Aplicación de Modelos Predictivos en la Evaluación de Solicitudes de Crédito: Un Estudio Comparativo

Presenta:

Diego Cardeña Sosa

201919711

Director de Tesis:

Dr. Bulmaro Juárez Hernández

Tesis presentada para obtener el grado

DE

LICENCIATURA EN ACTUARÍA

Facultad de Ciencias Físico Matemáticas

Benemérita Universidad Autónoma de Puebla

Puebla, Pue.

Octubre, 2024

AGRADECIMIENTOS

Me gustaría agradecer en primer lugar a mi madre. Gracias a ti pude terminar mis estudios. Tus sacrificios, tus esfuerzos y tu capacidad para sacarnos adelante han sido mi motivación. Gracias por enseñarme a no rendirme y por estar siempre apoyándome, mostrandome mi capacidad para superar los obstáculos que se presenten.

A mi abuela y mis hermanos por estar siempre ahí.

A mi hermano de otra madre, Sebastián Alonso, por acompañarme durante la carrera y estar siempre prestándome su apoyo y motivándome en los malos momentos. A Andrea Morales por su apoyo y motivación constante a lo largo de esta tesis. Les agradezco por estar siempre pendientes de mí y por no dejar que me detuviera por ningún motivo, ni siquiera en los peores momentos.

A los amigos que me acompañaron en la carrera, las largas horas de estudio, los trabajos en equipo y los momentos de relajación dieron sus frutos... Gracias, Sarita, Jenny, Erick, por las salidas en los mejores momentos para preparar nuestros trabajos grupales. Siempre listos para el estudio, su compañía y apoyo me permitieron disfrutar el camino. A Ana, Silvia y Ale por los jueves semanales de estudio para ahogarnos en conocimiento y risas. Gracias por los buenos momentos que me permitieron recargar energías.

Finalmente, a los profesores con los que pude contar para mi desarrollo durante la carrera y que lograron transmitirme sus conocimientos, su dedicación y consejos, lograron inspirarme a continuar. Gracias.

RESUMEN

En este estudio se creó un modelo de regresión logística y un modelo de bosques aleatorios para clasificar solicitudes de crédito. Sus objetivos fueron encontrar los patrones o características que toman más peso en la aprobación de un crédito, esto mediante el uso de datos históricos de solicitudes anteriores ya clasificadas. El primer paso fue el procesamiento de los datos, eliminando valores nulos, corrigiendo entradas en los datos y codificando la variable objetivo.

El implementar el modelo de regresión logística permitió interpretar los coeficientes para las variables independientes en términos de probabilidad, además de poder identificar la manera en que estas variables influyen en la clasificación. Mientras que el modelo de bosques aleatorios resaltó en su habilidad de manejar relaciones complejas, medir la importancia individual de las variables y reducir el riesgo de sobreajuste del modelo.

Algunas de las métricas utilizadas para evaluar los modelos fueron la precisión, la exactitud, el puntaje f1 y el área bajo la curva ROC. Ambos modelos obtuvieron puntuaciones satisfactorias y en ellas se vieron reflejadas las fortalezas de ambos modelos, los bosques aleatorios con puntuaciones significativamente mejores, y la regresión logística con su facilidad de interpretar los resultados y la influencia de cada variable.

Así, comprobando, que el uso de modelos predictivos es una herramienta útil para las entidades financieras a la hora de gestionar el riesgo crediticio en relación con las decisiones crediticias.

Aclaración de ilustraciones: Todas las ilustraciones de este artículo que no están acreditadas explícitamente son ilustraciones de mi propiedad creadas específicamente para este proyecto.

Índice general

AGRADECIMIENTOS II

RESUMEN III

1 INTRODUCCIÓN 1

1.1 Contexto y Justificación 1
1.2 Objetivos de la Investigación 2
1.2.1 Objetivos Específicos 2
1.3 Alcance y limitaciones 2

2 MARCO TEÓRICO 4

2.1 Sistema de Créditos 4
2.1.1 Riesgo de Crédito 6
2.1.2 Evaluación del riesgo de crédito 8
2.2 Modelos en la evaluación del Riesgo Créditicio 10
2.2.1 Modelos Clásicos 10
2.2.2 Integrando Modelos de Machine Learning ML 12
2.3 Regresión logística 15
2.3.1 Condiciones de uso 16
2.3.2 Función Sigmoide y los Odds 16
2.3.3 El Logit 17
2.3.4 El Modelo Logístico 17
2.3.5 Ajuste de un Modelo Logístico 18
2.4 Bosques Aleatorios 19
2.4.1 Árboles de Decisiones 19
2.4.2 Algoritmo y Funcionamiento 20
2.5 Métricas de Evaluación de Modelos 21
2.5.1 Significancia del Modelo 21
2.5.2 La curva ROC (Receiver Operating Characteristic) 22
2.5.3 Optimizando los Modelos 24

3 METODOLOGÍA **29**
3.1 Descripción de los Datos 30
3.2 Limpieza y preprocesamiento de Datos 32
3.3 Desarrollo de modelos 32
3.3.1 EDA 32
3.3.2 Modelo de Regresión Logística 45
3.3.3 Modelo de Bosques Aleatorios 49

4 RESULTADOS **52**

5 DISCUSIÓN **56**
5.1 Interpretación de Resultados 56
5.2 Implicaciones Prácticas 56
5.3 Limitaciones del Estudio 57

6 CONCLUSIONES **58**

BIBLIOGRAFÍA **61**

A Repositorio de Algoritmos **65**

Capítulo 1

INTRODUCCIÓN

1.1. Contexto y Justificación

Para el sector financiero, la evaluación del riesgo de crédito es esencial, ya que esta determina si un crédito será otorgado. Las instituciones financieras utilizan modelos predictivos para determinar que tan viable es otorgar un crédito a determinado solicitante, permitiéndoles tomar decisiones informadas.

Para el caso de los modelos realizados, fueron diseñados para una financiera que se dedica a ofrecer prestamos destinados a emprendimientos con cierto tiempo de antiguedad, por lo que sirven como un apoyo para el crecimiento de pequeños y medianos negocios. Esta ofrece productos diseñados para grupos de 5 personas o para emprendimientos con al menos 2 años de antiguedad, llamados crédito “solidario” y crédito “productivo individual”. Ambos créditos están relacionados con los objetivos de la financiera para fomentar el desarrollo económico local y brindar acceso a financiamiento para la comunidad.

La digitalización y los grandes volumenes de datos han permitido utilizar técnicas de machine learning para lograr mejorar la precisión y reducir la tasa de error de los modelos predictivos, entre ellos la regresión logística y los bosques aleatorios. La regresión logística resalta por su capacidad para modelar probabilidades en base a un conjunto de variables linealmente relacionadas. Y los bosques aleatorios, al construirse a partir de múltiples árboles de decisión, logran controlar mejor la variabilidad de los datos, siendo capaces de conseguir mejores resultados de precisión.

La justificación para este estudio se encuentra en la necesidad de las instituciones financieras para reducir el riesgo que conlleva otorgar un crédito, esto mediante la mejora continua de precisión e interpretabilidad de sus modelos predictivos.

Los modelos tradicionales de evaluación no han sido capaces de captar la complejidad de los datos totalmente, ni de qué manera la interacción de las variables influye en la aprobación de créditos. Aquí los modelos basados en técnicas de machine learning pueden aportar un valor adicional.

1.2. Objetivos de la Investigación

Desarrollar y evaluar modelos predictivos utilizando técnicas de machine learning para clasificar solicitudes de crédito para optimizar el proceso de aprobación y, además, respaldar la toma de decisiones para la financiera.

1.2.1. Objetivos Específicos

- Analizar y procesar datos históricos de solicitudes de crédito previamente clasificadas para buscar y corregir inconsistencias.
- Implementar modelos de machine learning para la clasificación y compararlos: modelos de regresión logística y bosques aleatorios.
- Evaluar los modelos entrenados mediante las métricas convencionales para modelos de clasificación: precisión, exactitud y la curva de ROC, comparándolos mediante su puntaje f1.
- Optimizar los modelos mediante técnicas pre y post entrenamiento, para asegurar puntuaciones confiables.
- Probar los modelos en un subconjunto de prueba y ajustar los parámetros necesarios para maximizar la precisión y minimizar el riesgo de sobreajuste de los modelos.

1.3. Alcance y limitaciones

- El estudio se basa en crear modelos predictivos con técnicas de machine learning para clasificarlos y compararlos. Utilizando datos proporcionados por la financiera para el entrenamiento y validación.
- Se analizan detalladamente los datos en búsqueda de problemas o inconsistencias, codificando la variable objetivo y garantizando calidad y efectividad de los modelos creados.

- Se determina el desempeño de los modelos mediante métricas de evaluación de modelos de clasificación como la exactitud y la curva de ROC.
- Este estudio provee recomendaciones para la financiera que proporcionó los datos, además de considerar el impacto de la toma de decisiones y la implementación de estos modelos para optimizar procesos.
- Los datos históricos, su calidad y disponibilidad se ven reflejados en los resultados de los modelos. Los datos incompletos o con irregularidades afectan negativamente el desempeño de estos.
- Los modelos pueden presentar diferentes limitaciones, desde la interpretabilidad hasta su capacidad de manejar relaciones complejas y no lineales.
- Los modelos pueden verse afectados por diferentes factores ajenos a los datos, tanto políticos y económicos, así también como por cambios internos en las políticas de aceptación.

Capítulo 2

MARCO TEÓRICO

2.1. Sistema de Créditos

La deuda (proveniente del latín “debitum”, que significa “lo que se debe”) representa una obligación contraída entre dos partes: acreedor y deudor. Donde el acreedor presta algo de valor al deudor, con la promesa de que este último lo devolverá en un plazo determinado.

A partir de la deuda surge el crédito. Este se define como el acceso, otorgado por el acreedor al deudor, a recursos financieros más allá de los medios actuales del deudor.

En esencia, el crédito es el intercambio de bienes existentes en una fecha futura, que pueden estar ya en existencia o pueden ser producidos en el futuro. Su importancia en el sistema financiero radica en que permite a los productores obtener capital, que es toda la riqueza generada con el objetivo de satisfacer necesidades o deseos, pero que en realidad se utiliza para aumentar la productividad del trabajo en el futuro [1].

¿Cómo funciona el crédito? En la apertura de un crédito, el acreedor se compromete a proporcionar al deudor una cantidad establecida de dinero de acuerdo a términos y condiciones establecidos, y el deudor se compromete a devolver al acreedor las cantidades acordadas o pagar el monto de la deuda en el momento oportuno, así como de pagar los intereses, prestaciones, gastos y costos acordados.

El crédito y el préstamo tienen diferente función. El primero es un acuerdo flexible entre un acreedor y un deudor. En este acuerdo, el deudor puede acceder a la cantidad máxima de dinero y solo pagar intereses sobre el monto que usa.

Por otro lado, un préstamo es un monto fijo de dinero otorgado a una persona o empresa con el compromiso de devolverlo en un plazo determinado. En la mayoría de los

préstamos, los pagos son programados e incluyen tanto la cantidad principal como los intereses. Estos sitios suelen estar instalados de forma permanente y se pueden utilizar con regularidad. Además, los préstamos suelen tener una finalidad específica, como financiar un proyecto empresarial concreto o adquirir un coche o una casa.

Podemos clasificar el crédito según su forma de disposición, según su emisor o según el destino del crédito. En el siguiente esquema, podemos observar como se clasifica el crédito según el destino [2]:

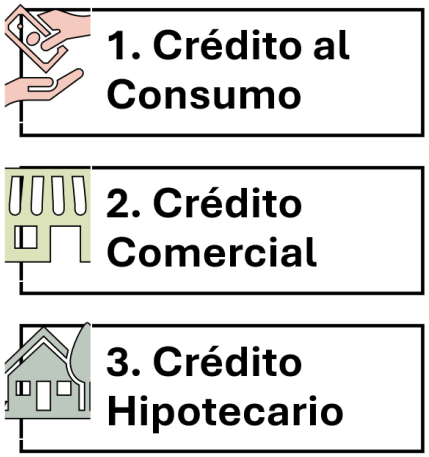


Figura 2.1: Clasificación del crédito [3].

Las definiciones, según [4], son las siguientes:

1. Crédito de Consumo

   Este es un tipo de crédito que está diseñado para financiar la adquisición de bienes de consumo y suplir diferentes necesidades: cubrir algunos gastos imprevistos, financiar parte de tus estudios, invertir en bienes y servicios, realizar un viaje, entre otros. El crédito de consumo es uno de los más populares entre las personas naturales.

Los créditos al consumo suelen ser los más fáciles de obtener a comparación de los hipotecarios o comerciales, debido a los montos más pequeños y los plazos más cortos. Estos créditos están destinados a financiar bienes y servicios personales, como compras, vacaciones, o mejoras en el hogar.

2. Crédito comercial

   Este tipo de préstamo es solicitado por personas naturales o jurídicas para financiar los costos del desarrollo de la actividad productiva y económica organizada.

