suministro de la información de clientes en las solicitudes de crédito y comportamiento de pagos (Trejo García et al., 2017).

Las tarjetas de crédito son muy importantes en la evolución y desarrollo de la economía, ya que éstas representan un mecanismo de crédito y una posibilidad de incrementar el nivel de ventas; esto quiere decir que es un beneficio tanto para los consumidores, como para las empresas (Harry & Torres, n.d.).

Las tarjetas de crédito sirven para múltiples propósitos: opción de pago, una fuente de crédito renovable un modo fácil de usar el pago. En la literatura se han discutido diferentes funciones que tiene el sistema financiero en la dinámica de una economía, como el reducir los costos de transacción, lo que moviliza recursos a aquellos proyectos más eficientes y productivos, o ayudar a mitigar y distribuir los riesgos asociados con proyectos individuales, industrias, regiones y países, mejorando la distribución de recursos. Además, el sistema financiero favorece el crecimiento de la productividad y el cambio tecnológico mediante la movilización de ahorros a proyectos productivos– lo que lleva a una mayor acumulación de capital físico y humano y, por consiguiente, a un mayor crecimiento económico (Acosta Mellado, Murillo Félix, & Almeida, 2021).

La administración del riesgo crediticio durante las últimas décadas ha sido una de las áreas con mayor crecimiento. Las técnicas de calificación más utilizadas para la administración del riesgo crediticio han sido el *credit scoring* (otorgamiento de crédito) y el *behavioral scoring* (comportamiento crediticio), así como varias herramientas para la estimación del riesgo financiero con relación a los préstamos o financiamientos al menudeo. En el mercado mexicano crediticio al consumo se tiene la siguiente clasificación en tres tipos de *pools* (carteras): créditos revolventes, créditos personales y créditos a la vivienda. Todos los solicitantes de crédito, así como los clientes consolidados (Trejo García et al., 2017).

Durante los años noventa, la aplicación del credit scoring, estaba centrada sobre todo en la evaluación de créditos hipotecarios y tarjetas de crédito (Mester, 1997). Sin embargo, la utilidad y aplicación de este método ha ido evolucionando

en el tiempo, y en la actualidad, muchas de las instituciones financieras y no financieras, utilizan el credit scoring para evaluar a sus posibles clientes. Pueden evaluar las solicitudes de los distintos tipos de créditos: personales, para negocios, hipotecarios; e inclusive, el otorgamiento de una tarjeta de crédito (Cruz y Villalta, 2017). El credit scoring presenta una alta capacidad de predicción, lo que puede permitir realizar un perfeccionamiento en el desarrollo de la evaluación crediticia (Schreiner, 2002).

El Credit Scoring tiene dos perspectivas: una perspectiva personal porque se centra en las particularidades propias de cada solicitante, es decir, en las cualidades personales del individuo; y tiene una perspectiva desde el ámbito financiero, ya que, se busca obtener información sobre préstamos u otros instrumentos de crédito pasados, en donde se pueda verificar su nivel de compromiso y responsabilidad en el cumplimiento de las condiciones del crédito (Harry & Torres, n.d.).

Entre las causas que restringen su desarrollo y crecimiento, las limitaciones a las fuentes de financiamiento formales que permitan un desarrollo adecuado de sus operaciones , Lo anterior, es debido a la falta o escaza información financiera provista por estas empresas según criterios contable financieros que permitan evaluar su capacidad de ser sujetos de crédito, como también por no disponer de activos de calidad que garanticen sus obligaciones ni disponer de un historial financiero que permita evaluar su capacidad de pago (Leal ET al, 2018).

Comúnmente los agentes económicos adoptan perfiles aversos al riesgo, motivo por el cual el accionar orientado a minimizar, transferir y/o mitigar los riesgos conducía a los bancos a rechazar aquellas operaciones que no ofrecían plenas garantías, la gestión moderna del riesgo de crédito establece como objetivo gestionar el riesgo de crédito para obtener una rentabilidad acorde con un nivel de pérdida esperada asumida, comprometiendo para ello una porción de su capital propio en cumplimiento de la normativa. Esto significa que una

operación crediticia con una mayor probabilidad de impago no necesariamente tiene que ser mal negocio, debe obtener una rentabilidad mayor que compense el riesgo de crédito asumido (Mostajo Castelú & Vargas Sánchez, 2015).

Este proceso de transformación de activos y pasivos o, en otras palabras, este proceso de intercambio (compra - venta) de riesgos en el que participan las entidades financieras está sujeto a una variedad de riesgos financieros y operativos. Uno de ellos el riesgo de crédito o de contraparte (counterparty risk), inherente a la gestión de carteras que tienen cuentas pendientes de cobro (Bambino Contreras & Morales Oñate, 2023).

La educación financiera y el uso de las tarjetas de crédito se encuentran correlacionadas mediante un análisis de varios factores, entre los más importantes se encuentran, el nivel de ingresos comparado con los egresos, determinantes al momento de evaluar la capacidad de pago del cliente, la edad de la persona, estabilidad laboral, historial crediticio, entre otros; por ello se ha visto necesario la implementación de una tarjeta de crédito dirigida a un grupo estratégico de personas, cuyos clientes potenciales sean todos los que no mantengan historial crediticio o incluso hayan tenido alguno negativo (Ormaza Andrade & Cevallos Jiménez, 2021).

Como parte de la gestión de riesgo crediticio es esencial conceptualizar la gestión de crédito, para conocer los mecanismos que se realizan ante un otorgamiento o disposición de crédito, además del control y manejo de este, como también la gestión de cobranza que es necesaria para evadir un riesgo de morosidad, por lo que las políticas de la empresa son importantes (Zambrano Molina, 2021).

Tabla 1. Matriz de investigaciones similares

Tema	Definición	Tipos de Datos	Metodologías Utilizadas	Resultados	Implicaciones Gerenciales
Modelos analíticos de scoring	Existe una variedad de modelos propuestos para evaluar el riesgo de crédito, y destacan entre estos los modelos de credit scoring los cuales proponen automatizar el proceso de gestión de créditos en cuanto a conceder o no una determinada operación crediticia sujeto a un conjunto de variables relevantes de decisión (Leal Et al., 2018).	Permitir la generación de alertas oportunas y consistentes con el fin de evitar el incumplimiento, y disminución de su probabilidad de ocurrencia, y gestionar adecuadamente este riesgo (Tulcanaza Aguilar, 2021).	Pueden ayudar a obtener la probabilidad de incumplimiento de un solicitante de crédito (Montalván Acaro, 2019).	El método de scoring para la selección de proyectos no debe considerarse una actividad aislada y sin ninguna conexión integral con los resultados de los proyectos seleccionados (Salazar Et al., 2019).	Optimizar y mejorar la gestión de riesgo, la optimización de la cartera de crédito, las mejoras operativas como automatización del proceso de aprobación de crédito, el cumplimiento regulatorio y una mejor experiencia para el cliente
Tarjetas de crédito	Se define también como un producto financiero emitido por una institución financiera bancaria, financiera no bancaria o entidades no financieras como las tiendas comerciales,	Servir como referencia crediticia al momento de adquirir un crédito y te permite controlar tu presupuesto si guardas los váuchers para administrar tus gastos (Caraguay Muñoz, Pesantez León, &	Poder acceder a dinero en efectivo de manera rápida inmediata solicitando los adelantos que se necesite. Los tiempos convenidos para el pago de los cuales pueden variar de	Desventajas al operar con tarjeta de crédito: inciden en el incremento de gastos, ya que el sistema estimula el consumo; el alto pago de intereses y gastos de administración que el usuario debe pagar en beneficio del sistema; la alta posibilidad de fraude, robo o pérdida de la tarjeta	

	destinadas a la adquisición de bienes o servicios para satisfacer necesidades (Cevallos Jiménez & Ormaza Andrade, 2021).	Elizalde Orellana, 2020).	acuerdo con la modalidad del sistema. Control de gastos ya que permite al usuario acceso a desgloses de orden financiero y tributario (Bayas Sánchez , 2020).	(Caraguay Muñoz, Pesantez León, & Elizalde Orellana, 2020).
Riesgo Crediticio	El riesgo de crédito representa la probabilidad de pérdida en la que incurre la empresa en el caso de una falla del socio comercial (Lapo Et al.,2021).	Maximizar el valor económico de la institución financiera, en un contexto de incertidumbre, además como objetivos de la gestión de riesgos se tiene: buscar el desarrollo y la estabilidad económica de la entidad, preservar el sistema de pagos y principalmente identificar, monitorear y mitigar los riesgos (Quindigalle Cuyo , 2018).	Al obtener ciertos riesgos en una empresa podría tener un alto desempeño, de tal forma, crear mejores ingresos, lo que lleva a su aumento y expansión (Elizalde Chapa, 2023).	Si el riesgo financiero no se tramita en el momento adecuado y con las estrategias consideradas, pueden lograr causar daños a las finanzas y la reputación de la entidad (Elizalde Chapa, 2023).

Defaults Crediticio	Funcionan como un instrumento de protección, donde, a cambio del pago de una prima, se da una protección al vendedor que promete pagar al comprador una garantía en caso de un evento de *default* (impago) u otro evento adverso para un contrato crediticio (Martínez Arroyo & Marín Rodríguez, 2021).	Estos modelos deben ser utilizados en la toma de decisiones de las entidades financieras, como en la admisión de clientes o en decisiones de negocio (Venosi , 2021).	El uso de derivados de crédito mejoraría la profundidad del mercado y, consecuentemente, reduciría la asimetría de información de los mercados financieros (Martínez Arroyo & Marín Rodríguez, 2021).	El riesgo de pérdida determina la pérdida como una fracción de la exposición al momento del default. Este factor se conoce como la pérdida dado el default o, por sus siglas en inglés, LGD (Venosi , 2021).

Fuente: Elaboración propia

II. IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Las economías en vías de desarrollo asientan su cuerpo financiero en la banca, sea esta pública o privada. Estas entidades son un conjunto único de empresas comerciales cuyas restricciones reglamentarias, funciones económicas y operaciones los convierten en un tema importante de investigación (Lapo Maza, Tello Sánchez, & Mosquera Camacas, 2021).

El riesgo de crédito representa la probabilidad de pérdida en la que incurre la empresa en el caso de una falla del socio comercial. Kliestik y Cug (2015), resulta esencial exponer diversos enfoques de riesgo crediticio que posteriormente han sido aplicados y mejorados para la valoración del riesgo específicamente en la industria bancaria.

En el Ecuador, 7 de los bancos con activos de más de 1.000 millones en conjunto poseen aproximadamente el 88% de activos totales y cartera de créditos, en depósitos poseen cerca del 89% del total del sector bancario (Superintendencia de Bancos, 2023).

El uso masivo de productos de créditos en todas sus características ha incrementado la preocupación por la seguridad de las transacciones y la protección contra el fraude, tanto en pagos en línea como en transacciones presenciales. Igualmente, el fácil acceso al financiamiento con tarjetas de crédito ha llevado a un aumento en el endeudamiento de los usuarios, lo que resalta la necesidad de una mayor educación financiera sobre el uso responsable del crédito.

La implementación de sistemas seguros y eficientes para procesar pagos, tanto en línea como en puntos de venta, requiere una infraestructura y tecnológica adecuada que pueda mantener al día las demandas del mercado.

La información recopilada y analizada por las entidades financieras puede resultar insuficiente para predecir de manera precisa y oportuna el comportamiento futuro de los usuarios. Esta falta de certeza plantea desafíos significativos para las entidades, que deben gestionar de manera efectiva el riesgo crediticio asociado con sus clientes, minimizando la morosidad y el impacto negativo en sus operaciones financieras.

En este contexto, es fundamental investigar las metodologías actuales de análisis de clientes utilizadas por las entidades proveedoras de crédito, para identificar deficiencias y áreas de mejora. Además, resulta imperativo explorar nuevas estrategias y herramientas que permitan a estas entidades mejorar la precisión de sus análisis y reducir la incertidumbre relacionada con el comportamiento crediticio de sus clientes. De esta manera, se contribuirá a una gestión más efectiva del riesgo y a una mejor atención de las necesidades financieras de sus clientes.

III. PLANTEAMIENTO DEL PROBLEMA

La creciente poblacional de consumidores y la sistematización de procesos de pagos en línea y pagos presenciales, hacen más frecuentes el uso de productos de créditos para los usuarios. Las entidades proveedoras de estos, aunque poseen metodologías de análisis de las características de los clientes, se pueden encontrar con algunos niveles de incertidumbre respecto a los usuarios que necesitan este producto – servicio, ya que cierta información que poseen no puede solucionar el tema de clientes impagos o que puedan caer en Default.

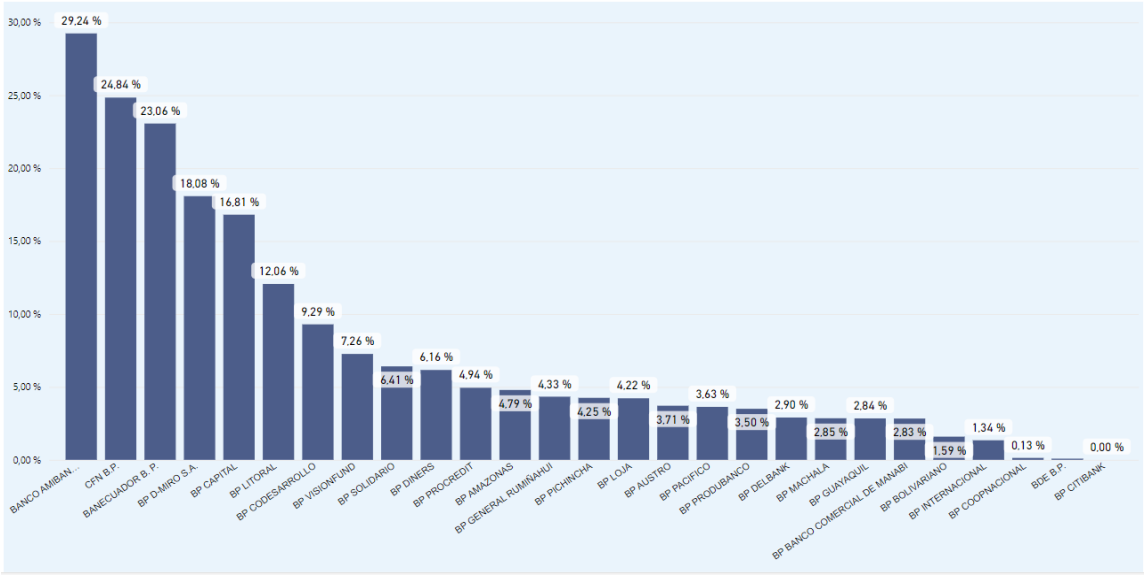
Las Pymes, son organizaciones empresariales que, por las características mismas de su concepción, son las que más fácilmente pueden adaptarse a los cambios; son el factor fundamental en el país de generar empleo (Lapo Maza, Tello Sánchez, & Mosquera Camacas, 2021).

En la realidad económica y comercial de nuestro medio, las organizaciones financieras en general están expuestas a riesgos; los niveles estratégicos y tácticos de las entidades financieras trabajan constantemente para buscar estrategias que disminuyan los riesgos crediticios empleo (Lapo Maza, Tello Sánchez, & Mosquera Camacas, 2021).

Los líderes bancarios en el Ecuador inherentemente se ven en la obligación de mejorar las prácticas de administración de riesgos.

A continuación, se muestra un gráfico del índice de morosidad bruta de la banca en el Ecuador. Lo que en algunos casos afecta de manera significativa a la liquidez y a la estructura de capital de esta.

Gráfico 1.Índice de Morosidad Bruta de la banca pública y privada del Ecuador



Fuente: Superintendencia de Bancos del Ecuador.

La cartera bruta se refiere al total de la cartera de crédito de una institución financiera sin deducir la provisión para créditos incobrables. La cartera improductiva son aquellos préstamos que no generan renta financiera a la institución, están conformados por la cartera vencida y la cartera que no devenga interés.

Por tal razón el índice de morosidad bruta mide la cartera improductiva frente a la cartera bruta, mientras mayor es el indicador, significa que las entidades están teniendo problemas en recuperar su cartera.

El índice de morosidad de ban Ecuador es del 23,06 % y de la CFN el 24,84 % (ambas bancas públicas), tienen un alto índice de morosidad con relación a los principales bancos privados del país, problema que se puede disminuir aplicando algunos métodos planteados en este estudio.

Dentro de la presente investigación se pretende resolver mediante diferentes métodos de analítica de datos un problema de aplicación y análisis de los perfiles de actuales y potenciales clientes que están activos dentro del sistema financiero, por lo que el contexto de la presente es disciplinar y tecnológico ya que aborda todas las fuentes de información posibles que se puedan estructurar dentro de una base de datos.

A partir de este planteamiento pretendemos resolver los siguientes problemas:

¿Cuál es la probabilidad de que utilizando métodos de regresión logística se pueda determinar con certeza la categoría de clientes prospectos de productos de crédito que puedan caer en default?

¿Cómo aplicando metodologías de Machine Learning, a través de algoritmos de árboles de decisión podremos enseñarles a las máquinas a tomar decisiones, para resolver problemas de regresión o de clasificación?

A partir de estas y otras preguntas que surjan en este proyecto de investigación resolveremos el problema, enmarcado en una investigación experimental con base en investigaciones de fuentes secundarias.

Objetivo General

Mediante la analítica predictiva conocer las variables de clientes potenciales de créditos que pudieran caer en default y proponer el modelo óptimo para la toma de decisiones en las instituciones financieras del Ecuador.

Objetivos Específicos

- Mediate el estudio de fuentes secundarias conocer las perspectivas de investigación del scoring por el que se sostiene la economía actual.

- Realizar un análisis exploratorio de datos de usuarios de productos de créditos de una base de datos de una institución financiera.

- Aplicar la técnica estadística del SMOTE para conocer su incidencia sobre el modelado base del presente estudio.

- Conocer cuáles son las variables (categóricas o numéricas) por las que se deben tomar las decisiones para otorgar créditos.

- Utilizando métodos de regresión logística y Machine Learning presentar una propuesta para automatizar la toma de decisiones en el otorgamiento de productos de crédito.

IV. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

1. Recolección de datos

El dataset contiene:

- 21 columnas tipo texto.
- 8 columnas decimal.
- 4 columnas tipo entero.
- De un total de 148.000 registros.

El ingente conjunto de datos consta de múltiples factores deterministas como los ingresos del prestatario, el género, el propósito del préstamo, etc.
El presente dataset está sujeto a una fuerte multicolinealidad y valores vacíos que serán necesario tratarlos para crear un clasificador sólido para predecir los resultados planteados en los objetivos.

La base de datos tomada de la Plataforma Kaggle, de título “Loan Default Dataset” (H, 2021).

2. Limpieza, preparación y procesamiento de datos

Para la limpieza de datos se procedió a modificar y eliminar campos de datos incompletos e incorrectos, se identificó y eliminó información duplicada y datos sin relación. Los valores nulos y vacíos de cada columna se imputaron con información de la media de sus valores porque el tamaño del dataset permite hacerlo sin que haya variaciones en sus resultados.

a. Tipos de variables

Tabla 2. Variables Cuantitativas

Variables cuantitativas	Detalle
Loan Amount	Valor total del préstamo
rate of interest	Tasa de interés
Interest_rate_spread	Diferencia tipo de interés respecto a la referencia
Upfront_charges	Cargos iniciales del préstamo
property value	Valor de la propiedad asociado al préstamos
income	Ingresos del solicitante del préstamo
credit score	Puntaje de crédito del solicitante
LTV (Loan to Value)	Relación monto del préstamo y valor de la propiedad
dtir1 (Debt to Income Ratio)	Relación deuda/ingresos del solicitante.

Fuente: Elaboración propia.

Tabla 3. Variables Cualitativas

Variables cualitativas	Detalle
loan_limit	Categoría de límite de préstamo
Gender	Género del solicitante del préstamo
approv_in_adv	Indica si el préstamo ha sido preaprobado.
loan_type	tipo de préstamo
loan_purpose	Objeto del préstamo
open_credit	Indica si el solicitante tiene otros créditos abiertos
construction_type	Tipo de construcción del inmueble
occupancy_type	Tipo de ocupación de la propiedad
Secured_by	Tipo de garantía de préstamo
age	Rango de edad del solicitante
Region	Región geográfica del préstamo
Security_Type	Tipo de garantía de préstamo

Fuente: Elaboración propia.

Tabla 4. Procesamiento de datos

```
df.head( )
✓ 0.0s                                                                    Python
```

	ID	year	loan_limit	Gender	approv_in:adv	loan_type	loan_purpuse	Credit_Worthiness	open_credit	business_or_commerc
0	24890	2019	cf	Sex Not Available	nopre	type1	p1	l1	nopc	nob
1	24891	2019	cf	Male	nopre	type2	p1	l1	nopc	b
2	24892	2019	cf	Male	pre	type1	p1	l1	nopc	nob
3	24893	2019	cf	Male	nopre	type1	p4	l1	nopc	nob
4	24894	2019	cf	Joint	pre	type1	p1	l1	nopc	nob

Fuente: Elaboración propia.

b. Conversión de Columnas con tipos de datos Texto a numéricos

Se convierte estas columnas a formato numérico usando pd.to_numeric. Si hay valores no numéricos, los manejaremos adecuadamente (por ejemplo, reemplazándolos con NaN).

Tabla 5. Limpieza de datos

Porcentaje de valores faltantes por columna:

ID	0.000000
year	0.000000
loan_limit	2.249277
Gender	0.000000
approv_in_adv	0.610749
loan_type	0.000000
loan_purpose	0.090133
Credit_Worthiness	0.000000
open_credit	0.000000
business_or_commercial	0.000000
loan_amount	0.000000
rate_of_interest	24.509989
Interest_rate_spread	24.644515
Upfront_charges	26.664425
term	0.027578
Neg_ammortization	0.081388
interest_only	0.000000
lump_sum_payment	0.000000
property_value	10.155378
construction_type	0.000000
occupancy_type	0.000000
Secured_by	0.000000
total_units	0.000000
income	6.154571
credit_type	0.000000
Credit_Score	0.000000
co-applicant_credit_type	0.000000
age	0.134526
submission_of_application	0.134526
LTV	10.155378
Region	0.000000
Security_Type	0.000000
Status	0.000000
dtir1	

Conteo de valores únicos por columna:

ID	148670
year	1
loan_limit	2
Gender	4
approv_in_adv	2
loan_type	3
loan_purpose	4
Credit_Worthiness	2
open_credit	2
business_or_commercial	2
loan_amount	211
rate_of_interest	131
Interest_rate_spread	22516
Upfront_charges	58271
term	26
Neg_ammortization	2
interest_only	2
lump_sum_payment	2
property_value	385
construction_type	2
occupancy_type	3
Secured_by	2
total_units	4
income	1001
credit_type	4
Credit_Score	401
co-applicant_credit_type	2
age	7
submission_of_application	2
LTV	8484
Region	4
Security_Type	2
Status	2
dtir1	57

dtype: int64

16.224524
dtype: float64

Fuente: Elaboración propia.