Por ejemplo, si una empresa necesita comprar materia prima para realizar sus productos, con un crédito comercial puede financiar esa compra. También está la compra de maquinaria, de mercancía para la venta, o el financiamiento de cuentas por cobrar cuando se hace una venta a plazos. Un aspecto importante de este tipo de crédito es que se debe contar con la certeza de las ganancias. Es decir, para conservar la liquidez, es necesario saber que recibirás dinero de la misma actividad productiva en la que estás invirtiendo.

Debido a la variabilidad y el riesgo asociado con las empresas, la facilidad de obtener un crédito comercial suele depender de la solidez financiera de la empresa y de la confianza del prestamista en la capacidad de la empresa para generar ingresos suficientes para pagar el préstamo.

3. Crédito Hipotecario

   Es un préstamo a mediano o largo plazo que se otorga para la compra, ampliación, reparación o construcción de una vivienda. La propiedad adquirida queda en garantía o "hipotecada" a favor del banco para asegurar el cumplimiento del crédito.

En general, se presentan requisitos estrictos para estos créditos debido a los grandes montos de dinero involucrados y los largos plazos de pago. La facilidad de obtener un crédito hipotecario depende en gran medida de la política del prestamista y de la situación financiera del solicitante. Las probabilidades de aprobación aumentan con elementos como un buen historial crediticio, ingresos estables y un bajo nivel de deuda.

Sintetizando las definiciones anteriores, tenemos que el crédito al consumo es un tipo de crédito a corto plazo que se otorga a personas físicas para la adquisición de bienes o servicios con un plazo de uno a dos años. Mientras que el crédito comercial es un contrato mediante el cual las instituciones financieras proporcionan recursos (dinero) destinados específicamente a la producción, y el crédito hipotecario para la vivienda se otorga con el propósito de comprar, ampliar, construir o remodelar una vivienda.

2.1.1. Riesgo de Crédito

El riesgo crediticio se refiere a la posibilidad de que un prestamista o contraparte no cumpla con sus obligaciones financieras, lo que podría resultar en pérdidas para el prestamista. En el mundo financiero, es una consideración importante porque impacta la rentabilidad y estabilidad de las entidades financieras.

Algunas de las variables que contribuyen al riesgo crediticio son: la solvencia del prestatario, el tipo de instrumento financiero y las circunstancias económicas generales. Una gestión crediticia efectiva está compuesta por una evaluación exhaustiva de los prestatarios, la diversificación de la cartera de préstamos y la implementación de estrategias de reducción de riesgos. Es fundamental que las instituciones financieras comprendan y manejen el riesgo crediticio para asegurar la solidez de las actividades de préstamo y mantener un balance general saludable.

El riesgo crediticio se compone por varias partes importantes. El nivel de riesgo varía según la naturaleza del instrumento financiero, como su prioridad, términos y colateral. La solvencia del prestatario es un factor importante en la evaluación del riesgo crediticio, esta incluye su historial financiero, ingresos y activos. Además, la solvencia de los prestatarios y el panorama general del riesgo crediticio pueden verse afectados por condiciones económicas generales, como las tasas de interés, la volatilidad del mercado y las tendencias de la industria.

Santander [5] nos ayuda a definir los siguientes tipos de riesgo crediticio:

- Riesgo de concentración:

  Este riesgo está relacionado con la posibilidad de que la institución financiera dependa de un solo prestatario o sector económico. La institución financiera podría sufrir pérdidas significativas en caso de que ese prestatario o sector experimente dificultades financieras.

- Riesgo de default o impago:

  De acuerdo con las agencias de calificación, un impago ocurre cuando no se paga durante al menos tres meses después de la fecha en que se tuvo que pagar.

- Riesgo de exposición:

  Es la incertidumbre sobre los pagos que se deben en el futuro. Es posible que esto se deba a la situación del prestamista o a factores del mercado, así como a la variación de los intereses.

- Riesgo de tasa de recuperación:

  Varía en función de si hay garantías o colaterales en la operación. Si existe, existe un riesgo de liquidez en la garantía.

2.1.2. Evaluación del riesgo de crédito

La evaluación del riesgo crediticio es un componente esencial del funcionamiento de las instituciones financieras porque tiene un impacto directo en su capacidad para reducir las pérdidas potenciales y garantizar la sostenibilidad de sus prácticas de préstamos. Para que las instituciones financieras tomen decisiones bien informadas, gestionen eficientemente sus carteras y, en última instancia, contribuyan a la estabilidad general del sistema financiero, es esencial comprender la importancia de la evaluación del riesgo crediticio y los factores principales que impulsan este proceso.

Chatterjee [6] define el riesgo de crédito como el asociado al incumplimiento de un pago contratado.

La evaluación del riesgo crediticio es importante porque ayuda a las instituciones financieras a identificar, medir y administrar los riesgos relacionados con las actividades de préstamo, lo que les permite tomar decisiones informadas, asignar recursos de manera eficiente y mantener un equilibrio saludable entre riesgo y rentabilidad.

Las instituciones financieras evalúan el riesgo crediticio con el fin de determinar la capacidad de los prestatarios para pagar los préstamos, o lo que es la probabilidad de incumplir sus obligaciones y evitar un impacto negativo en la salud de la institución.

En conjunto, estos cinco factores: capacidad, capital, colateral, carácter y cumplimiento, proporcionan una evaluación integral de la solvencia del prestatario. Al evaluar exhaustivamente cada uno de estos elementos, las instituciones financieras pueden tomar decisiones informadas sobre si extender el crédito y bajo qué términos. Comprender cómo los prestamistas utilizan los Cinco C's ayuda a los prestatarios a prepararse y posicionarse para obtener el crédito que necesitan.

Las instituciones financieras pueden identificar prestatarios de alto riesgo al evaluar el riesgo crediticio y decidir negarles el crédito o ofrecerlo a una tasa de interés más alta para compensar el riesgo aumentado. Este proceso permite a las organizaciones establecer límites de crédito apropiados, monitorear las carteras de préstamos y tomar decisiones informadas sobre la aprobación de préstamos, fijación de precios y estrategias de cobro.

Además, Chatterjee [6] señala que las instituciones bancarias, en particular, crean modelos de riesgo de crédito debido a la necesidad de determinar la cantidad de capital económico necesaria para mantener las operaciones de toma de riesgo de un banco.

Finalmente, para lograr objetivos estratégicos, mantener una cartera de préstamos saludable y contribuir a la estabilidad general del sistema financiero, las instituciones

financieras deben realizar una evaluación precisa del riesgo crediticio. A continuación, en la Figura 2.2, se observa la distribución de la pérdida de una cartera de crédito:

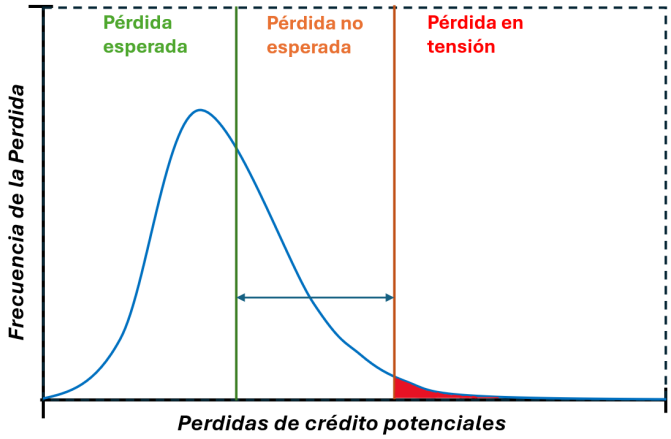


Figura 2.2: Distribución de la pérdida de una cartera de crédito.

A esta distribución, según Chatterjee [6], se le conoce como una técnica que relaciona el capital económico que en general se requiere para el riesgo de crédito con la función de densidad de probabilidad (fdp) de pérdidas crediticias de la cartera, con el propósito de determinar la cantidad de capital económico necesaria para mantener las actividades con riesgo de crédito.

Como podemos observar en la Figura 2.2, esta distribución se descompone en tres tipos de pérdidas:

- **Pérdida Esperada**:

La pérdida esperada es la cantidad de dinero que una institución financiera anticipa que perderá debido a préstamos que no se pagarán. Como se menciona en su nombre, ayuda a los bancos a prepararse para las pérdidas que se esperan y a asegurarse de que tienen suficiente dinero en reserva para cubrir estas pérdidas. Según Bonás Piella et al. [7], la pérdida esperada es el resultado del producto de tres variables:

1. Probabilidad de Incumplimiento (PD): ¿Qué tan probable es que el prestatario no pague el préstamo?
2. Exposición al Momento del Incumplimiento (EAD): ¿Cuánto dinero se debe en el momento en que el prestatario incumple?
3. Pérdida Dada el Incumplimiento (LGD): ¿Qué porcentaje del préstamo se perderá después de intentar recuperar el dinero?

- **Pérdida Inesperada**:

La pérdida no esperada es el dinero extra que una institución podría perder debido a eventos inesperados o situaciones inusuales. Su propósito es ayudar a los bancos a tener una reserva de capital adicional para cubrir pérdidas imprevistas que superen las expectativas normales. Los siguientes son factores de riesgo que pueden afectar las pérdidas imprevistas [7]:

+ La volatilidad de exposición.
+ La volatilidad de probabilidad de default.
+ La volatilidad de la severidad.
+ La concentración de la exposición.
+ Las correlaciones.

- **Pérdida en Tensión**:

La pérdida en tensión es una estimación de cuánto podría perder un banco en condiciones extremas, como una crisis económica o financiera. Esta ayuda a los bancos a evaluar su capacidad para lidiar con situaciones extremadamente difíciles y asegurarse de que tienen suficiente capital para mantenerse solventes incluso durante momentos de gran estrés financiero.

Para mantener sus licencias operativas y evitar sanciones regulatorias, las instituciones financieras deben demostrar su capacidad para gestionar el riesgo de manera efectiva. Por lo tanto, la evaluación del riesgo crediticio es crucial para cumplir con los requisitos regulatorios.

2.2. Modelos en la evaluación del Riesgo Créditicio

2.2.1. Modelos Clásicos

Para obtener la aprobación de un crédito, las instituciones financieras evalúan cuidadosamente factores clave, algunos de estos factores son considerados dentro los primeros modelos, tales como los que comprenden el modelo de las “ Cinco C’s del crédito”. Comprender estos criterios puede ayudar a asegurar solvencia crediticia.

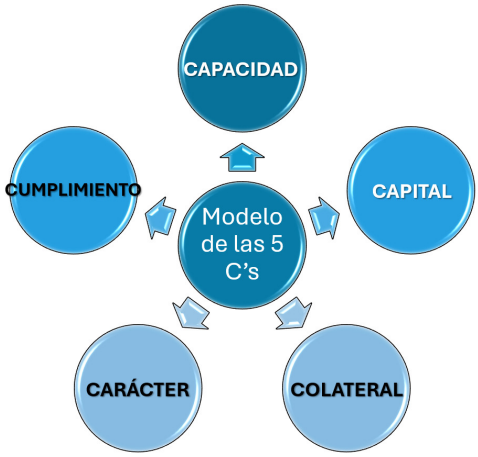


Figura 2.3: Cinco C's del crédito.

La capacidad es la habilidad del prestatario para hacer pagos puntuales y conscientes del préstamo [8]. Esto se evalúa mediante un análisis de los ingresos del solicitante, su historial laboral y sus obligaciones de deuda actuales.

El capital examina el patrimonio neto del prestatario y la cantidad de capital o ahorros que tiene disponibles, ya que esto demuestra su estabilidad financiera y su capacidad para enfrentar eventos inesperados [9].

El colateral es cualquier activo, como bienes raíces o propiedades personales valiosas, que puede usarse como garantía para el préstamo [9]. Esto brinda al prestamista una opción de respaldo en caso de que el prestatario no cumpla con sus obligaciones.

El carácter evalúa la reputación del solicitante, su solvencia crediticia y su historial de manejo de responsabilidades de deuda. Finalmente, el cumplimiento considera factores externos que podrían impactar la capacidad del prestatario para reembolsar, incluyendo las condiciones económicas generales, las tendencias de la industria y las propias políticas y regulaciones del prestamista [10].

A diferencia del modelo anterior que no podría clasificarse propiamente como estadístico, los siguientes dos modelos han ganado amplio reconocimiento en este ámbito. Son el modelo de Z-Score o Credit Scoring y el modelo ZETA, cada uno ofreciendo un enfoque distinto para evaluar la probabilidad de incumplimiento.

El modelo Z-Score, pionero de Edward Altman en la década de 1960, es un modelo matemático multivariable que combina varios ratios financieros para producir una única puntuación, que sirve como indicador de la salud financiera de una empresa y su probabilidad de incumplimiento. Este modelo se compone de las siguientes variables o ratios [11, 12, 13]:

Figura 2.4: Modelo Z - Score.