Tabla 6. Preparación y procesamiento de datos

Datos después de eliminar columnas

#	Column	Non-Null Count	Dtype
0	ID	148670 non-null	int64
1	loan_limit	145326 non-null	object
2	Gender	148670 non-null	object
3	approv_in_adv	147762 non-null	object
4	loan_type	148670 non-null	object
5	loan_purpose	148536 non-null	object
6	Credit_Worthiness	148670 non-null	object
7	open_credit	148670 non-null	object
8	business_or_commercial	148670 non-null	object
9	loan_amount	148670 non-null	int64
10	rate_of_interest	112231 non-null	float64
11	Interest_rate_spread	112031 non-null	float64
12	Upfront_charges	109028 non-null	float64
13	term	148629 non-null	float64
14	Neg_ammortization	148549 non-null	object
15	interest_only	148670 non-null	object
16	lump_sum_payment	148670 non-null	object
17	property_value	133572 non-null	float64
18	construction_type	148670 non-null	object
19	occupancy_type	148670 non-null	object
20	Secured_by	148670 non-null	object
21	total_units	148670 non-null	object
22	income	139520 non-null	float64
23	credit_type	148670 non-null	object
24	Credit_Score	148670 non-null	int64
25	co-applicant_credit_type	148670 non-null	object
26	age	148470 non-null	object
27	submission_of_application	148470 non-null	object
28	LTV	133572 non-null	float64
29	Region	148670 non-null	object
30	Security_Type	148670 non-null	object
31	Status	148670 non-null	int64
32	dtir1	124549 non-null	float64

dtypes: float64(8), int64(4), object(21)

Datos después de llenar valores faltantes

#	Column	Non-Null Count	Dtype
0	ID	148670 non-null	int64
1	loan_limit	148670 non-null	object
2	Gender	148670 non-null	object
3	approv_in_adv	148670 non-null	object
4	loan_type	148670 non-null	object
5	loan_purpose	148670 non-null	object
6	Credit_Worthiness	148670 non-null	object
7	open_credit	148670 non-null	object
8	business_or_commercial	148670 non-null	object
9	loan_amount	148670 non-null	int64
10	rate_of_interest	148670 non-null	float64
11	Interest_rate_spread	148670 non-null	float64
12	Upfront_charges	148670 non-null	float64
13	term	148670 non-null	float64
14	Neg_ammortization	148670 non-null	object
15	interest_only	148670 non-null	object
16	lump_sum_payment	148670 non-null	object
17	property_value	148670 non-null	float64
18	construction_type	148670 non-null	object
19	occupancy_type	148670 non-null	object
20	Secured_by	148670 non-null	object
21	total_units	148670 non-null	object
22	income	148670 non-null	float64
23	credit_type	148670 non-null	object
24	Credit_Score	148670 non-null	int64
25	co-applicant_credit_type	148670 non-null	object
26	age	148670 non-null	object
27	submission_of_application	148670 non-null	object
28	LTV	148670 non-null	float64
29	Region	148670 non-null	object
30	Security_Type	148670 non-null	object
31	Status	148670 non-null	int64
32	dtir1	148670 non-null	float64

dtypes: float64(8), int64(4), object(21)

Fuente: Elaboración propia.

c. Eliminación de Columnas No Útiles

Identificaremos las columnas que no contribuyen a nuestro análisis o entrenamiento del modelo. Estas podrían ser columnas con un solo valor único, un alto porcentaje de valores faltantes, o información irrelevante por ejemplo los identificadores

Columnas con un solo valor único: Estas columnas no proporcionan variabilidad y, por tanto, no son útiles para el análisis o modelado, en este ejemplo la columna “ID” y “Year”.

Columnas con Alto Porcentaje de Valores Faltantes (>50%): Imputar estas columnas puede introducir sesgo significativo o ruido, por lo que es mejor eliminarlas., en este caso no tenemos columnas que cumplan con esta limitación.

Columna ID: Esta columna contiene valores únicos para identificar cada fila del conjunto de datos, pero no aporta información útil para la predicción del objetivo. Su presencia podría añadir ruido y afectar el rendimiento del modelo.

Columna Year: Esta columna tiene solo un valor por lo que no resulta relevante.

d. Manejo de Valores Faltantes

Los valores faltantes se pueden manejar de diversas maneras según la naturaleza de los datos y su extensión. Las estrategias comunes incluyen:

Eliminar filas o columnas con valores faltantes.

- Columnas Numéricas: Llenar valores faltantes con la mediana es robusto ante valores atípicos y preserva la distribución de los datos.
- Columnas Categóricas: Llenar valores faltantes con la moda (valor más frecuente) asegura que la imputación sea consistente con la categoría más común en los datos.

e. Renombrar columnas para mejor entendimiento

Las columnas que tienen abreviaciones se proceden a renombrar para tener una mejor visualización en gráficos o asociar una etiqueta a valores comunes de mejor entendimiento.

- LTV se renombra por LoanToValueProperty
- Dtir1 se renombra por DebTo IncomeRatio

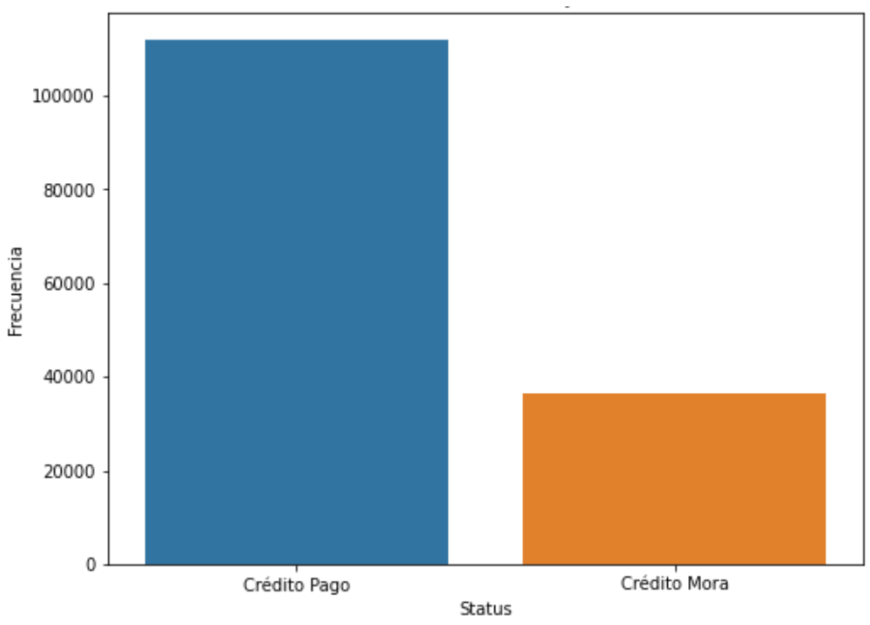
f. Identificación de Variable Objetivo

Predicción del Estado del Préstamo: La variable “Status” es crucial porque representa el resultado que estamos tratando de predecir. En un análisis de riesgo de crédito, por ejemplo, predecir si un préstamo será aprobado o caerá en incumplimiento (Default) es fundamental.

Evaluación de Modelos: Al construir modelos de clasificación, la precisión y otras métricas de rendimiento se evaluarán en función de la capacidad del modelo para predecir correctamente los valores de “Status”.

Variable Objetivo “Status” encontrada:
Valores únicos en “Status”: [1,0]

Gráfico 2. Distribución de la Variable Objetivo (Status)

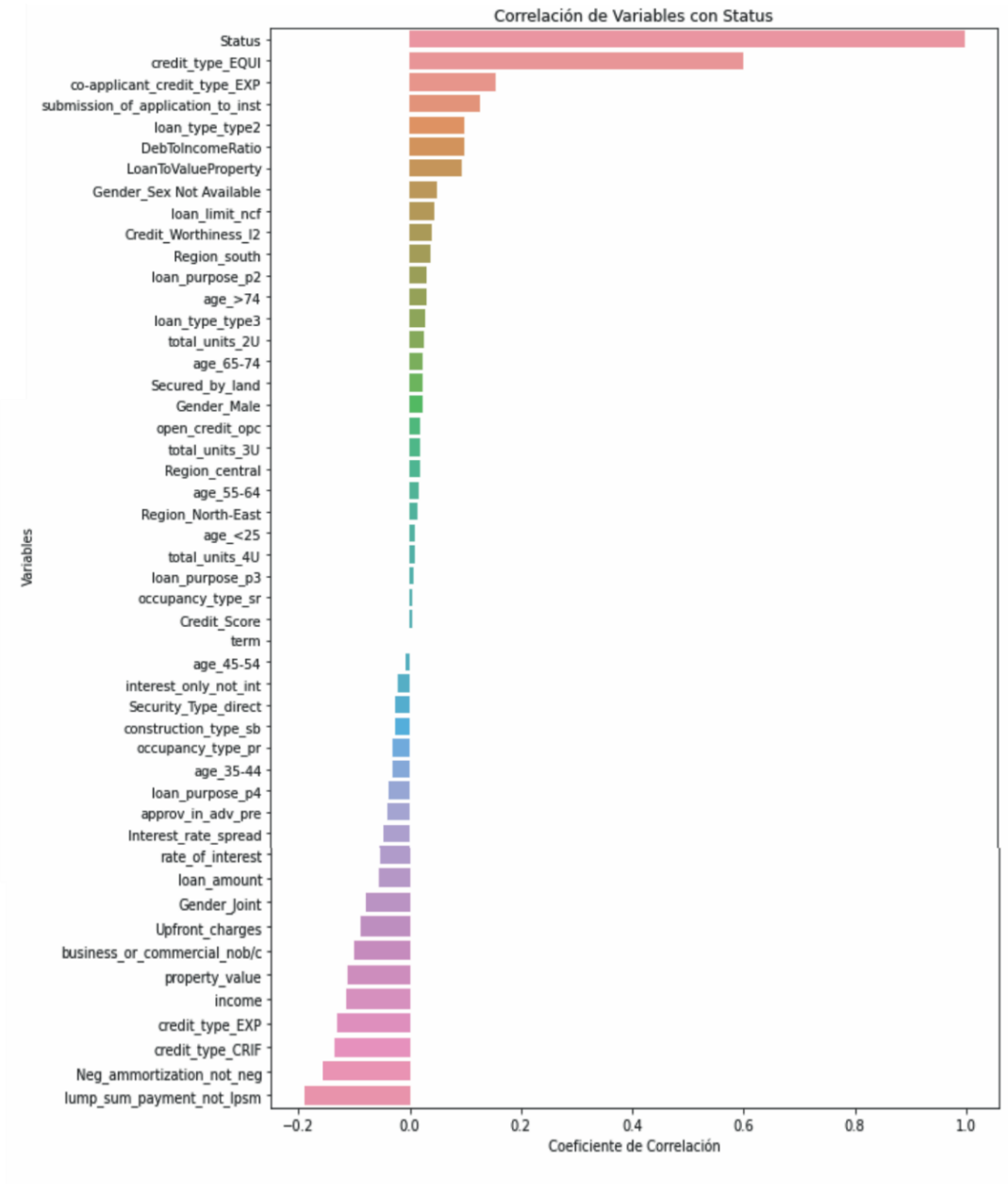


Fuente: Elaboración propia.