1. La relación de capital de trabajo a activos totales, que mide la capacidad de una empresa para cumplir con sus obligaciones a corto plazo y su liquidez general;
2. la relación de ganancias retenidas a activos totales, que refleja la rentabilidad de la empresa y su capacidad para reinvertir ganancias;
3. la relación de ganancias antes de intereses e impuestos a activos totales, que indica la eficiencia operativa de la empresa y su capacidad para generar beneficios;
4. la relación del valor de mercado del capital a la contabilidad del valor de los pasivos totales, que representa la capacidad de la empresa para soportar dificultades financieras;
5. y la relación de ventas a activos totales, que mide la utilización de los activos de la empresa .

El modelo Z-Score ha sido ampliamente adoptado por instituciones financieras, ya que proporciona una herramienta valiosa para identificar empresas financieramente en apuros, guiar estrategias de recuperación e informar decisiones legales relacionadas con la quiebra [14].

En conjunto con el modelo ZETA basado en el modelo Z-Score desarrollado en 1970 por Altman y sus colegas, estos modelos fueron pioneros en la mejora de la evaluación del riesgo crediticio, brindando herramientas útiles a las instituciones para la mejora en la precisión de la evaluación.

2.2.2. Integrando Modelos de Machine Learning ML

Clasificamos a los modelos de machine learning en dos categorías principales de aprendizaje: modelos de aprendizaje supervisado y modelos de aprendizaje no supervisado.

Los modelos de aprendizaje supervisado se caracterizan por utilizar conjuntos de datos etiquetados para el entrenamiento de los modelos.

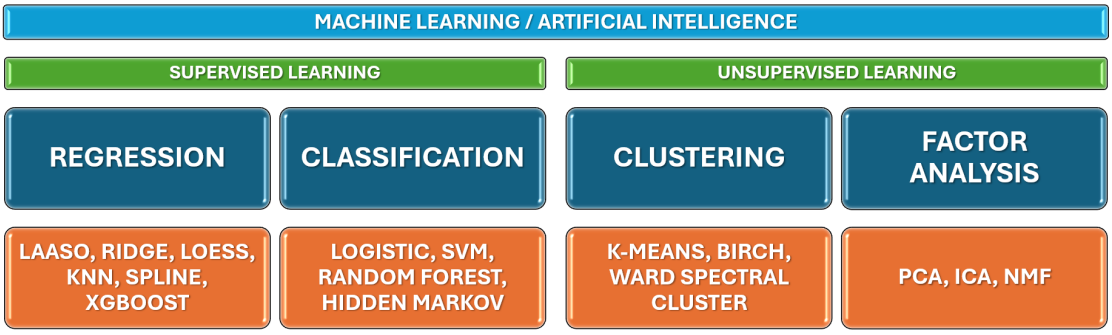


Figura 2.5: Tipos de Aprendizaje [15].

Por datos etiquetados nos referimos a aquellos que ya cuentan con el resultado correcto, distinguiéndolos de prestatarios que cumplen con sus obligaciones de los que no. Los modelos de aprendizaje no supervisado utilizan conjuntos de datos no etiquetados; en este caso se busca que el modelo logre descubrir patrones ocultos o estructuras en los datos que puedan señalar comportamientos de riesgo poco evidentes a simple vista.

Para abordar el riesgo de crédito, se han utilizado tradicionalmente técnicas estadísticas; sin embargo, es difícil aplicarlas a grandes cantidades de datos, lo que puede poner en peligro su estimación si se utilizan de manera exclusiva. Para evitarlo, contamos con diferentes técnicas de evaluación del riesgo de crédito (Figura 2.6).

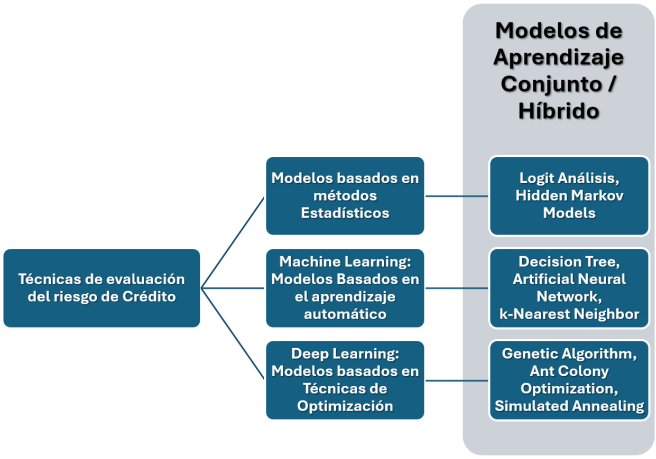


Figura 2.6: Técnicas de Evaluación [16].

Los siguientes son ejemplos puntuales de algunas de las técnicas aplicadas en la evaluación del riesgo de crédito.

1) Construcción de ScoreCard para evaluar la probabilidad de impago [17].

a) Estudio:

Hand & Henley describieron el proceso estándar para construir una ScoreCard, involucrando técnicas como la regresión lineal, la regresión logística y el análisis discriminatorio lineal. Estas técnicas permiten evaluar la probabilidad de impago de los consumidores de manera efectiva.

Se mencionan las diferentes condiciones y situaciones que afectan esta manera de evaluar el riesgo, como al establecer límites muy conservadores para la cesión de un crédito puede implicar el pago siempre puntual de los prestatarios. Se debe tener en cuenta que, aunque esto podría parecer un objetivo agradable, el propósito principal es maximizar las ganancias y esto no está únicamente relacionado con el riesgo. No se podrán aplicar intereses adicionales en un prestatario que no tenga ningún retraso; esto disminuirá el beneficio que se obtenga.

b) Conclusión:

La aplicación de estas técnicas en la construcción de ScoreCards ha demostrado ser efectiva para predecir el riesgo de impago. En la evaluación del riesgo crediticio, los métodos descritos por Hand y Henley han mejorado significativamente la precisión de las evaluaciones del riesgo de impago.

Al encontrar un equilibrio que maximice las ganancias mediante la gestión adecuada del riesgo, la evaluación del riesgo no se limita en evitar impagos. Con este método, las instituciones financieras pueden tomar decisiones más informadas sobre la concesión de créditos, modificando sus políticas para maximizar la estabilidad financiera y la rentabilidad.

2) Detección de Fraude en Tarjetas de Crédito utilizando SVM (Support Vector Machines) [18].

a) Estudio:

En este contexto, el fraude se refiere a la extracción ilegal de fondos de una cuenta de crédito o a la adquisición de bienes sin pagar por ellos. Se pueden emplear métodos de prevención o detección, identificando transacciones que difieren del patrón de compra “normal” del cliente mediante la utilización de la información histórica de los clientes como datos de entrenamiento.

Las SVM (Support Vector Machines) son un método comúnmente empleado para la clasificación y el reconocimiento de patrones en la detección de fraude. Este método

tiene como objetivo maximizar la distinción entre las dos categorías: transacciones no fraudulentas y transacciones fraudulentas.

Se construye un margen en base a la distancia entre dos hiperplanos, que reduce el error de clasificación, se construye como la función de Lagrange para los datos lineales y una función Kernel para los datos no lineales. Las transacciones que difieren del comportamiento habitual de compra se consideran fraudulentas según el patrón de comportamiento de los clientes.

b) Conclusión:

El uso de SVM para detectar fraude en tarjetas de crédito ha demostrado ser efectivo; logra una alta precisión en la detección de transacciones fraudulentas. Ajustándose a los patrones de comportamiento de los usuarios y mejorando constantemente a medida que se acumulan más datos históricos, los algoritmos de aprendizaje automático como SVM permiten una detección de fraude más efectiva y precisa.

2.3. Regresión logística

La regresión logística es una herramienta estadística fundamental para analizar y modelar situaciones donde la variable dependiente es binaria, es decir, tiene dos posibles resultados, como éxito/fracaso o presencia/ausencia. Cuando deseamos relacionar un conjunto de variables X con una variable dependiente, estamos abordando un problema multivariable. En el análisis de este tipo de problemas, generalmente se utiliza este tipo de modelo matemático para manejar las complejas interrelaciones entre numerosas variables.

Según Camarero Rioja, Almazán Llorente y Mañas Ramírez [19], la regresión logística tiene una gran importancia en la investigación sociológica debido a su capacidad para analizar las relaciones entre variables categóricas.

“Logistic regression is by far the most popular modeling procedure used to analyze epidemiologic data when the illness measure is dichotomous.” [La regresión logística es ampliamente preferida como el procedimiento principal de modelado para analizar datos epidemiológicos cuando la variable de resultado es dicotómica.] [20].

Este modelo se suele utilizar en distintas disciplinas, esto debido a su capacidad para manejar variables categóricas y la interpretabilidad facilitada en términos de probabilidades y odds ratios.

2.3.1. Condiciones de uso

La regresión logística, además de ser considerada una técnica de machine learning, ésta tiene sus bases en la estadística y la probabilidad. Por lo mismo, a diferencia de otros, modelos, este se ve limitado por condiciones o supuestos que deben cumplir los datos a utilizar. Estos supuestos son los siguientes:

- La variable de respuesta solo toma dos resultados posibles.
- Las observaciones son independientes.
- Se deben evitar correlaciones altas entre las variables independientes con el fin de evitar multicolinalidad y provocar incertidumbre sobre la influencia de cada variable.
- Los datos deberán contener solo observaciones validas sin valores atipicos o significativos.
- Debe existir una relación lineal entre cada variable explicativa y el logit de la variable de respuesta.
- La muestra del conjunto de datos debe ser lo suficientemente grande como para llegar a conclusiones válidas sobre el modelo de regresión logística ajustado.

2.3.2. Función Sigmoide y los Odds

La regresión logística utiliza una función cuyo resultado siempre está entre 0 y 1 para transformar el valor devuelto por la regresión lineal. La función logística (también conocida como función sigmoide) es una de las funciones más comunes que cumplen esta descripción:

$$sigmoide = \theta(y) = \frac{1}{1 + e^{-y}}$$

En la regresión logística, el concepto fundamental de “odd” es crucial para entender la probabilidad de eventos. El “odd” representa la relación entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra [19], expresada matemáticamente como

$$Odd = \frac{p}{1 - p}.$$

Donde p es la probabilidad de ocurrencia del evento. Este valor proporciona una medida directa de la probabilidad relativa de que ocurra el evento en comparación con que no ocurra.

2.3.3. El Logit

El logit es otra transformación crucial en la regresión logística. Es el logaritmo del odd; la base de la función logaritmo usada en este estudio es poco importante siempre que sea mayor que 1. Sin embargo, se usa con frecuencia el logaritmo natural con base e. La función logit transforma la proporción de ocurrencia de un evento en una escala que puede tomar cualquier valor real, lo cual es útil para el análisis estadístico.

Esta transformación permite modelar la relación entre las variables independientes y la probabilidad del evento de una manera lineal. Según Dangeti [21], la regresión logística aplica la estimación de máxima verosimilitud después de transformar la variable dependiente en una variable logit (logaritmo natural de las probabilidades de que ocurra o no la variable dependiente) con respecto a las variables independientes.

Esto nos permite suponer un modelo lineal sin restricciones, gracias a que el logit puede tomar cualquier valor real, $Rangeof Log(odds) = [-\infty, \infty]$.

$$Log(odds) = Log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + ... + \beta_k x_k.$$

2.3.4. El Modelo Logístico

La capacidad del logit para transformarse linealmente permite derivar la función logística, fundamental en el modelo logístico:

$$P(x) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i x_i)}}.$$

En esta ecuación, $z = \alpha + \sum \beta_i x_i$ que contiene los términos α y β_i representan parámetros desconocidos que se estiman utilizando los datos de las variables x's y el resultado de la variable dicotómica.

Los modelos de regresión logística tienen tres objetivos principales [22]:

- Cuantificar la relación entre cada variable independiente y la variable dependiente.

- Detectar interacciones y confusiones entre las variables independientes en relación con la variable dependiente (mediante odds ratios de x_i).
- Clasificar dentro de las categorías de la variable dependiente.

Así, la función logística transforma las predicciones en probabilidades (Figura 2.7), determinando el valor de la variable categórica objetivo. Esta transformación es crucial porque permite modelar de manera lineal la relación entre las variables independientes y la probabilidad de ocurrencia de un evento en la escala del logit.

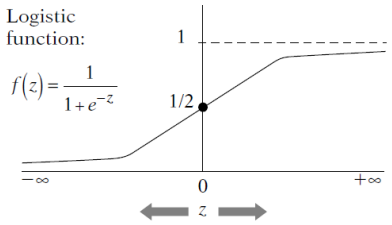


Figura 2.7: Gráfica de la función logística [20].

2.3.5. Ajuste de un Modelo Logístico

Ajustar un modelo de regresión logística implica determinar los valores de los coeficientes que mejor representan la relación entre las variables independientes y la variable dependiente en los datos observados. Como mencionamos antes, este proceso se realiza utilizando el método de máxima verosimilitud, que busca encontrar los parámetros que maximizan la probabilidad de que los datos observados se hayan producido bajo el modelo especificado.

Maximizar la función de verosimilitud $L(\theta)$ generalmente equivale a maximizar el logaritmo natural de $L(\theta)$, lo cual simplifica los cálculos. Los componentes de θ se determinan resolviendo ecuaciones de derivadas parciales e igualando a cero.

Al buscar ajustar el modelo, estamos considerando una función de verosimilitud

$$L(\theta) = \frac{\prod_{l=1}^{n} e^{(\alpha+\sum_{i=1}^{k}\beta_i x_i l)}}{\prod_{l=1}^{n}[1+e^{(\alpha+\sum_{i=1}^{k}\beta_i x_i l))}]}.$$

Minimizando la diferencia entre las predicciones del modelo y los valores reales observados, logrando así una representación precisa de la probabilidad de ocurrencia del evento binario en función de las variables explicativas.

2.4. Bosques Aleatorios

Un bosque aleatorio es una colección de árboles de decisión que se entrenan utilizando diferentes subconjuntos de datos y características. El conjunto de técnicas $Bootstrap$ y $Aggregation$ o mejor conocido como $Bagging$ hace que el modelo sea más preciso y más fuerte.

La idea en los bosques aleatorios es mejorar la reducción de varianza del bagging reduciendo la correlación entre los árboles, sin aumentar demasiado la varianza [23].

Los bosques aleatorios suelen ser más precisos que los árboles de decisión individuales debido a la combinación de múltiples árboles. Cada árbol en un bosque aleatorio hace predicciones diferentes basadas en diferentes subconjuntos de datos y subconjuntos de características. Al promediar o votar sobre las predicciones de todos los árboles, los bosques aleatorios tienden a producir resultados más precisos y confiables.

2.4.1. Árboles de Decisiones

Los árboles que componen el Bosque son estructuras jerárquicas donde cada nodo de decisión representa una "pregunta" en base a una característica y cada hoja representa una "respuesta" o clase. Un modelo que puede anticipar el valor de una variable objetivo mediante el uso de reglas de decisión elementales derivadas de los datos.

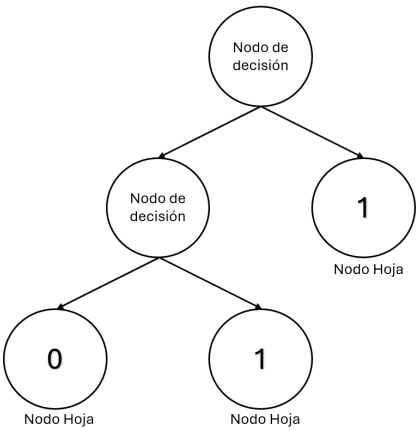


Figura 2.8: Esquema de Arbol de Decisión [3].

La creación de árboles de decisión cuenta con el beneficio de asemejarse a poder tomar decisiones basadas en un conjunto de preguntas formuladas consecutivamente. El árbol comienza con un nodo raíz y va pasando por ramas hasta llegar a los nodos terminales o a las hojas, donde se toma la decisión final.

El primer paso es identificar el nodo raíz, que contiene el conjunto de datos completo. Luego se analizan todas las características disponibles para determinar cuál divide los datos en dos o más grupos homogéneos en relación con la variable objetivo. La mejor división está determinada por varios factores, incluida la ganancia de información, la reducción de impurezas de Gini o la reducción de la varianza, de las cuales se hablará más adelante.

Este proceso continúa hasta que se cumple un criterio de detención, como alcanzar una profundidad máxima predefinida, tener un número mínimo de muestras en un nodo, o no poder mejorar la homogeneidad de los datos con más divisiones. En los nodos terminales, se asigna una clase para la tarea de clasificación o un valor promedio para la tarea de regresión.

2.4.2. Algoritmo y Funcionamiento

Como mencionamos anteriormente, el funcionamiento de los bosques aleatorios se basa en dos conceptos clave: el muestreo aleatorio de datos o "Bootstrap", cuya idea básica según [24] es extraer aleatoriamente conjuntos de datos con reemplazo de los datos de entrenamiento, cada muestra del mismo tamaño que el conjunto de entrenamiento original.

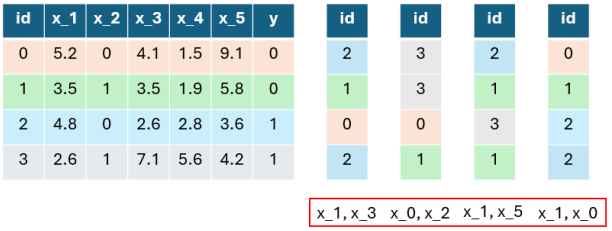


Figura 2.9: Método de Remuestreo "Bootstrap".

Una vez que todos los árboles del bosque se han entrenado en base a una selección aleatoria de características por cada conjunto de entrenamiento $Bootstrap$, en el caso de una tarea de clasificación, cada árbol del bosque emite una "votación" por la clase predicha, y la clase final predicha por el bosque es aquella que recibe la mayoría de los votos. A dicha combinación o promedio de resultados la llamamos $Aggregation$ y finalmente al uso de las 2 técnicas mencionadas anteriormente lo denominamos $Bagging$.

El $bootstrap\,aggregation$, o $bagging$, es un procedimiento de propósito general para reducir la varianza de un método de aprendizaje estadístico [23].

2.5. Métricas de Evaluación de Modelos

Al evaluar el rendimiento de un modelo de regresión logística, hay varias métricas clave a considerar. Las pruebas de significancia del modelo contra el modelo nulo y los predictores individuales son cruciales para evaluar el ajuste general del modelo y la importancia de las variables independientes. Además, los índices de bondad de ajuste, como las medidas descriptivas e inferenciales, pueden proporcionar información valiosa sobre la capacidad del modelo para representar con precisión los datos observados [25].

Otro aspecto importante es la capacidad del modelo para hacer predicciones precisas, lo cual puede evaluarse mediante el examen de las probabilidades predichas [25].

Más allá de estas métricas comúnmente utilizadas, hay una variedad de otras medidas de rendimiento que pueden emplearse para evaluar los modelos de regresión logística. Estas incluyen métricas como el error cuadrático medio, el error absoluto medio, el coeficiente de correlación de Pearson y el coeficiente de determinación, entre otras [26].

Cada una de estas métricas tiene sus propias ventajas, desventajas y limitaciones, y la elección de cuál usar depende de los objetivos específicos de la investigación y las características de los datos [26].

2.5.1. Significancia del Modelo

La razón de verosimilitud se utiliza en la modelización estadística para comparar el ajuste de dos o más modelos anidados. Para realizar esta prueba, se compara la probabilidad de obtener los valores observados ($log\,likelihood$) con el modelo de interés ($m1$) en comparación con la probabilidad de obtener los valores observados con un modelo sin predictores ($m0$), o lo que sería;

$$LLR = -2log\left(\frac{L(m0)}{L(m1)}\right) = 2(logL(m1) - logL(m0)).$$

La razón de verosimilitud sigue una distribución chi-cuadrada y los grados de libertad son iguales al número de parámetros en los que difieren los dos modelos [27]. La hipótesis nula y alternativa de esta prueba generalmente se describen como: $H_0 : \beta_1 = ... = \beta_i = 0 \quad \& \quad H_a$: almenos un $\beta_i \neq 0$.

Si la prueba es significativa, esto indica que el modelo es útil, pero no el mejor. Es posible que alguno de sus predictores no fuera necesario.

El pseudo R-cuadrado se utiliza como una alternativa al R-cuadrado en el contexto de regresiones logísticas. El pseudo R-cuadrado mide, de manera similar al R-cuadrado, la proporción de variabilidad en las respuestas observadas que es explicada por las variables explicativas. A diferencia del R-cuadrado, el pseudo R-cuadrado es solo una medida relativa; no proporciona una medida absoluta del ajuste del modelo [28].

Para calcularlo, se compara la principal contribución del modelo con la contribución realizada por un modelo nulo. El más conocido es el propuesto por McFadden's:

$$R^2_{Mcf} = 1 - \frac{ln\hat{L}(modelo)}{ln\hat{L}(modelo\,nulo)}.$$

Siendo $\hat{L}$ el valor de $likelihood$ de cada modelo. La comparación de las contribuciones se realiza a través de la comparación de las sumas de los errores al cuadrado y las verosimilitudes, lo que corresponde a una comparación de probabilidades mediante proporciones [29]. Si el modelo candidato no mejora al modelo nulo, R^2_{Mcf} tiene valor 0 y tiene valor 1 si se ajusta perfectamente a los datos.

2.5.2. La curva ROC (Receiver Operating Characteristic)

La curva ROC es una representación gráfica que ilustra la capacidad de un modelo de clasificación binaria para distinguir entre clases. Se obtiene trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de decisión.

Para concebir de mejor manera estos conceptos (TPR, FPR), comenzamos definiendo a la matriz de confusión, una matriz que permite visualizar el desempeño del modelo, mostrando los valores de las predicciones correctas e incorrectas en cada clase.

	Predicción: SI	**Predicción: NO**
Valor Real: SI	TP: VERDADERO POSITIVO	FN: FALSO NEGATIVO
Valor Real: NO	FP: FALSO POSITIVO	TN: VERDADERO NEGATIVO

Figura 2.10: Matriz de confusión.

A partir de estos valores podemos medir la precisión "qué tan bien el modelo predice

una categoría específica" y la exactitud "cuántas veces el modelo acertó sus predicciones en general";

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precisión = \frac{TP}{TP + FP}.$$

Mientras que la matriz de confusión proporciona una visión detallada de las predicciones para un umbral específico, la curva ROC ofrece una perspectiva global del rendimiento del modelo en todos los posibles umbrales utilizando las métricas;

$$Tasa\,de\,Verdaderos\,Positivos \text{ (TPR ó Sensibilidad)} = \frac{TP}{TP + FN}.$$

TPR es la relación entre las predicciones positivas correctas y el número total de predicciones positivas.

$$Puntaje\,F1 = \frac{2 * \text{Precisión} * \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}.$$

Este puntaje es la medida armónica de la sensibilidad y la precisión. Mejor puntaje, mejor modelo.

$$Tasa\,de\,falsos\,positivos \text{ (FPR ó "1 - Especificidad")} = \frac{TN}{FP + TN}.$$

La curva ROC tendrá el siguiente aspecto:

Al variar el umbral que determina si una instancia pertenece a la clase positiva, se obtienen diferentes combinaciones de TPR y FPR, que se representan en el gráfico ROC. Entonces el área bajo la curva ROC (AUC) es una medida agregada del rendimiento del modelo en todos los umbrales posibles. Un modelo perfecto tiene un AUC de 1, mientras que un modelo que no tiene capacidad discriminativa (equivalente a adivinar) tiene un AUC de 0.5. Y el punto en la curva ROC más cercano al vértice superior izquierdo (donde TPR es 1 y FPR es 0) es considerado el punto de operación óptimo, balanceando correctamente los verdaderos positivos y falsos positivos .

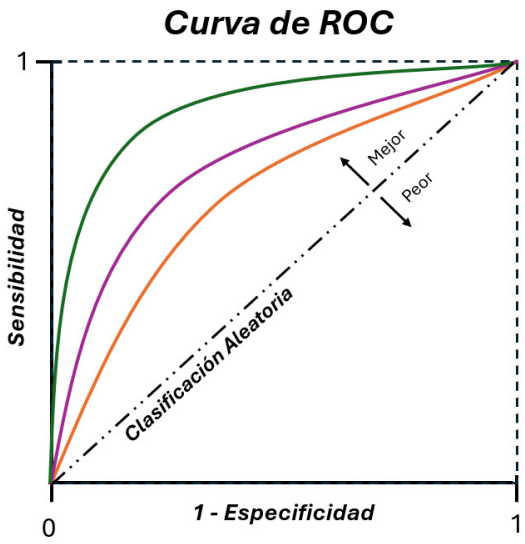


Figura 2.11: Curva de ROC.

2.5.3. Optimizando los Modelos

- INFORMATION VALUE

  El valor de la Información, o Information Value (IV), es una métrica crucial en el análisis predictivo y la modelización estadística, especialmente en contextos como el credit scoring y la evaluación de riesgos. Esta métrica evalúa la aptitud predictiva de una variable en función de la variable objetivo.

  El objetivo principal del Information Value consiste en mostrar cuánta información contribuye cada una de las variables predictoras sobre la variable objetivo

$$InformationValue = ln\left(\frac{\%Bueno}{\%Malo}\right) * (\%Bueno - \%Malo).$$

  Este valor se obtiene a través de la proporción de eventos positivos y la proporción de eventos negativos en cada nivel de la variable predictora.

  Es aquí donde esta medida se convierte en una parte fundamental para la fase de selección de variables para la construcción del modelo, puesto que mientras mayor sea el Information Value, mayor será la conexión predictiva de la variable con la variable objetivo.

Valor de Información	Poder Predictivo
<0.02	No es útil
0.02 a 0.01	Débil
0.01 a 0.3	Medio
0.3 a 0.5	Fuerte
>0.5	Poder Predictivo SOSPECHOSO

Figura 2.12: Clasificación de Poder Predictivo.