Existen más de 100.000 registros o transacciones donde se muestra que hay créditos que se han pagado correctamente, mientras que existe 40.000 transacciones con las características de préstamos en mora.

g. Análisis de correlación

Gráfico 3. Correlación de Variables con Status



Fuente: Elaboración propia.

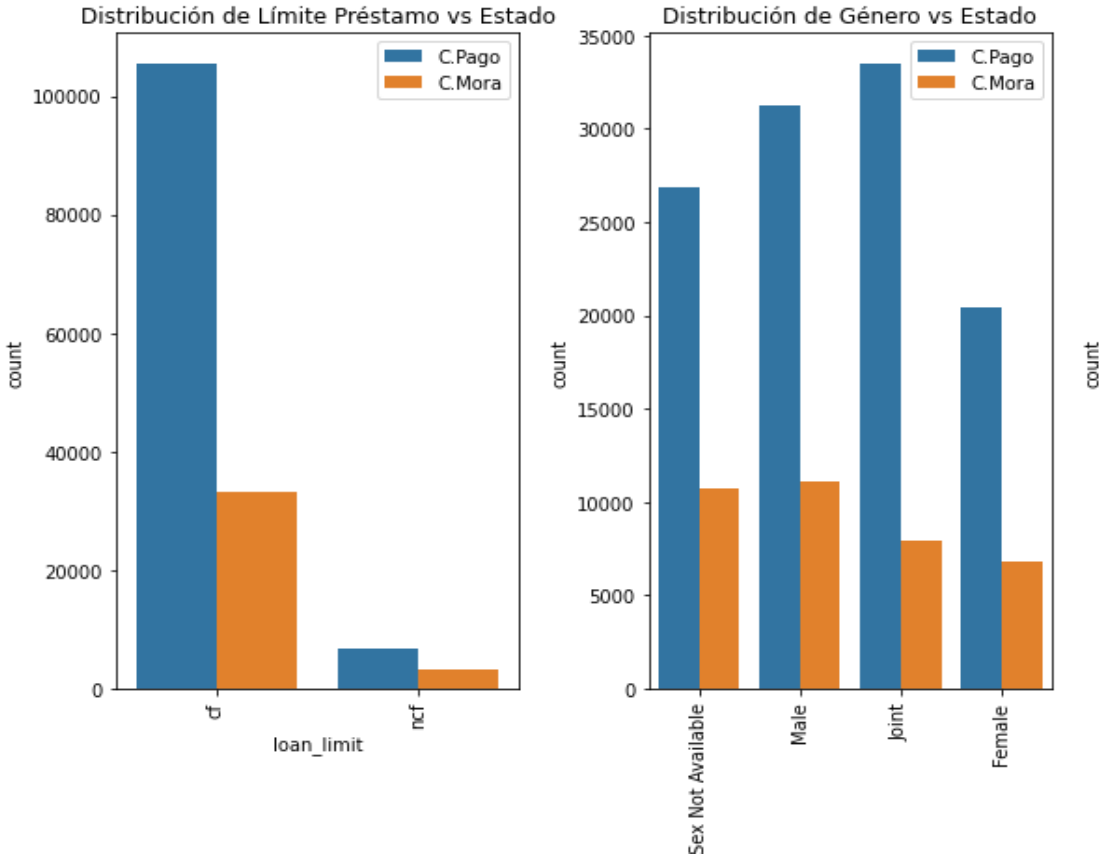
De acuerdo con la matriz de correlación, las variables con mayor relación tienen es el "property_value" con el "loan_amount", es decir la propiedad del valor con el monto del préstamo. La siguiente relación es entre el "interes_rate_spread"

con "rate_of_interest" que es la diferencia de interés con respecto al punto de referencia versus la tasa de interés.

h. Análisis de gráficos basados en el EDA

Los gráficos que se generan como parte del Análisis exploratorio de datos, con la visualización, se puede lograr identificar importantes hallazgos sin necesidad de recurrir hacia algoritmos o códigos para interpretar los resultados.

Gráfico 4. Distribución de Limites de Préstamo y Género



Fuente: Elaboración propia.

Estado del Préstamo por Limite: La mayoría de los préstamos están clasificados como Facilidades de crédito (CF), mientras que una minoría de pagos son los sin facilidad de crédito (NCF) ambos tipos de préstamos tienen los presamos pagados, la proporción de préstamos en estado de Mora es menor que ambas

categorías, pero más notablemente menor en la categoría de Facilidades de pago (CF).

La mayoría de los préstamos se otorgan bajo condiciones de CF, lo que podría implicar mejores condiciones de crédito. Los préstamos en NCF son menos comunes, lo que sugiere condiciones de crédito más restrictivas.

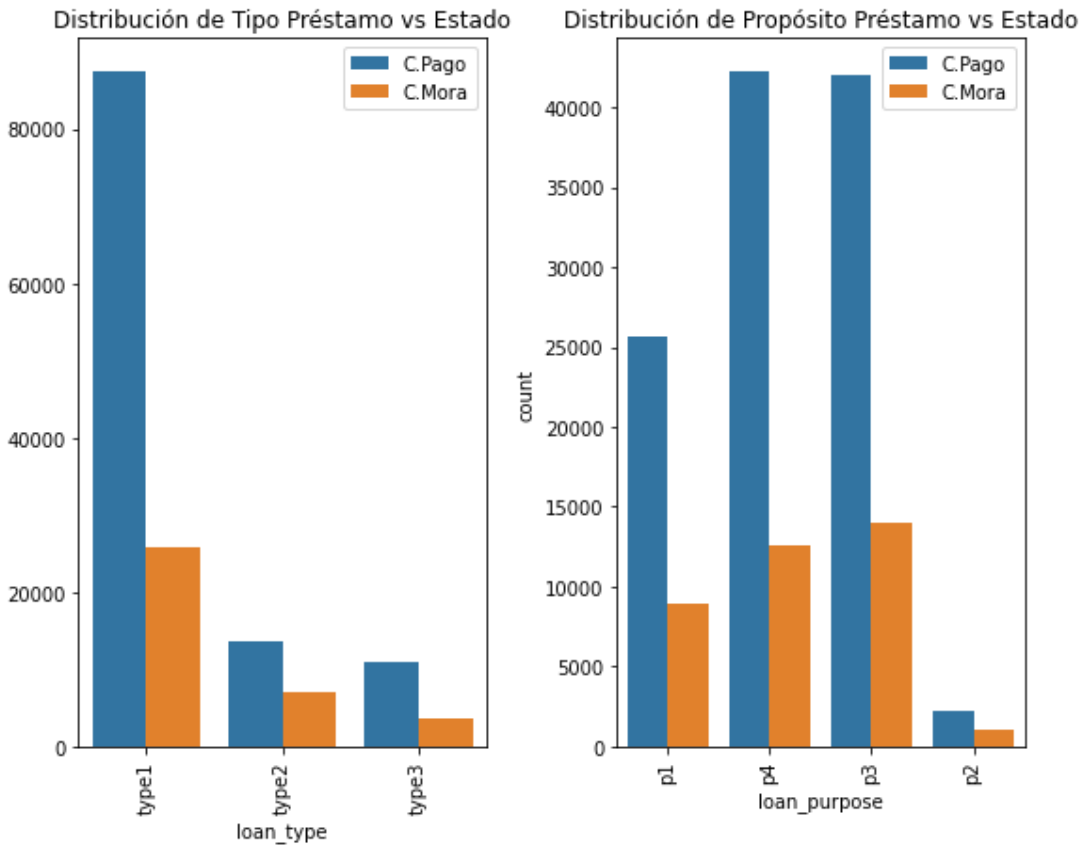
En ambas categorías (CF y NCF), los prestatarios tienden a cumplir con sus pagos, dado que la mayoría de los préstamos están en estado de crédito pagado.

La proporción de préstamos en mora es menor, pero más significativa en NCF, lo que podría implicar un mayor riesgo asociado con esta categoría.

Se debe continuar monitoreando los préstamos en NCF para identificar patrones de riesgo y desarrollar estrategias de mitigación. Evaluar las condiciones de CF para asegurarse de que están siendo efectivas en promover el cumplimiento de pagos sin incurrir en demasiados riesgos.

Estado del Préstamo por Género: Las mujeres y las cuentas conjuntas tienden a tener un mejor comportamiento de pago, con una mayor proporción de préstamos en estado de pago. Los hombres y la categoría de género no disponible muestran un mayor riesgo de préstamos en mora, lo que podría ser un área de enfoque para las estrategias de mitigación de riesgo.

Gráfico 5. Distribución de Tipo y Propósito de Préstamo



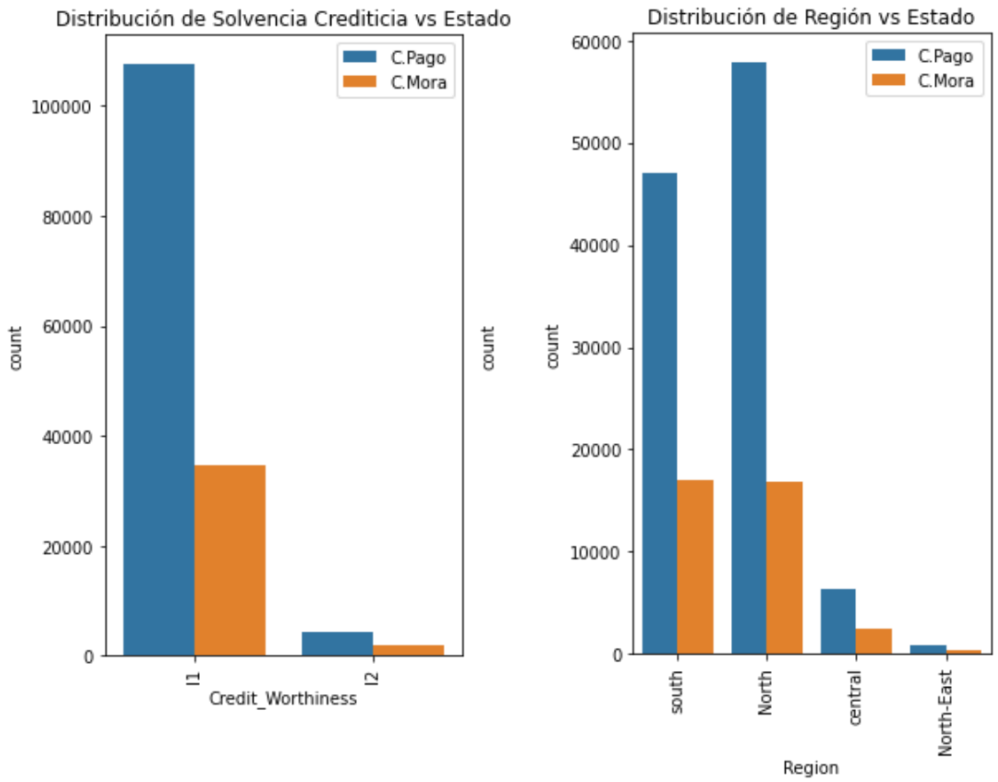
Fuente: Elaboración propia.