- P VALUES

  Recordemos que el valor p nos indica la probabilidad de tener nuestros resultados bajo la hipótesis nula. Resulta de mucha utilidad para evaluar la significancia estadística de las variables dentro de un modelo de regresión logística [25, 30].

  El estadístico de Wald mide la relación entre el coeficiente de regresión con su error estándar [25]. además sigue una distribución chi-cuadrada, Como bien sabemos, en un modelo de regresión logística el valor p se calcula en base de este estadístico, ya que se obtiene comparando la distribución chi cuadrada con el estadístico de Wald [31].

  Específicamente, la estadística de Wald se calcula de la siguiente manera [31]:

$$WaldStatistic = \frac{\widehat{\beta}}{SE(\widehat{\beta})}.$$

  El valor p se determina comparando la estadística de Wald con una distribución chi-cuadrado con un grado de libertad.

  Si el valor p para una variable predictora es menor que el nivel de significancia elegido (por ejemplo, 0.05), indica que la variable es estadísticamente significativa y tiene una contribución significativa al modelo de regresión logística.

- FACTOR DE INFLACION DE LA VARIANZA

  El Factor de Inflación de la Varianza (VIF, por sus siglas en inglés) es una métrica estadística fundamental para detectar la presencia de multicolinealidad en un conjunto de datos. La multicolinealidad ocurre cuando dos o más variables independientes en un modelo de regresión están altamente correlacionadas, lo que puede complicar la interpretación de los coeficientes de regresión y afectar la estabilidad del modelo.

El VIF mide cuánto se inflan las varianzas de los coeficientes de regresión debido a la multicolinealidad. En otras palabras, indica cuánto de la variabilidad de una variable independiente se explica por las demás variables independientes del modelo. Cuanto mayor sea el VIF, mayor será la multicolinealidad presente en el modelo, y por lo tanto, más problemático puede ser.

Para calcular el VIF de una variable independiente, se realiza una regresión auxiliar donde esa variable se convierte en la variable dependiente y todas las demás variables independientes se utilizan como predictores. El VIF para cada variable se puede calcular utilizando la fórmula

$$VIF(\widehat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

donde $R^2_{X_j|X_{-j}}$ es el R^2 de una regresión de X_j sobre todos los demás predictores [24].

La interpretación del VIF es relativamente directa. Un VIF menor o igual a 1 indica que no hay multicolinealidad, mientras que valores entre 1 y 4 sugieren que la multicolinealidad es tolerable. Para valores mayores a 4 el VIF nos indica una multicolinealidad considerable; para un VIF mayor a 10 este sugiere ya una severa multicolinealidad. Para contextos financieros o bancarios se suele optar por un VIF superior a 2 para eliminar una variable.

Las consecuencias de la multicolinealidad en los modelos se ven reflejadas en los valores inflados de las varianzas de los coeficientes de regresión, reduciendo la precisión de las estimaciones. Otra consecuencia de las malas estimaciones es sobre las variables, haciendo parecer significativas a variables que no lo son y viceversa. Estas consecuencias contribuyen a que el modelo se vuelva inestable, haciendo que pequeños cambios en los datos provoquen grandes variaciones en los resultados.

- IMPORTANCIA DE CARACTERÍSTICAS

El índice Gini es una métrica que mide la impureza o la heterogeneidad de una muestra. En el contexto de los árboles de decisión, se utiliza para evaluar la calidad de una división o partición de los datos. El índice Gini se define como:

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2$$

donde p_i es la proporción de observaciones de la clase i en el conjunto de datos D.

El índice Gini lo utilizamos para calcular la importancia individual de cada variable en el modelo, esto a través de la importancia de Gini, midiendo la disminución del índice Gini al eliminar una variable del modelo [32]. Para el contexto del modelo de bosques aleatorios, esta importancia de Gini suele ser utilizada para clasificar las variables predictoras según su impacto en el modelo [33].

Los modelos en los que se suele utilizar el índice Gini se llaman CART o también conocidos como Classification and Regression Trees. Para cada nodo en el árbol, el índice Gini se calcula para cada posible división, seleccionando la que minimice la impureza del nodo. La reducción dada por la división provocada por cierta característica es lo que determina su importancia.

Una métrica a considerar para el cálculo de importancia de una variable, además del índice Gini, es la conocida cross-entropy [34].

También conocida como entropia cruzada, o perdida logaritmica o log loss, esta medición mide al igual que el índice Gini la calidad de las divisiones en árboles de decisión y otros modelos de clasificación. Matematicamente se define como:

$$Cross\,Entropy = -\sum_{i=1}^{n} p_i log(p_i).$$

En los modelos de árboles de decisión, es utilizada para distinguir las mejores divisiones en los nodos de los árboles. Calculandose en base de la reducción en la entropia lograda al dividir un nodo con cierta característica.

- CROSS VALIDATION

  La validación cruzada en el contexto de machine learning es una técnica que estima el error dada una prueba, un error asociado al ajuste de algún modelo de aprendizaje dado un conjunto de datos. Esta técnica consiste en dividir al conjunto en dos partes: un conjunto de entrenamiento y otro de validación o prueba. El modelo se ajusta en el conjunto de entrenamiento, y el modelo ajustado se utiliza para predecir las respuestas de las observaciones en el conjunto de validación.

  La validación cruzada con K dobleces (k-fold cross-validation) es una de las formas más populares de validación cruzada. En este método, los datos se dividen en K subconjuntos (o “dobleces”) de tamaño aproximadamente igual, la estimacion se calcula promediando los siguientes valores

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$

El modelo se entrena K veces, cada vez utilizando K-1 dobleces como datos de entrenamiento y el doblez restante como datos de validación. Este proceso se repite K veces, cambiando cada vez el doblez de validación.

Capítulo 3

METODOLOGÍA

Este estudio es de tipo cuantitativo y utiliza una técnica analítica. Se evaluó el riesgo crediticio mediante la aplicación de modelos estadísticos y de aprendizaje automático, específicamente la regresión logística y los bosques aleatorios. Utilizando datos de solicitudes de crédito anteriores, el análisis cuantitativo permitió medir, evaluar y comparar los resultados de los modelos.

El enfoque de esta investigación se centra en el uso de técnicas de aprendizaje automático para la prevención y la ayuda ante el riesgo crediticio. Para determinar cuál modelo ofrece una mayor precisión y robustez en la clasificación para la prevención del riesgo de incumplimiento, los modelos de regresión logística y los modelos de bosques aleatorios han sido implementados y comparados.

El estudio tiene un horizonte temporal transversal, ya que analiza datos históricos de solicitudes de crédito en un punto específico del tiempo. Los datos utilizados cubren un período de tiempo específico, lo que permite evaluar el rendimiento de los modelos durante un período de tiempo determinado.

El proceso de análisis de la muestra se realizó utilizando herramientas de software estadístico y de análisis de datos como $Python$, con librerías especializadas como $matplotlib$, $seaborn$, $sklearn$, pandas y $numpy$. Estas herramientas facilitaron la gestión del volumen de datos y permitieron un análisis detallado y preciso de la muestra seleccionada.

Criterios de Inclusión

- Solicitudes de crédito realizadas, proporcionadas por la institución de crédito. Datos completos y válidos de las variables necesarias para el análisis (p. ej. historial crediticio, ingresos, empleo, etc.).
- Solicitudes tanto aprobadas como rechazadas para obtener una visión completa del comportamiento crediticio.

Criterios de exclusión

- Solicitudes con datos incompletos o inconsistentes que no puedan ser corregidos mediante técnicas de preprocesamiento.
- Casos con información irrelevante para el análisis del riesgo crediticio, como solicitudes de crédito con fines no tradicionales o excepcionales.

3.1. Descripción de los Datos

La base de datos a evaluar contiene información sobre los usuarios que solicitaron crédito de una empresa de financiamiento. Esta información es histórica y se encuentra en el periodo de 2022 al 2023.

La base de datos está compuesta de 9,869 observaciones, que incluye 28 variables independientes y una variable objetivo. La siguiente es una descripción de cada variable en la Tabla 3:

Cuadro 3.1: Descripción de las Variables

Nombre	**Política**	**Tipo**	**Descripción**
sucursal		ABC	Nombres de las sucursales
nombre		ABC	Nombre del solicitante
fecha_consulta		#	Fechas de consultas (2022-2023)
version_analisis		#	Versión del análisis
status_generado		ABC	0 = RECHAZADO, 1 = APROBADO
scorecc	✓	#	Puntaje crediticio
acreedores	✓	#	

Continúa en la siguiente página

Cuadro 3.1 – continuación

Nombre	**Política**	**Tipo**	**Descripción**
cuentas_activas	✓	#	Cant. de cuentas activas
emprende_activos		#	
emprende_cerrados		#	
quebrantos	✓	#	
atrasos	✓	#	Cant. de atrasos en pagos
cobranza	✓	#	
creditosFraudulentos	✓	#	Implica rechazo
credIntegranteMora	✓	#	Implica rechazo
credIntegranteSubsidiado	✓	#	Implica rechazo
credLocalizacion	✓	#	Implica rechazo
credCedidoTercero	✓	#	Implica rechazo
credConvenioFiniquito	✓	#	Implica rechazo
posiblesCredCerrados	✓	#	
credOtros		#	Información adicional
cerrados		#	
consultas		#	Cant. de consultas realizadas
domicilios		#	Cant. de domicilios
empleos		#	Cant. de empleos
credAbiertos		#	Cant. de creditos con los que cuenta
otrosCredAnalizar		#	Información adicional
credAbiertosQuebrantosNoCerrados		#	Información adicional
edad		#	Edad en años

3.2. Limpieza y preprocesamiento de Datos

Al examinar la base de datos y sus características, se descubrieron irregularidades y algunos valores que no estaban incluidos en su descripción. Y para la manipulación de datos en la creación de los modelos, se redujeron las variables a tomar en cuenta. Como resultado, se decidió realizar algunas modificaciones, que incluyen los siguientes valores:

- La variable Edad contiene valores debajo de la edad mínima de 18 años; por lo tanto, se procedió a tratar los outliers como la edad mínima requerida como decisión conjunta con la empresa, evitando reducir el número de muestras.
- Las siguientes variables fueron retiradas del dataframe por no tener influencia sobre la variable objetivo: “sucursal“, “nombre”, “fecha_de_consulta”, “versión_análisis”.
- Se comprobó y eliminó registros duplicados.
- La variable objetivo fue codificada como se muestra en su descripción.

3.3. Desarrollo de modelos

3.3.1. EDA

El objetivo del EDA es entender las características de los datos y preparar un conjunto de datos limpio y estructurado que permita construir modelos de machine learning robustos y precisos. Para encontrar la variabilidad, la media y observar cómo se distribuyen las variables, éste análisis se basa principalmente en la estadística descriptiva y los gráficos.

Tras la preparación de los datos, el primer objetivo de este análisis exploratorio es examinar a grandes rasgos las variables para observar su comportamiento y si es posible y necesario purgar aquellas variables que puedan presentar una deficiencia al modelo. En la Figura 3.1 podemos apreciar los estadísticos de las variables.

	count	mean	median	std	min	25%	50%	75%	max	skewness	kurtosis
status_generado	9869	0.527	1	0.499	0	0	1	1	1	-0.106545	-1.988648
scorecc	9869	559.471	572	155.842	0	521	572	650	826	-2.309642	6.341336
acreedores	9869	1.738	1	1.587	0	1	1	3	11	1.080969	1.360776
cuentas_activas	9869	2.811	2	2.593	0	1	2	4	37	1.905251	7.746458
emprende_activos	9869	0.736	1	0.764	0	0	1	1	4	0.702324	-0.218792
emprende_cerrados	9869	8.103	3	10.686	0	0	3	14	59	1.412493	1.299085
quebrantos	9869	0.537	0	0.979	0	0	0	1	9	2.482895	8.434324
atrasos	9869	0.73	0	1.29	0	0	0	1	13	2.703932	10.382332
cobranza	9869	0.206	0	0.504	0	0	0	0	7	3.066422	13.627113
creditosFraudulentos	9869	0.006	0	0.074	0	0	0	0	1	13.283146	174.441968
creditosIntegranteCausanteDeMora	9869	0.014	0	0.122	0	0	0	0	2	8.657448	78.058329
creditosIntegranteSubsidiadoParaEvitarMora	9869	0	0	0	0	0	0	0	0	nan	nan
creditosEnLocalizacion	9869	0.015	0	0.146	0	0	0	0	4	11.657820	168.887693
creditosCedidoAUnTercero	9869	0.003	0	0.053	0	0	0	0	2	21.386327	498.742023
creditosConvenioDeFiniquito	9869	0	0	0	0	0	0	0	0	nan	nan
posiblesCreditosCerrados	9869	0.879	1	1.11	0	0	1	1	11	1.684294	4.155992
creditosOtros	9869	0.01	0	0.102	0	0	0	0	2	9.967200	100.926361
cerrados	9869	27.983	21	33.848	0	7	21	39	845	6.824097	102.554179
consultas	9869	10.328	8	7.957	0	6	8	13	101	2.145906	8.858310
domicilios	9869	2.821	3	0.61	0	3	3	3	3	-3.800544	13.951476
empleos	9869	0.214	0	0.522	0	0	0	0	3	2.740077	8.019337
creditosAbiertos	9869	3.348	3	2.94	0	1	3	5	37	1.702527	5.891311
otrosCreditosParaAnalizar	9869	27.023	20	33.542	0	6	20	38	845	6.962876	105.769195
creditosAbiertosQuebrantosNoCerrados	9869	3.265	3	2.929	0	1	3	5	37	1.773036	6.151852
edad	9869	40.77	40	12.948	18	30	40	50	87	0.383557	-0.488055

Figura 3.1: Resumen estadístico de las variables.