Estado del Préstamo por Tipo de Préstamo: Los préstamos de type1 son los más numerosos y tienen una proporción considerable de préstamos en mora, lo que sugiere que, aunque sean más accesibles o populares, también llevan un riesgo considerable. Los type2 y type3 tienen menos préstamos en general, pero muestran una proporción significativa de préstamos en mora, lo que puede indicar que estos tipos de préstamos tienen condiciones más restrictivas o son solicitados por prestatarios con perfiles de riesgo más altos lo que hace que sea una relación de 2 a 1 entre los préstamos pagados y los que están en mora.

Estado del Préstamo por Propósito del Préstamo: Los préstamos con propósito p3 y p4 son los más numerosos y muestran una alta proporción de préstamos en estado de pagado, lo que sugiere que estos propósitos están asociados con prestatarios que tienden a cumplir con sus obligaciones.

El propósito p2, aunque tiene una menor cantidad de préstamos, muestra una mayor proporción de préstamos en mora en relación con su cantidad total, lo que puede indicar un mayor riesgo asociado a este propósito de préstamo.

Gráfico 6. Distribución de Solvencia Crediticia y Región.



Fuente: Elaboración propia.

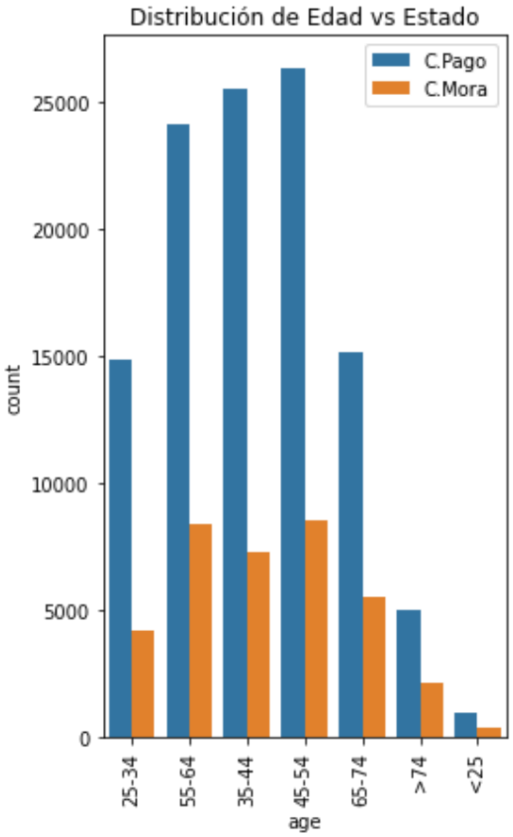
Estado del Préstamo por Solvencia Crediticia: Los prestatarios con solvencia crediticia l1 tienden a tener una mayor cantidad de préstamos, tanto en estado de pago como en mora, lo que puede indicar que esta categoría abarca una amplia gama de perfiles de riesgo.

Los prestatarios con solvencia crediticia l2, aunque menos numerosos, muestran una proporción de mora significativa, lo que puede implicar un mayor riesgo asociado con esta categoría. Las categorías de solvencia crediticia muestran diferentes proporciones de estados del préstamo.

Estado del Préstamo por Región: En todas las regiones, la mayoría de los préstamos están en estado de pago (C. Pago). Las regiones North y South muestran una alta actividad crediticia con una mayor cantidad de préstamos tanto en estado de pago como en mora, lo que sugiere que estas regiones son económicamente más activas.

Las regiones Central y North-East, aunque menos activas en términos de volumen de préstamos, muestran proporciones significativas de mora, lo que podría indicar áreas de mayor riesgo relativo.

Gráfico 7. Distribución de Edad



Fuente: Elaboración propia.

Estado del Préstamo por segmento de edad: Los prestatarios en los grupos de edad 35-44, 45-54 y 55-64 tienden a tener una mayor cantidad de préstamos, tanto en estado de pago como en mora, lo que puede indicar que estos grupos

son económicamente más activos y tienen una mayor necesidad de financiamiento.

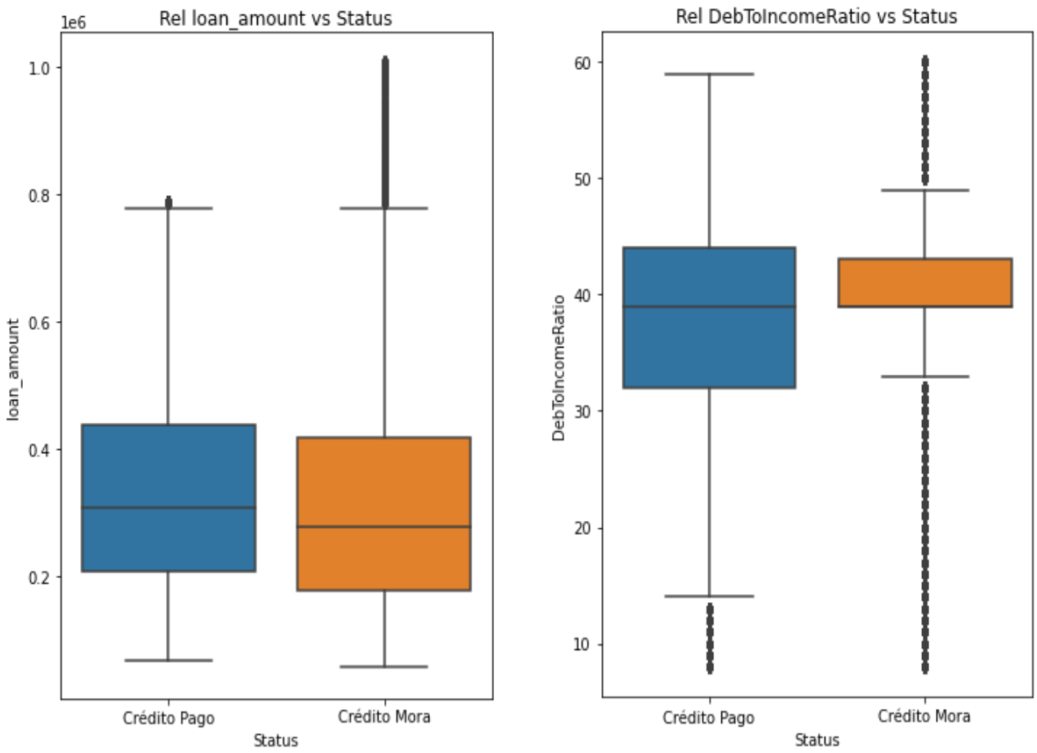
Los prestatarios en el grupo de edad 25-34 también muestran una actividad significativa, pero con una menor cantidad de préstamos en mora en comparación con los grupos mayores.

Los grupos de edad <25 y >74 tienen menos préstamos en general, pero siguen mostrando una mayor cantidad de préstamos en estado de pago.

i. Tratamiento de valores atípicos o Outliers

Eliminar valores atípicos es una parte importante del preprocesamiento de datos, especialmente en aplicaciones como el credit scoring, donde los valores atípicos pueden distorsionar los resultados del modelo.

Gráfico 8. Ejemplo de Valores Atípicos



Fuente: Elaboración Propia.

Justificación para Eliminar Valores Atípicos

En la elaboración propia de todos los gráficos de caja, se puede observar que existen varios valores atípicos por cada variable (Cedeño&Pacheco, 2024), por lo que necesario aplicar un modelo y justificar las razones de eliminación de valores atípicos que cumplen diferentes objetivos como:

Mejorar del Rendimiento del Modelo: Los valores atípicos pueden sesgar las estimaciones del modelo y reducir su precisión. Al eliminar valores atípicos, el modelo puede generalizar mejor los datos y mejorar su rendimiento en datos no vistos.

Reducción de la Varianza: Los valores atípicos pueden aumentar la varianza del modelo, haciéndolo menos robusto. Eliminar estos valores ayuda a estabilizar las estimaciones y reducir la varianza.

Simplificación del Modelo: Los valores atípicos pueden complicar el modelo, haciendo que sea más difícil de interpretar y explicar. Un conjunto de datos más limpio y sin valores atípicos permite construir modelos más simples y comprensibles.

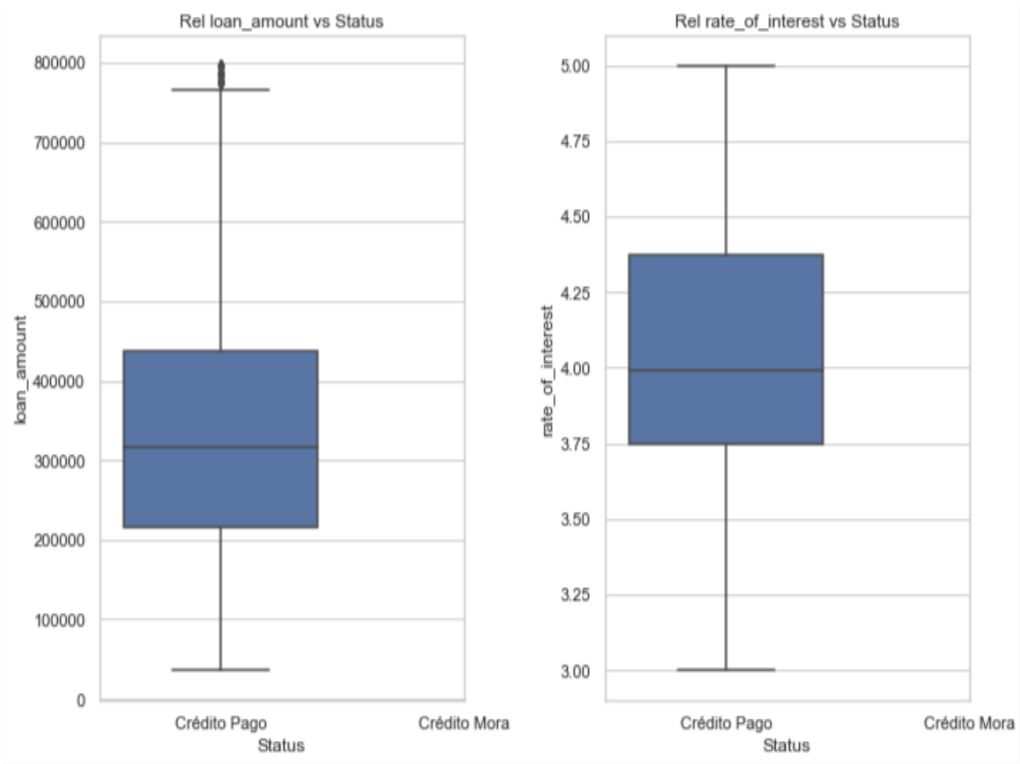
Rango Intercuartílico

Rango Intercuartílico (IQR): Los datos se consideran valores atípicos si están por debajo del primer cuartil (Q1) menos 1.5 veces el IQR o por encima del tercer cuartil (Q3) más 1.5 veces el IQR.