A continuación se presenta un análisis del comportamiento de algunas variables, así como ilustraciones de sus distribuciones, realizando inferencias a partir de los estadísticos mostrados en la Figura 3.1:

1. status_generado

   Con una media de 0.527 y una mediana de 1, indican que más del 50 % de los datos están en el valor máximo de 1 o en los créditos que se denominan “APROBADO”. Se muestra una variabilidad moderada en los datos con una desviación estándar de 0.499. Los valores extremos son menos comunes de lo esperado, según la asimetría negativa (-0.1) y la curtosis negativa (-1.98), que indican que la distribución está ligeramente sesgada hacia la izquierda y es más plana que una distribución normal.

   En la Figura 3.2 se observa como se distribuyen las observaciones, entre los créditos aprobados (1) y los rechazados (0).

   En conclusión, esta información indica que, a pesar de que hay una pequeña mayoría de créditos aprobados, el sistema de aprobación de créditos suele ser equi-

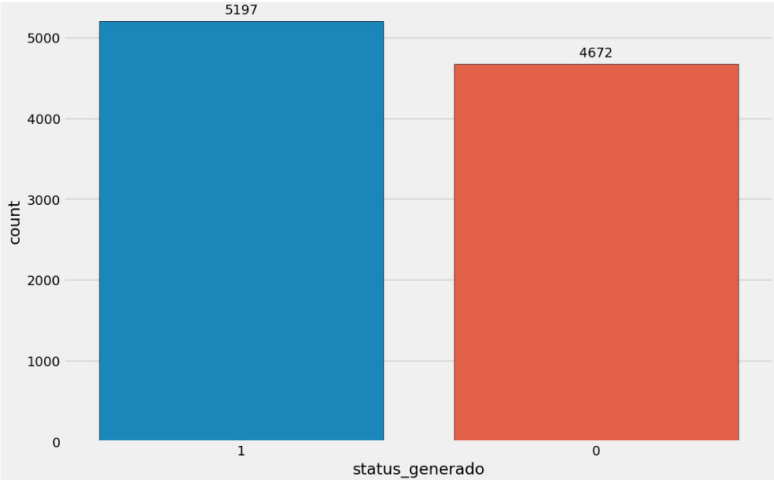


Figura 3.2: Countplot de la variable “status generado”.

librado, con una distribución uniforme entre los créditos aprobados y rechazados. La baja asimetría y la curtosis negativa indican que no hay una gran cantidad de valores extremos, lo que puede implicar que el proceso de aprobación es consistente y no tiende a producir resultados altamente inusuales.

2. scorecc

   Con una media de 559.471 y una desviación estándar de 155.842, la variación entre el valor máximo (826) y el valor mínimo (0) es notable. La curtosis positiva (6.34) muestra una distribución con colas largas que indica la presencia de valores atípicos elevados; por otro lado, la asimetría negativa (-2.3) indica que los valores tienden a acumularse en el extremo inferior de la escala.

   Según la distribución, una gran parte de las personas tienen un puntaje de 0, lo cual podría indicar que son parte de la población que recién comienza a utilizar los sistemas crediticios. Sin embargo, también hay una gran cantidad de personas con puntajes muy altos. La diferencia entre personas con un historial crediticio amplio y aquellas con un historial crediticio nulo se ilustra mejor en la Figura 3.3.

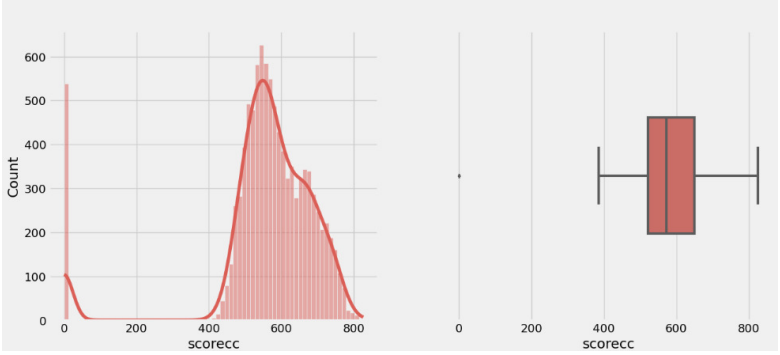


Figura 3.3: Histograma y Boxplot de la variable “scorecc”.

Aunque la mayoría de los puntajes crediticios están concentrados en el rango inferior, hay suficientes valores altos para crear una considerable variabilidad y una distribución sesgada hacia la izquierda con valores extremos elevados. Esto permitirá al modelo ser capaz de diferenciar claramente entre individuos con diferentes perfiles de riesgo, aunque una gran parte tiende a tener un perfil de riesgo mayor.

3. acreedores

   La variable presenta una media de 1.738 y una mediana de 1, lo que indica que la mayoría de los sujetos tiene uno o dos acreedores. La desviación estándar es de 1.587, mostrando una variabilidad moderada. La curtosis positiva (1.36) y la asimetría positiva (1.08) muestran una distribución sesgada hacia la derecha con una concentración de valores alrededor del mínimo.

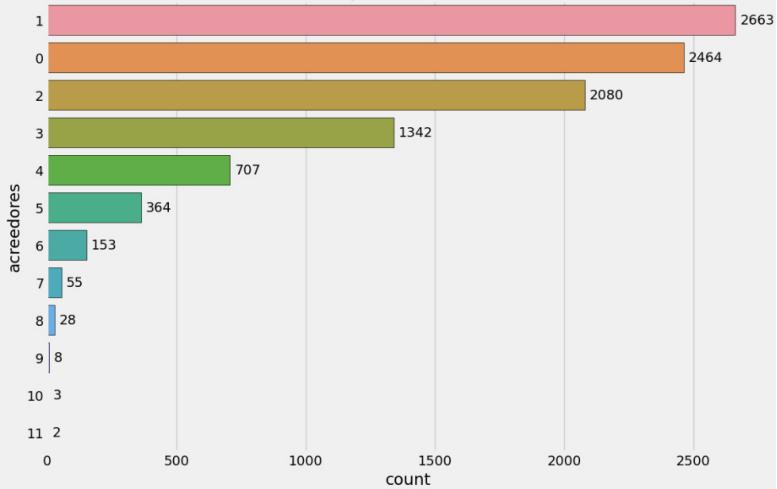


Figura 3.4: Countplot de la variable “Acreedores”.

4. cuentas_activas

   La media de las “cuentas activas” es 2.811, con un valor de la mediana de 2 y un valor máximo de 37. La desviación estándar es de 2.593, lo que sugiere una amplia dispersión de los datos. Según la curtosis positiva (7.46) y la asimetría positiva (1.9), la mayoría de los valores son bajos; sin embargo, algunos valores muy altos tienen un impacto en la distribución. Podemos visualizar su comportamiento en la Figura 3.5.

5. emprende_activos

   La variabilidad es moderada, y la distribución está ligeramente sesgada hacia valores más bajos. La curtosis negativa indica una dispersión relativamente uniforme sin muchos valores extremos (Figura 3.6).

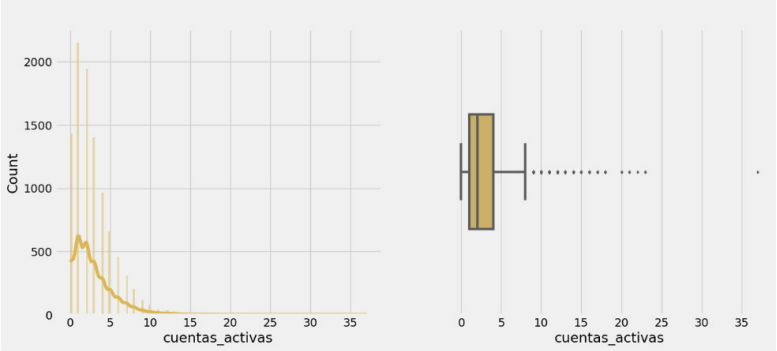


Figura 3.5: Histograma y Boxplot de la variable “cuentas_activas”.

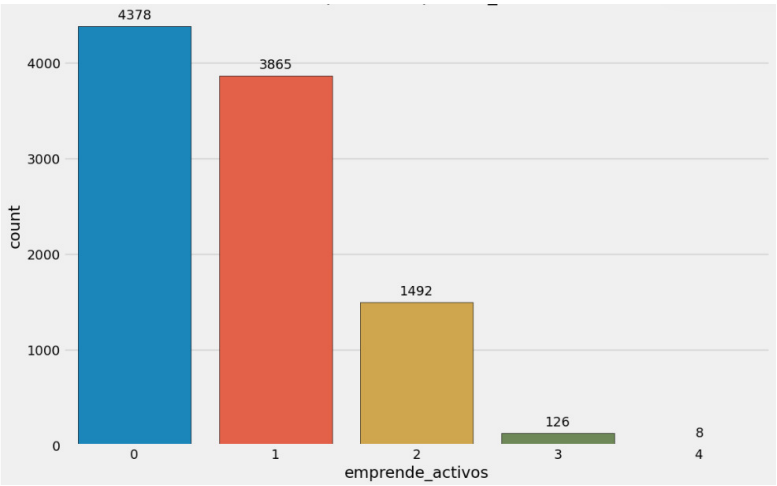


Figura 3.6: Countplot de la variable “emprende_activos”.

6. emprende_cerrados

   La asimetría positiva y la curtosis negativa sugieren que hay pocos valores extremadamente altos, y la mayoría de los datos están dispersos alrededor de la media (Figura 3.7).

7. quebrantos

   La media de “quebrantos” es 0.537, con una mediana de 0, lo que sugiere que más de la mitad de los sujetos no han experimentado quebrantos. La desviación estándar de 0.979 indica poca variabilidad en los datos. La asimetría positiva (2.48) y la curtosis alta (8.43) sugieren que aunque la mayoría de los valores son bajos, hay algunos sujetos con quebrantos muy altos (Figura 3.8).

   La distribución está altamente sesgada hacia la derecha, con una gran cantidad de sujetos sin quebrantos y algunos casos extremos que elevan considerablemente los valores de quebrantos. Esto podría indicar que, aunque las fallas crediticias no son comunes, cuando ocurren, tienden a ser sustanciales y significativas.

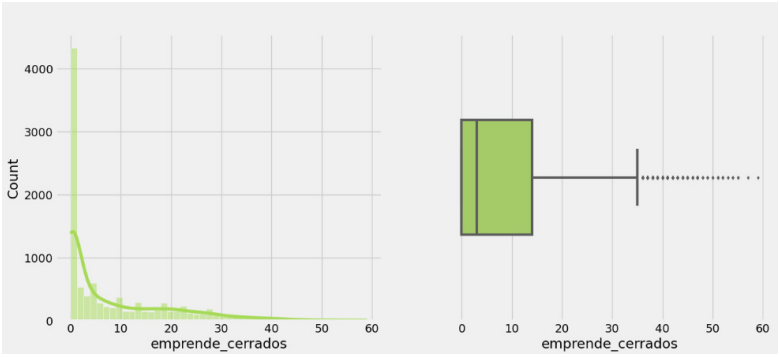


Figura 3.7: Histograma y Boxplot de la variable “emprende_cerrados”.

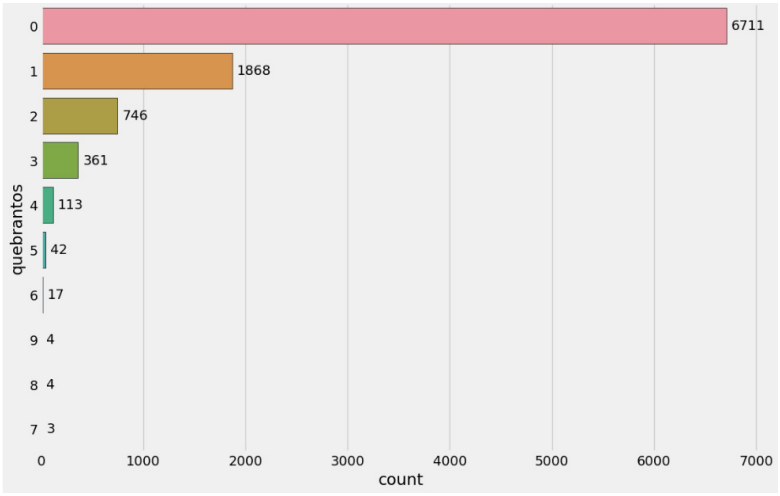


Figura 3.8: Countplot de la variable “quebrantos”.