Fórmula 1. Rango Intercuartílico

$$Q1 = datos.quantile\ (0.25)$$
$$Q3 = datos.quantile\ (0.75)$$
$$IQR = Q3 - Q1$$

Gráfico 9. Representación de eliminación de variables atípicos



Fuente: Elaboración Propia.

Eliminar valores atípicos de manera demasiado agresiva puede llevar a la eliminación de datos críticos para el análisis, como todos los registros con un "status" específico. Esto puede ocurrir si los valores de esas filas se consideran atípicos en varias de las columnas numéricas.

Para abordar este problema, es importante:

- Verificar la distribución de la variable objetivo después de la eliminación de outliers para asegurarse de que no se ha distorsionado significativamente.
- Aplicar técnicas más robustas para la detección de outliers, que no eliminen todos los registros de una categoría crítica.
- Realizar la detección de outliers de manera condicional, es decir, considerar la variable objetivo al decidir qué valores son outliers.

- Al revisar los boxplot se identifican que todos los outliers se eliminan de aquellos registros que están por esta razón es necesario mejorar el tratamiento de los outliers.
- Aplicar una técnica más robusta para la detección de outliers (por ejemplo, z-score condicional a la variable objetivo).

Z-score condicional a la variable objetivo

Podemos usar el z-score condicional para detectar outliers, asegurándonos de no eliminar todos los registros de una categoría crítica.
A la vez que cuando es muy estricta la eliminación es necesario el ajuste del Umbral del Z-Score y Aplicación de la Eliminación de Outliers.

Al ajustar el umbral del z-score a 3.5 para hacer la eliminación de outliers menos estricta y verificar las estadísticas descriptivas antes y después de la eliminación de outliers se concluye que aun así se eliminan la mayoría de los datos.

Al parecer que hay un problema subyacente en los datos o en la forma en que se está calculando el z-score. Podemos probar un enfoque diferente para eliminar outliers que podría ser más robusto.

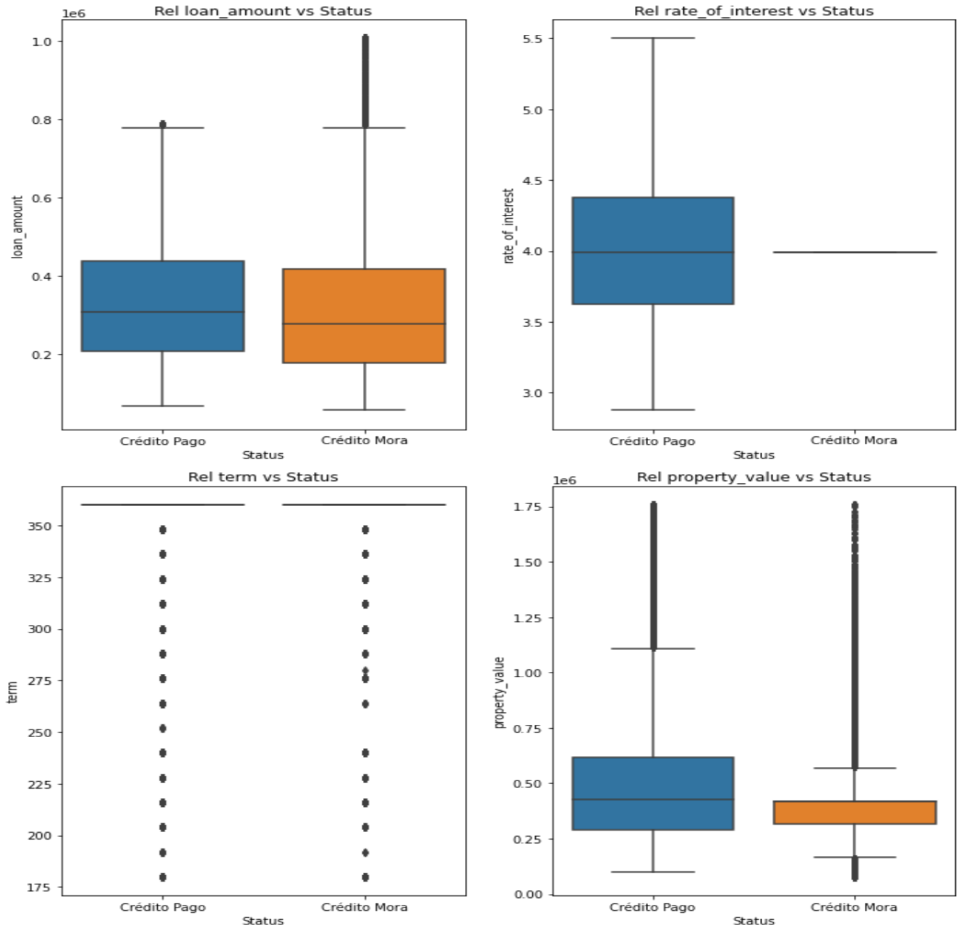
Usar percentiles para identificar Outliers

En lugar de usar z-scores, podemos usar percentiles para identificar y eliminar outliers. Esto es menos sensible a la distribución de los datos y puede ser más efectivo en conjuntos de datos con distribuciones no normales.

En lugar de usar z-scores, podemos usar percentiles para identificar y eliminar outliers. Esto es menos sensible a la distribución de los datos y puede ser más efectivo en conjuntos de datos con distribuciones no normales.

Es necesario utilizar la técnica de percentiles, para que el modelo sea menos sensible.

Gráfico 10. Eliminación de Outliers con Percentiles



Fuente: Elaboración Propia.

3. Modelado de los Datos

a. Definición de Regresión logística, formas y usos:

La regresión logística es una elección sólida para el modelado de datos de Credit Scoring debido a su capacidad para manejar problemas de clasificación binaria, su interpretabilidad, eficiencia, y capacidad para proporcionar probabilidades de incumplimiento, lo que ayuda a las instituciones financieras a tomar decisiones de crédito más informadas y justas.

La regresión logística es una técnica ampliamente utilizada para el modelado de datos de Credit Scoring debido a varias razones que la hacen adecuada y eficaz en este contexto presentamos algunas de las principales razones:

1. Naturaleza Binaria del Problema

El Credit Scoring generalmente se trata de un problema de clasificación binaria, donde el objetivo es predecir si un solicitante de crédito incumplirá o no (default/no default). La regresión logística es una herramienta ideal para problemas de clasificación binaria, ya que estima la probabilidad de que un evento ocurra.

2. Interpretabilidad

La regresión logística proporciona coeficientes que pueden interpretarse directamente, lo que es crucial en el contexto de Credit Scoring, donde los bancos y las instituciones financieras necesitan entender claramente cómo cada variable afecta la probabilidad de incumplimiento. Esto permite una mayor transparencia y explicabilidad en las decisiones de crédito.

3. Escalabilidad

La regresión logística es computacionalmente eficiente y puede manejar grandes conjuntos de datos, lo cual es común en aplicaciones de Credit Scoring. Su implementación y ajuste son relativamente rápidos en comparación con otros modelos más complejos.

4. Manejo de Multicolinealidad

Aunque la multicolinealidad puede ser un problema, la regresión logística tiene técnicas como la regularización (Ridge o Lasso) que pueden mitigar sus efectos, mejorando la estabilidad y la precisión del modelo.

5. Estabilidad y Robustez

La regresión logística es menos propensa a sobre ajustar los datos en comparación con modelos más complejos como las redes neuronales, lo que la hace más robusta y confiable para la toma de decisiones en el ámbito financiero.

6. Facilidad de Implementación

Las herramientas y bibliotecas para la regresión logística están bien desarrolladas y ampliamente disponibles en lenguajes de programación como Python y R. Esto facilita su implementación y despliegue en sistemas de producción.

7. Evaluación de Probabilidades

La regresión logística no solo clasifica las observaciones, sino que también proporciona probabilidades de incumplimiento, lo cual es valioso para evaluar el riesgo y tomar decisiones más informadas. Estas probabilidades pueden usarse para segmentar a los solicitantes en diferentes niveles de riesgo.

8. Flexibilidad con Variables

La regresión logística puede manejar tanto variables continuas como categóricas, permitiendo una mayor flexibilidad en la inclusión de diferentes tipos de datos relevantes para el Credit Scoring

b. Aplicación del modelo de regresión logística.

Para proceder con la regresión logística, realizaremos los siguientes pasos:

- Codificar las variables categóricas: Convertir las variables categóricas en variables numéricas utilizando la codificación one-hot, esto es crucial para que los modelos de aprendizaje automático puedan utilizar esta información. Cada categoría única de una variable categórica se convierte en una columna separada con valores binarios (0 o 1).

- Balance de Clases: Al entrenar un modelo de regresión logística, es importante considerar el balance de clases en la variable objetivo. Si una clase está

subrepresentada (como en este caso, los préstamos en mora), puede ser necesario utilizar técnicas de sobre muestreo, submuestreo o ajuste de los pesos de las clases para mejorar el rendimiento del modelo.

- División de Datos: La división en conjuntos de entrenamiento y prueba permite evaluar el rendimiento del modelo en datos no vistos durante el entrenamiento, proporcionando una medida más realista de su desempeño.

- Evaluar el modelo: Analizar el rendimiento del modelo utilizando el conjunto de prueba y métricas de evaluación como la matriz de confusión, precisión, recall (Cedeño&Pacheco, 2024).

La división de los datos se muestra en la siguiente manera:
Forma de X_train y X_test:
X_train tiene 91,457 registros y 48 columnas.
X_test tiene 39,197 registros y 48 columnas.
Forma de y_train: (91457,)
Forma de y_test: (39197,)

Hemos codificado las variables categóricas y dividido los datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento tiene 91,457 registros y 48 columnas, mientras que el conjunto de prueba tiene 39,197 registros y las mismas 48 columnas.

Cada columna en X_train y X_test representa una variable predictora. Las 48 columnas incluyen tanto las variables numéricas originales como las variables categóricas codificadas.
Forma de y_train y y_test:
y_train y y_test contienen la variable objetivo para los conjuntos de entrenamiento y prueba, respectivamente.
y_train tiene 91,457 valores y y_test tiene 39,197 valores, cada uno correspondiente a un registro en X_train y X_test.

V. RESULTADOS

1. Implementación del modelo de regresión logística, modelo base

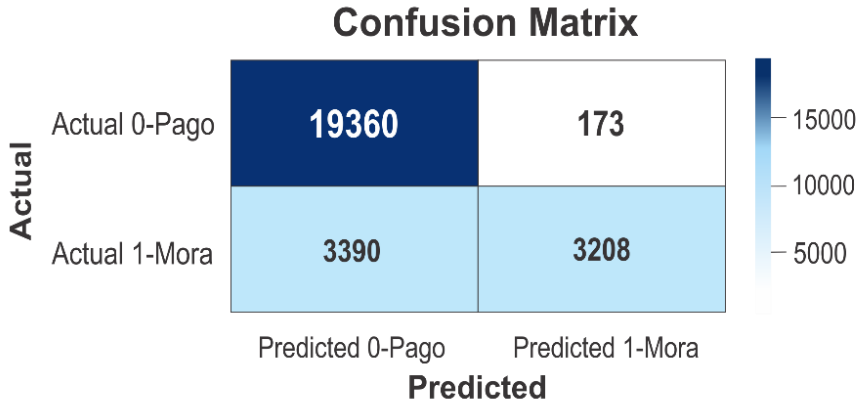
El resultado a continuación es la ejecución del código Python para la generación del reporte de regresión logística y se incluye la matriz de confusión.