8. atrasos

   La variable “atrasos” tiene una media de 0.73 y una mediana de 0, indicando que la mayoría de los sujetos no tienen atrasos. La desviación estándar es 1.29, mostrando variabilidad en los datos. La asimetría positiva (2.703932) y la curtosis alta (10.382332) sugieren que la mayoría de los valores son bajos, pero hay algunos sujetos con altos niveles de atrasos.

   Según la distribución (Figura 3.9), la mayoría de las personas tienen un buen historial de pagos sin atrasos; sin embargo, hay una minoría considerable que tiene altos niveles de demora en los pagos. La presencia de valores inusuales en personas con múltiples retrasos se destaca por la alta curtosis y la distribución sesgada hacia la derecha; esto tiene un impacto en la variabilidad general de los datos.

   Esto podría significar que, a pesar de que la mayoría de los sujetos manejan sus pagos de manera adecuada, hay un grupo específico con problemas de retraso graves que requiere una atención particular para la gestión del riesgo.

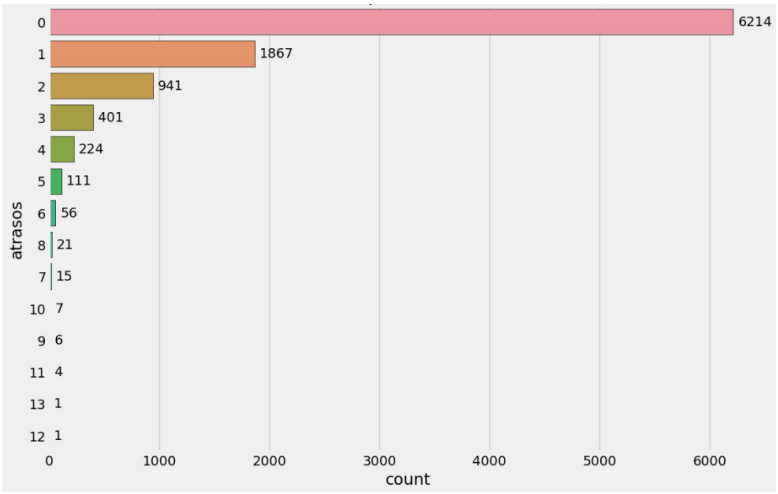


Figura 3.9: Countplot de la variable “atrasos”.

9. cobranza

Para “cobranza”, la media es 0.206 y la mediana es 0, indicando que la mayoría de los sujetos no tienen cobros pendientes. La desviación estándar es 0.504, y la asimetría positiva (3.066422) junto con la curtosis alta (13.627113) sugieren una distribución donde la mayoría de los valores son bajos, con algunos sujetos presentando valores excepcionalmente altos.

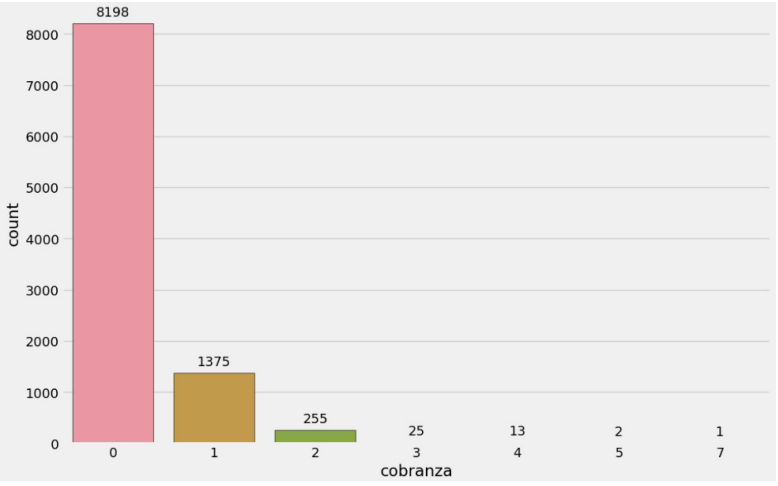


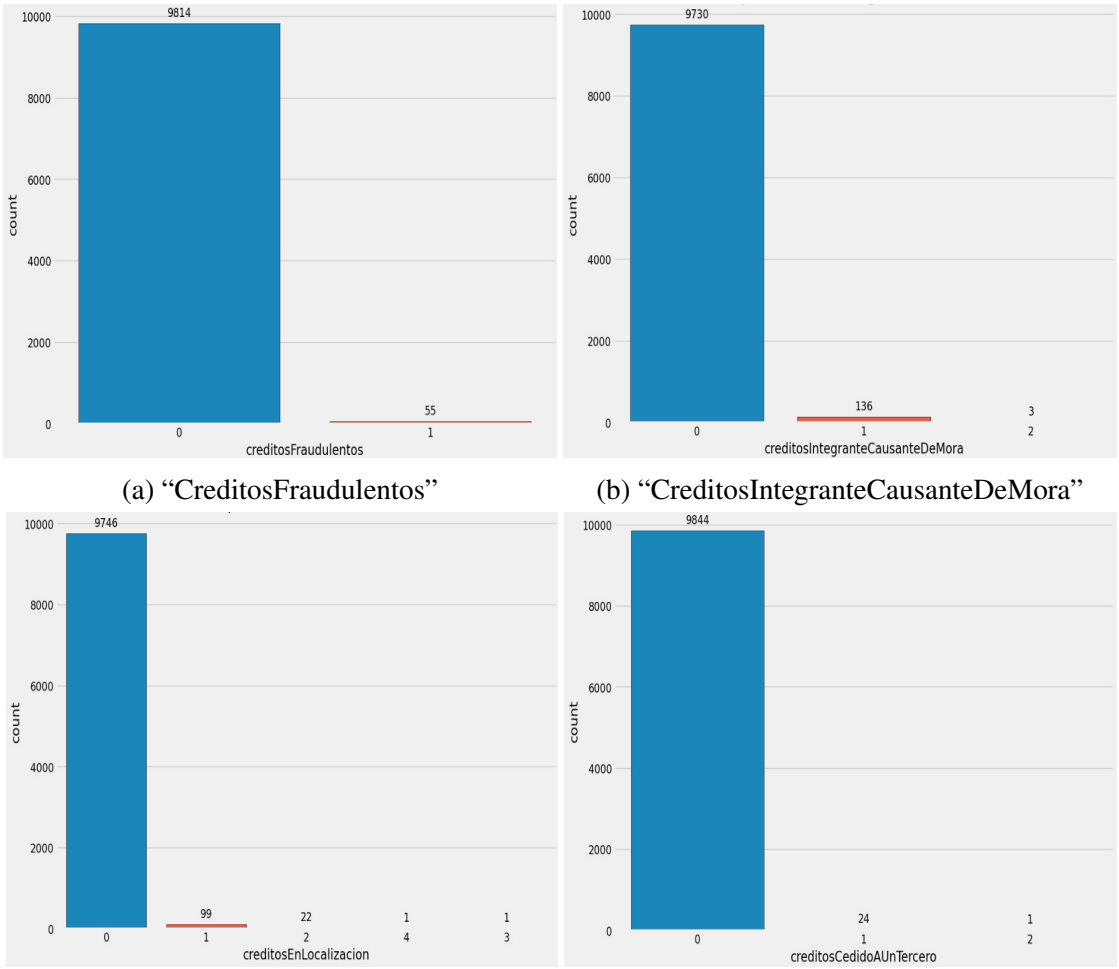
Figura 3.10: Countplot de la variable “cobranza”.

Parece que la mayoría de los sujetos no tienen cobros pendientes, pero hay una pequeña cantidad de sujetos con cobros pendientes excepcionalmente altos.

Esto resulta en una distribución altamente asimétrica y leptocúrtica, indicando la presencia de valores extremos que afectan significativamente la dispersión y forma de la distribución. Este patrón podría sugerir que, a pesar de que el problema de cobros pendientes no es muy común, suele ser bastante severo cuando surge.

10. Variables de Créditos

    Las variables de crédito “CreditosFraudulentos”, “CreditosIntegranteCausante-DeMora”, “CreditosEnLocalizacion” y “CreditosCedidoAUnTercero” muestran una distribución similar, como se ve en la Figura 3.11. Este comportamiento nos muestra que son muy pocas las personas que son rechazadas por tener una de estas variables.



(a) “CreditosFraudulentos” (b) “CreditosIntegranteCausanteDeMora”

(c) “CreditosEnLocalizacion” (d) “CreditosCedidoAUnTercero”

Figura 3.11: Countplots de las variables de Créditos.

11. posiblesCreditosCerrados

    La desviación estándar es 1.11, indicando una variabilidad considerable. La asimetría positiva (1.684294) y la curtosis positiva (4.155992) indican una distribución sesgada hacia valores más bajos con algunos valores altos. Para tener un mayor conocimiento de cómo se distribuye la variable Límite de crédito, a continuación, en la Figura 3.12 se presenta su gráfica de frecuencia.

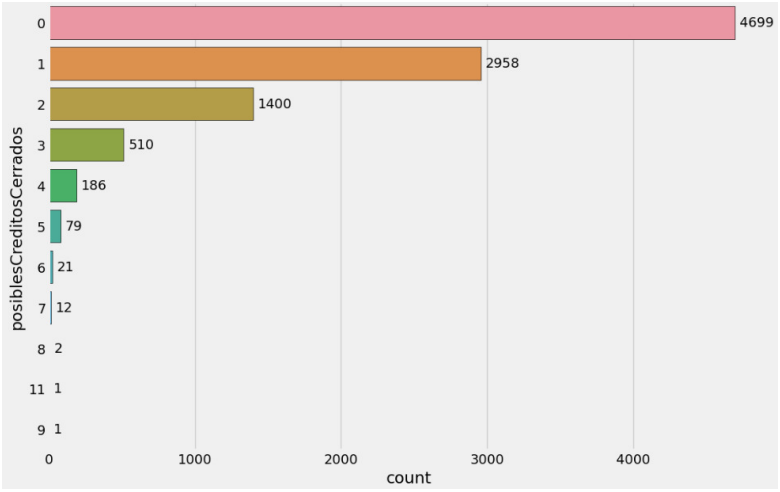


Figura 3.12: Countplot de la variable “PosiblesCreditosCerrados”.

12. creditosOtros

    La media es 0.01 y la mediana es 0, mostrando que otros tipos de créditos son poco comunes. Su gráfica de frecuencia se encuentra en la Figura 3.13 a continuación.

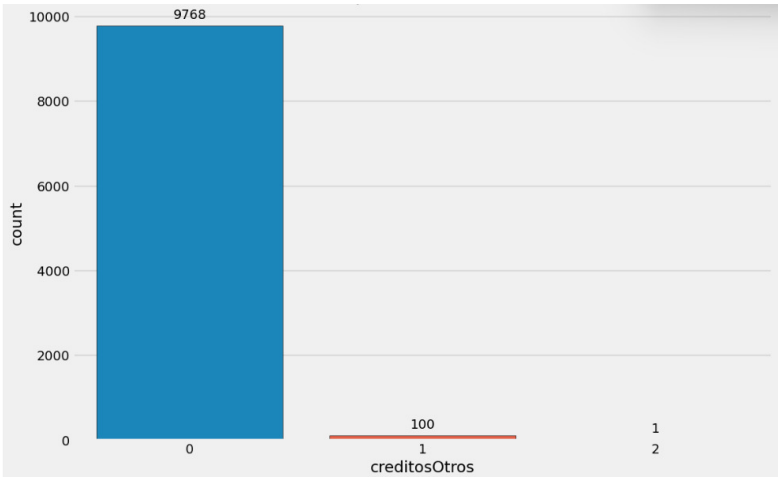


Figura 3.13: Countplot de la variable “CreditosOtros”.

13. cerrados

    La asimetría positiva de 6.824097 muestra que la distribución está sesgada hacia la derecha.

    La muy alta curtosis de 102.554179 sugiere una distribución con picos extremos y colas largas. Esto implica que, además de la mayoría de los valores concentrados alrededor de la mediana, hay un número significativo de valores atípicos elevados.

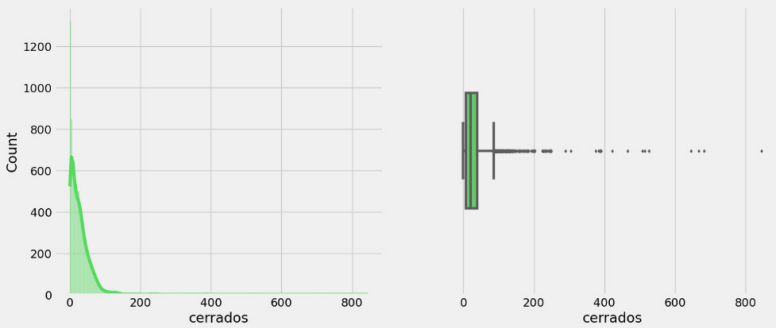


Figura 3.14: Histograma y Boxplot de la variable “cerrados”.