Tabla 7. Regresión Logística resultado Modelo Base

	precision	recall	f1-score	support
0	0.85	0.99	0.92	19533
1	0.85	0.49	0.64	6598
accuracy			0.86	26131
macri avg	0.90	0.74	0.78	26131
weighted avg	0.88	0.86	0.86	26131

Fuente: Elaboración propia.

Gráfico 11. Matriz Regresión logística resultado Modelo Base



Fuente: Elaboración Propia.

El modelo tiene un buen rendimiento en la predicción de la clase mayoritaria (0) con un F1-Score de 0.92.

El rendimiento para la clase minoritaria (1) es más bajo, con un F1-Score de 0.64, lo que indica que el modelo tiene dificultades para predecir correctamente las instancias de esta clase.

La exactitud global del modelo es alta (0.86), pero dado el desbalance en las clases, es importante considerar otras métricas como el recall y el F1-Score para evaluar el rendimiento del modelo en la clase minoritaria.

Así, el recall que muestra la sensibilidad, mide la proporción de verdaderos positivos que son correctamente identificados por el modelo, es decir, de todas las instancias que realmente pertenecen a la clase positiva.

Fórmula 2.- Medición de Recall

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Recall para la clase 1 (minoritaria): 0.49
Esto significa que el modelo solo está identificando correctamente el 49% de las instancias de la clase minoritaria. El 51% restante son falsos negativos (instancias que el modelo predijo incorrectamente como clase 0).

Así, el F1-Score es la media armónica entre la precisión y el recall, proporcionando una única métrica que balancea ambos aspectos. Es especialmente útil cuando se necesita un balance entre la precisión (qué tan correcto es el modelo cuando predice la clase positiva) y el recall (qué tan bien el modelo captura la clase positiva).

Fórmula 3. Representación del F1-Score

$$F1\ Score = 2 * (\frac{Precision * Recall}{Precision\ + Recall})$$

F1-score para la clase 1 (minoritaria): 0.64

El F1-score más bajo para la clase minoritaria refleja que, aunque el modelo puede tener una precisión razonable (95%), no está capturando suficientes instancias de la clase positiva (recall bajo), lo que reduce el F1-score.

Para mejorar el rendimiento en la clase minoritaria se puede:

- Aplicar técnicas de balance de clases como sobre muestreo de la clase minoritaria o submuestreo de la clase mayoritaria.
- Ajustar los umbrales de decisión del modelo.
- Utilizar técnicas de modelado más avanzadas como Random Forest y Gradient Boosting (J. Monroy-de-Jesús, A. Guadalupe-Ramírez, J.C. Ambriz-Polo, E. López-González, 2018).

2. Modelo Regresión implementado SMOTE

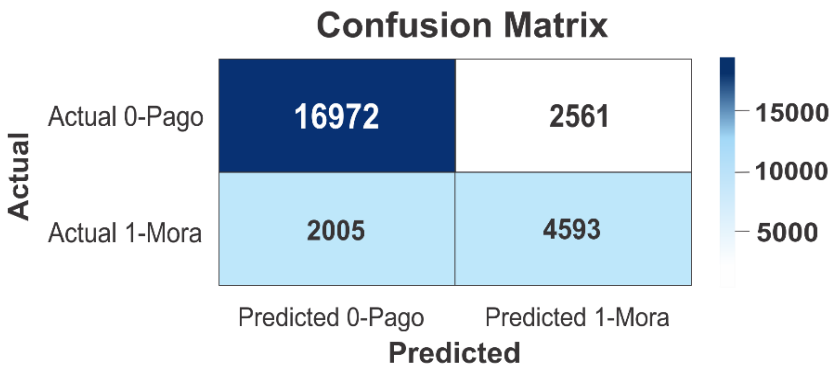
Al aplicar las técnicas de balance de clases una de ella es SMOTE, la cual genera instancias sintéticas para la clase minoritaria y posteriormente entrenar el modelo nuevamente con la regresión logística.

Tabla 8. Regresión logística – SMOTE

	precision	recall	f1-score	support
0	0.89	0.87	0.88	19533
1	0.64	0.70	0.67	6598
accuracy			0.83	26131
macri avg	0.77	0.78	0.77	26131
weighted avg	0.83	0.83	0.83	26131

Fuente: Elaboración Propia.

Gráfico 12. Matriz Regresión logística – SMOTE



Fuente: Elaboración Propia.

El recall de la clase minoritaria (1) ha aumentado significativamente de 0.49 a 0.70, lo que indica que el modelo ahora está capturando más instancias positivas.
La precisión de la clase minoritaria ha disminuido de 0.95 a 0.64, lo que refleja un incremento en los falsos positivos.

El F1-score de la clase minoritaria ha mejorado de 0.64 a 0.67, lo que muestra un mejor balance entre precisión y recall.

La exactitud global del modelo es ligeramente menor (0.83) comparada con el modelo no balanceado (0.86), lo que es un resultado esperado debido al enfoque en mejorar la detección de la clase minoritaria.

El balanceo de clases con SMOTE ha mejorado significativamente la capacidad del modelo para detectar instancias de la clase minoritaria, a expensas de una ligera disminución en la precisión y exactitud global. Este es un compromiso común en problemas de desbalance de clases, donde mejorar el recall de la clase minoritaria puede ser más importante que mantener una precisión global alta.

3. Evaluar rendimiento de modelo con Random Forest.

Para garantizar la robustez del modelo, utilizaremos la validación cruzada. La validación cruzada es una técnica que divide el conjunto de datos en varios subconjuntos (folds) y entrena el modelo en varios ciclos, asegurando que cada subconjunto se utilice tanto para el entrenamiento como para la validación.
Vamos a utilizar la validación cruzada con 5 subconjuntos para evaluar el rendimiento del modelo balanceado con SMOTE. Esto nos permitirá obtener una estimación más confiable del rendimiento del modelo en datos no vistos.

Tabla 9. Resultado de Random Forest y validación cruzada

F1 - Score promedio :	0.637920852179077
Desviación Standard del F1 - Score:	0.0038805076938629862

Fuente: Elaboración Propia.

El F1-score promedio de 0.63 indica que, en general, el modelo tiene un rendimiento moderado en términos de balance entre precisión y recall. El F1-score se calcula como la media armónica de la precisión y el recall, por lo que este valor sugiere:

Moderada Precisión: El modelo tiene una tasa razonable de predicciones correctas para la clase positiva.

Moderado Recall: El modelo también es razonablemente capaz de identificar las instancias positivas.

Un F1-score de 0.63 puede considerarse adecuado en contextos donde se espera un rendimiento balanceado entre detectar verdaderos positivos y minimizar falsos positivos y negativos. Sin embargo, hay margen para mejorar.

Mientras que, la desviación Estándar del F1-Score de 0.003 es extremadamente baja, lo que indica que el rendimiento del modelo es muy consistente entre las

diferentes particiones del conjunto de datos en la validación cruzada. El modelo proporciona resultados similares en cada partición, por ende, el modelo es robusto y no está sobre ajustado.

Se puede tener confianza en que el rendimiento promedio del modelo (F1-score de 0.63) es representativo y no fluctúa significativamente dependiendo de cómo se divida el conjunto de datos.

4. Optimización de modelo Random Forest.

Para optimizar el modelo de Random Forest, será necesario ajustar algunos hiperparámetros clave:

- n_estimators: Número de árboles en el bosque.
- max_depth: Profundidad máxima de cada árbol.
- min_samples_split: Es el número mínimo de muestras requeridas para dividir un nodo.
- min_samples_leaf: Número mínimo de muestras que debe tener un nodo hoja.
- max_features: Es el número de características a considerar al buscar la mejor división.

Exploración de algunas combinaciones de hiperparámetros importantes donde se evalúa el rendimiento.

1. Número de Árboles (n_estimators): 100, 200, 300
2. Profundidad Máxima (max_depth): 10, 20, 30
3. Muestras Mínimas para Dividir (min_samples_split): 2, 5, 10
4. Muestras Mínimas en Hojas (min_samples_leaf): 1, 2, 4
5. Características Máximas (max_features): 'sqrt', 'log2'

El mejor modelo que se ajuste después de varias simulaciones se genera con los siguientes datos.

- n_estimators: 200
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2
- max_features: 'sqrt'
- **F1-Score**: Approximately 0.65
- **Standard Deviation**: Approximately 0.002

Mejora el rendimiento con el F1-score alrededor de 0.65, una mejora sobre el inicial de 0.63, esto muestra que los hiperparámetros seleccionados solo los mejores para el problema del conjunto de datos, la baja desviación estándar de 0.002 indica un rendimiento consistente a lo largo de diferentes subconjuntos asegurando la robustez del modelo.

VI. DISCUSION DE LOS RESULTADOS Y PROPUESTA DE VALOR

En este análisis, hemos desarrollado y evaluado varios modelos para predecir el estado de los préstamos (Status) en un conjunto de datos desbalanceado. Los pasos principales involucraron:

Resultados

Regresión Logística del modelo base:

En este se observaron problemas con la detección de la clase minoritaria de las transacciones que caen en default o mora en pago (Status = 1), reflejados en un F1-score bajo para esta clase.

- F1-Score Promedio: Bajo
- Recall: Bajo para la clase minoritaria
- Precisión: Alta para la clase mayoritaria

Regresión Logística con SMOTE:

Se utilizó SMOTE para balancear las clases del conjunto de datos, Se reentrenó el modelo de regresión logística con los datos balanceados.

- F1-Score Promedio: Mejorado en comparación con el modelo inicial.
- Recall: Aumentado significativamente para la clase minoritaria.
- Precisión: Se redujo ligeramente debido al aumento de falsos positivos, pero se logró un mejor equilibrio.

Modelo Random Forest con Validación Cruzada y SMOTE:

Se entrenó un modelo de Random Forest utilizando los datos balanceados con SMOTE.

- F1-Score Promedio: 0.63
- Desviación Estándar del F1-Score: 0.003, indicando alta consistencia.
- Precisión y Recall: Balanceados con mejoras significativas en comparación con la regresión logística.

Modelo Random Forest Optimizado:

Se optimizaron los hiperparámetros del modelo Random Forest con los hiperparámetros ideales.

- F1-Score Promedio: 0.65
- Desviación Estándar del F1-Score: 0.002, indicando una consistencia aún mayor.
- La Matriz de Confusión finalmente generada mostró un equilibrio adecuado entre las predicciones correctas de ambas clases, con una mejora notable en la detección de la clase minoritaria.

A continuación, el cuadro comparativo de los modelos ejecutados como variable de análisis el Accuracy de cada uno de ellos.

Tabla 10.- Comparación de Modelos

Macro Average/ Modelo	**Precisión**	**Recall**	**F1-score**
Regresión Logística base	0.90	.74	0.78
Regresión Logística SMOTE	0.77	0.78	0.77
Random Forest	1	1	1
Random Forest Optimizado	1	1	1

Fuente: Elaboración Propia

Tabla 11.- Comparación del Accuracy de Modelos

	Regresión Logística Base	Regresión logística SMOTE
Accuracy	0.86	0.83

Fuente: Elaboración Propia

Propuesta de Valor

La implementación de SMOTE y el uso de modelos avanzados como Random Forest, junto con la optimización de hiperparámetros, proporcionan una mejora clara en la detección de la clase minoritaria en problemas de clasificación desbalanceada.

1. Mejora en la Detección de Préstamos Riesgosos:

La aplicación de SMOTE y el uso de Random Forest optimizado aumentan significativamente la capacidad de detectar préstamos que están en riesgo de incumplimiento, reduciendo la tasa de falsos negativos.
Esto permite a las instituciones financieras identificar y gestionar mejor los riesgos asociados con los préstamos.

2. Consistencia y Robustez del Modelo:

La baja desviación estándar del F1-score en la validación cruzada sugiere que el modelo es confiable y produce resultados consistentes.
Esto es crucial para la toma de decisiones basada en datos, asegurando que el modelo se comporte de manera predecible en diferentes subconjuntos de datos.

3. Base para Mejoras Futuras:

La metodología aplicada proporciona una base sólida para futuras optimizaciones, como el ajuste de hiperparámetros o la exploración de otros algoritmos de ensamble.
Las técnicas avanzadas como Gradient Boosting o XGBoost pueden ser consideradas para mejorar aún más el rendimiento.

4. Beneficios Operacionales:

La capacidad de identificar y mitigar riesgos de manera más efectiva puede resultar en una reducción significativa de las pérdidas por incumplimiento de préstamos.

Además, permite la asignación más eficiente de recursos para la gestión de cartera de préstamos y la toma de decisiones estratégicas.

VII. CONCLUSIONES Y RECOMENDACIONES.

Conclusiones

Según fuentes secundarias, el default crediticio puede indicar problemas económicos profundos, como la incapacidad para pagar deudas y la falta de liquidez de una institución financiera, hay diferentes tipos de default crediticio, como el de pago, el de incumplimiento y el de crédito. Cada tipo puede tener consecuencias diferentes para las partes involucradas. Afectando de manera directa a la estructura de capital de las instituciones financieras e impidiendo el financiamiento a nuevos clientes.

Gracias al análisis exploratorio de datos mediante la regresión logística logramos sacar el máximo provecho de la variable dependiente status, ya que identificamos las variables de mayor afectación que nos dieron a conocer de manera más precisa la predicción del Default aumentando la certidumbre en un 25% versus los sistemas actuales lo que representa algunos miles de dólares en beneficio a la liquidez de las instituciones financieras y que dependiendo de su tamaño se puede traducir en millones.

La integración de técnicas de balanceo de clases como SMOTE y el uso de algoritmos avanzados como Random Forest, junto con la optimización de hiperparámetros, han demostrado ser efectivos en mejorar el rendimiento de los modelos de clasificación en conjuntos de datos desbalanceados. Estos enfoques no solo mejoran la capacidad de detección de la clase minoritaria, sino que también aseguran la consistencia y robustez del modelo, proporcionando un valor significativo para la gestión de riesgos en instituciones financieras. Continuar con la optimización y exploración de técnicas avanzadas puede llevar a mejoras adicionales y proporcionar un marco robusto para la toma de decisiones informadas.

La implementación de un sistema de puntuación más detallado y preciso, que considere variables categóricas y numéricas, y otorgue puntuaciones mayores a

las variables positivas y clientes con capacidades de pago seguras, puede mejorar significativamente la evaluación de la solvencia de los clientes. El uso de métodos avanzados como bosques aleatorios puede aumentar la precisión en la predicción de resultados y la toma de decisiones. Esto permite a las instituciones financieras:

Incremento de la capacidad de crédito: La automatización puede permitir a los prestamistas evaluar y aprobar créditos de manera más rápida y eficiente, lo que puede aumentar la capacidad de crédito.

Recomendaciones

Es fundamental que las instituciones financieras implementen medidas proactivas para identificar y mitigar el riesgo de default crediticio, como:

- Monitorear de cerca la solvencia y liquidez de los deudores
- Diversificar la cartera de créditos para minimizar el riesgo
- Establecer políticas de crédito estrictas y transparentes
- Realizar análisis de crédito exhaustivos antes de otorgar préstamos
- Mantener reservas y provisiones adecuadas para cubrir posibles pérdidas
- Desarrollar planes de contingencia para abordar situaciones de default crediticio.

Usar datos limpios y estandarizados es la clave de la calidad de los datos que se presenten para la evaluación en los modelos es importante para tener la certeza de los algoritmos funcionen correctamente, así se puede usar todo el conjunto de datos, si necesidad de completar con la media, la moda o eliminar filas por falta de datos para que no impacte a los modelos aplicados.

Utilizar técnicas de balanceo de clases como SMOTE o incluso usar otras variantes más avanzadas para asegurar que el modelo mantenga su capacidad de detección de clases minoritarias en futuros datos.

Se recomienda usar un modelo Random Forest optimizado en el entorno de detección de riesgo de incumplimiento para préstamo, es fundamental monitorear continuamente el rendimiento del modelo y ajustar los hiperparámetros según sea necesario para mantener su eficacia.

Otorgar nuevos tipos de valoraciones de puntuación en aspectos relevantes que pesan en la solvencia de los clientes, haciendo un análisis más preciso en las variables categóricas vs las variables numéricas. otorgando puntuaciones mayores a las variables positivas o clientes con capacidades de pagos más seguras cuyos criterios están fuera del nivel de endeudamiento por el que se decide la capitalización. Usando métodos más precisos como bosques aleatorios que determinan con mayor precisión los resultados para la toma de decisiones.

Implementar sistemas de aprendizaje de máquina en las instituciones financieras sobre todo a las pertenecientes al sector público, ya que la decisión debe tener análisis técnicos y no políticos, lo que ayudaría a tener ingentes ahorros de capital para mejora de los servicios. Así optimizar el tiempo de respuesta y ayudar a tener una mejor atención a los clientes de manera presencial o remota.

Bibliografía

Bambino Contreras, C., & Morales Oñate, V. (2023). Exposición al Default: Estimación para un Portafolio de Tarjeta. *Revista Politécnica*, 71.

Bayas Sánchez , D. L. (2020). *Factores que influyen en el endeudamiento con tarjeta de crédito en los tarjeta avientes de clase socioeconómica media en Guayaquil.* Guayaquil : Universidad Politecnica Salesiana .

Caraguay Muñoz, K. P., Pesantez León, S. A., & Elizalde Orellana, M. V. (2020). *Impacto de las tarjetas de crédito en el crecimiento económico ecuatoriano, período 2010-2020.* Machala: Universidad Técnica de .

LEAL FICA, A. L., ARANGUIZ CASANOVA, M. A., & GALLEGOS MARDONES, J. (2018). ANÁLISIS DE RIESGO CREDITICIO, PROPUESTA DEL MODELO CREDIT SCORING. *Revista Facultad de Ciencias Económicas: Investigación y Reflexión*, 3.

Salazar , J. C., Ramirez , J., Pinzón , L., & Rosemberg, C. (2019). *Estudio del modelo de Scoring de Ruta N.* CAF.

Venosi , C. L. (2021). *La nueva definición de default según la Autoridad Bancaria Europea: Impactos en los requerimientos de capital .* Buenos Aires : Universidad de San Andrés .

Acosta Mellado, E. I., Murillo Félix, C. A., & Almeida, L. d. (2021). ANÁLISIS DEL USO DE TARJETAS DE CRÉDITO DEL PERSONAL DEL H. AYUNTAMIENTO DE LA CIUDAD DE NAVOJOA, SONORA. *Revista inclusiones*, 10.

Bambino Contreras, C., & Morales Oñate, V. (2023). Exposición al Default: Estimación para un Portafolio de Tarjeta. *Revista Politécnica*, 71.

Bayas Sánchez , D. L. (2020). *Factores que influyen en el endeudamiento con tarjeta de crédito en los tarjeta avientes de clase socioeconómica media en Guayaquil.* Guayaquil : Universidad Politecnica Salesiana .

Caraguay Muñoz, K. P., Pesantez León, S. A., & Elizalde Orellana, M. V. (2020). *Impacto de las tarjetas de crédito en el crecimiento económico ecuatoriano, período 2010-2020.* Machala: Universidad Técnica de .

CedeñoPacheco. (07 de 2024). *LoanDefaultv1*. Obtenido de Github: https://github.com/sanfer1ec/ProyectoUdla/blob/main/LoanDefaultv1.ipynb

Cevallos Jiménez, A. B., & Ormaza Andrade, J. E. (2021). Estudio de factibilidad para la implementación de una tarjeta de crédito con depósito inicial en la ciudad de Cuenca. *Revista Cientifica Dominio de las ciencias*, 1092.

Elizalde Chapa, C. K. (2023). *Ventajas y Desventajas de Los Riesgos Financieros.* México .

H, M. Y. (2021). *Loan Default Dataset*. Obtenido de Kaggle: https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data

Lapo Maza, M. d., Tello Sánchez, M. G., & Mosquera Camacas, S. C. (2021). Rentabilidad, capital y riesgo crediticio en bancos ecuatorianos. *Scielo*, 54.

LEAL FICA, A. L., GALLEGOS MARDONES, J., & ARANGUIZ CASANOVA, M. A. (2018). *ANÁLISIS DE RIESGO CREDITICIO, PROPUESTA DEL MODELO CREDIT SCORING1.*

Martínez Arroyo, J. L., & Marín Rodríguez, N. J. (2021). RELACIÓN DINÁMICA ENTRE LOS CREDIT DEFAULT SWAPS Y LA DEUDA PÚBLICA. ANÁLISIS EN EL CONTEXTO LATINOAMERICANO. *Scielo*, 16.

Montalván Acaro, C. O. (2019). *Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito.* Quito: Universidad Andina Simón Bolívar.

Mostajo Castelú, S., & Vargas Sánchez, A. (2015). MEDICIÓN DEL RIESGO CREDITICIO MEDIANTE LA APLICACIÓN DE MÉTODOS BASADOS EN CALIFICACIONES INTERNAS. *Investigación & Desarrollo*, 14.

Ormaza Andrade, J. E., & Cevallos Jiménez, A. B. (2021). Estudio de factibilidad para la implementación de una tarjeta de crédito con depósito inicial en la ciudad de Cuenca. *Revista Cientifica Dominio de las Ciencias*, 1089.

Quindigalle Cuyo , N. L. (2018). *ANÁLISIS DEL RIESGO CREDITICIO Y SU INCIDENCIA EN LA RENTABILIDAD.* Latacunga : UNIVERSIDAD TÉCNICA DE COTOPAXI.

Tulcanaza Aguilar, M. Á. (2021). *Propuesta de un modelo de score de originación para la cartera de consumo de una cooperativa de ahorro y crédito del segmento 3 en el Ecuador.* Quito: Universidad Andina Simón Bolívar.

Zambrano Molina, M. A. (2021). *Estrategia para optimizar la gestión del riesgo crediticio para el manejo de la tasa de morosidad en empresas del sector comercial pertenecientes a la asociación de electrodomésticos del Ecuador.* Guayaquil: Universidad Tecnológica Empresarial de Guayaquil.