14. consultas

    Para “consultas”, la media es 10.328 y la mediana es 8, indicando que los sujetos tienen un número moderado de consultas en promedio. La desviación estándar es 7.957, mostrando variabilidad en los datos. La asimetría positiva (2.145906) y la curtosis alta (8.858310) sugieren que hay algunos sujetos con un número muy alto de consultas.

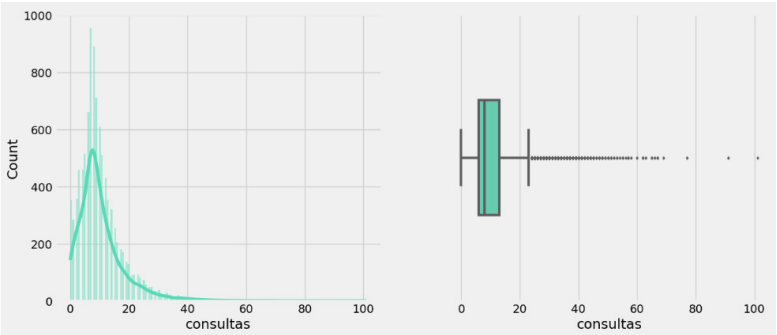


Figura 3.15: Histograma y Boxplot de la variable “Consultas”.

15. domicilios

    La variable “domicilios” tiene una media de 2.821 y una mediana de 3, indicando que la mayoría de los sujetos tienen cerca de tres domicilios.

    La alta curtosis indica que, a pesar de que la mayoría de los datos se concentran cerca de la media, existen suficientes valores extremos para generar colas largas. Aunque hay algunos casos excepcionales que se desvían considerablemente de la norma, esta distribución indica que, en general, el número de domicilios por persona es bastante uniforme.

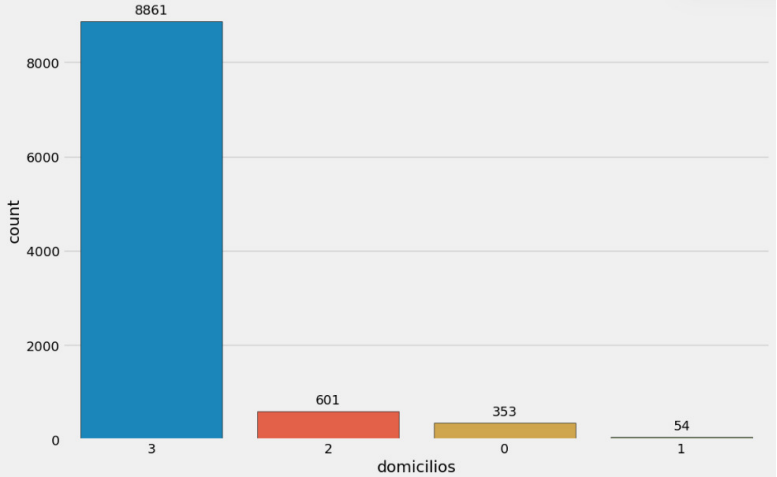


Figura 3.16: Countplot de la variable “Domicilios”.

16. edad

    La variable “edad” tiene una media de 40.77 y una mediana de 40, indicando que los sujetos están en promedio en sus cuarentas. La desviación estándar es de 12.948, mostrando una variabilidad considerable en la edad de los sujetos.

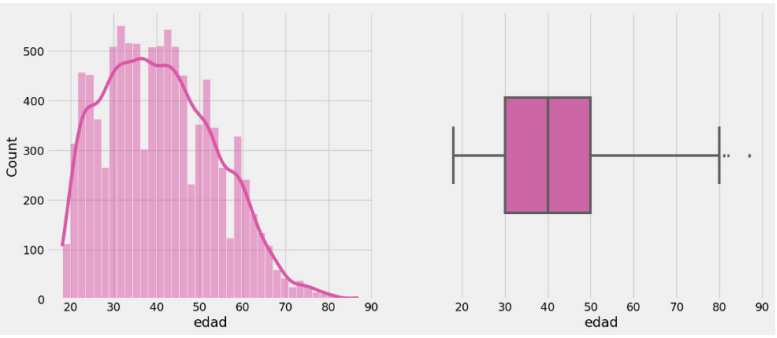


Figura 3.17: Histograma y Boxplot de la variable “edad”.

17. Para las variables empleos, creditosAbiertos, otrosCreditosParaAnalizar, creditosAbiertosQuebrantosNoCerrados, tenemos sus gráficas de frecuencia y Countplots en la Figura 3.18.

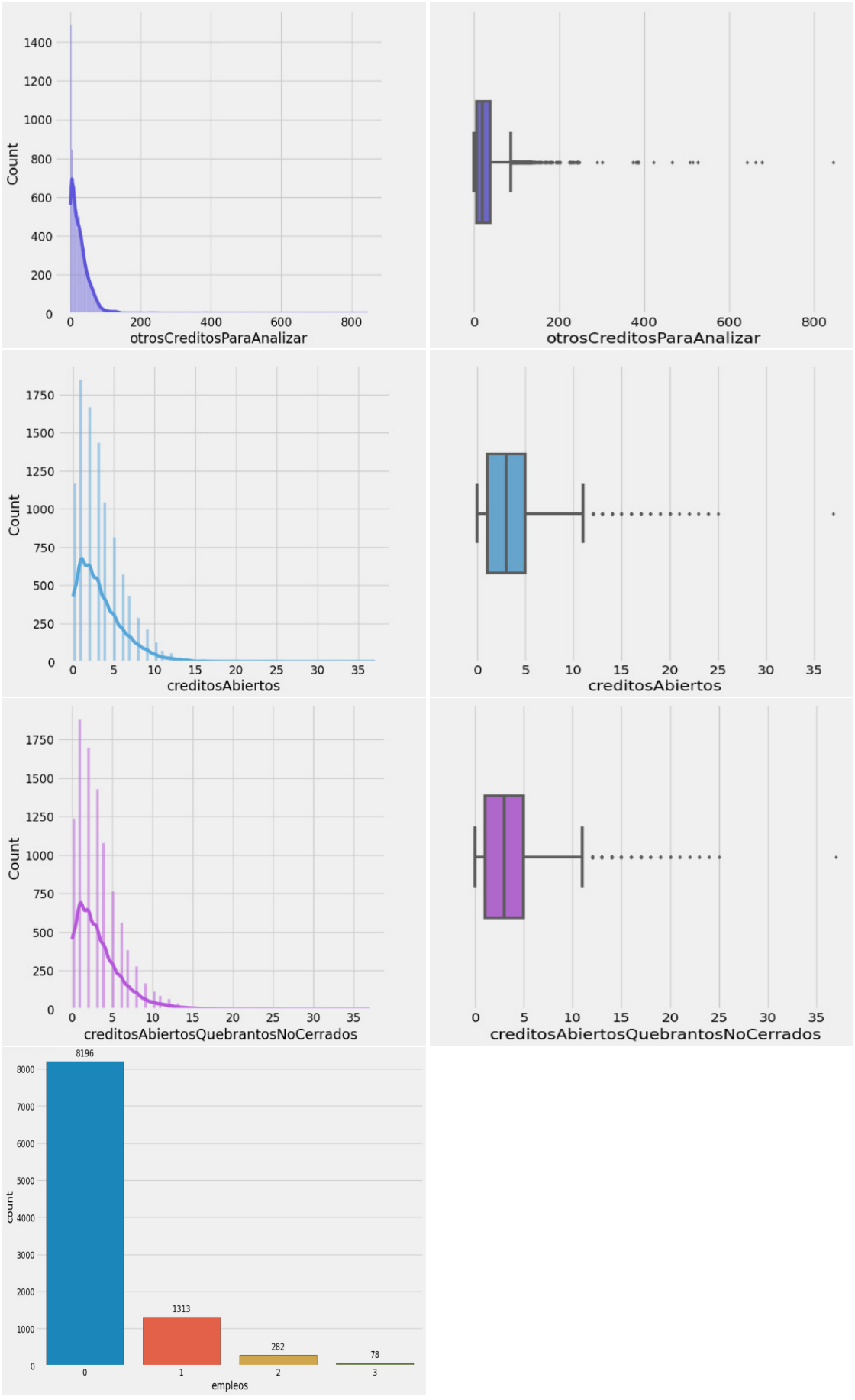


Figura 3.18: Histogramas, Boxplots y Countplots.

Finalmente observamos la correlación entre las variables, en la Figura 3.19. Analizar este mapa de calor de la correlación nos permite identificar variables problemáticas previo al entrenamiento del modelo, como habíamos mencionado con anterioridad.

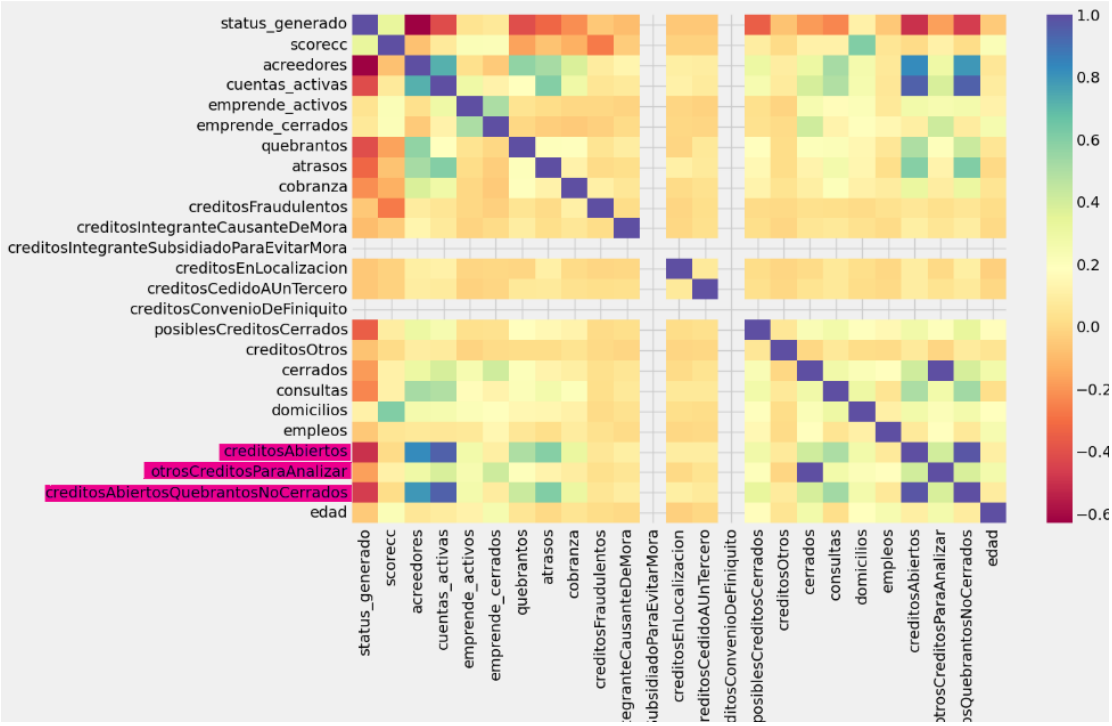


Figura 3.19: Correlación entre variables.

Se ha tomado la decisión de eliminar las variables siguientes con el fin de aumentar la eficiencia del modelo y evitar la multicolinealidad:

- creditosAbiertos.
- otrosCreditosParaAnalizar.
- creditosAbiertosQuebrantosNoCerrados.

En colaboración con el equipo de la institución crediticia, se decidió que las variables seleccionadas fueran eliminadas, ya que no explicaban su información más que sus contrapartes.

3.3.2. Modelo de Regresión Logística

Debido a su amplia gama de habilidades y su versatilidad en el manejo y el análisis de grandes cantidades de datos, Python se ha utilizado para el análisis de datos y la creación de modelos predictivos. Python es un lenguaje de programación popular en la comunidad de ciencia de datos, facilitando el procesamiento y análisis de datos, la visualización y la creación de modelos de aprendizaje automático.

Para este estudio, la biblioteca Scikit-Learn (también conocida como $sklearn$), fue fundamental en este proyecto para la creación y entrenamiento de modelos. Y fue se utilizada en variedad de tareas:

- División de Datos: Con la función train_test_split se realizo la separación de los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para evaluar el rendimiento del modelo.
- Creación y Entrenamiento de Modelos: $sklearn$ incluye implementaciones de numerosos algoritmos de aprendizaje automático, como la regresión logística, árboles de decisión y bosques aleatorios, que fueron utilizados para construir los modelos predictivos.
- Evaluación del Modelo: Herramientas como cross_val_score y métricas como accuracy_score permiten evaluar el rendimiento del modelo.

A partir del algoritmo descrito en la Figura 3.20 se procede a calcular el IV o Information Value mencionado con anterioridad en la sección 2.5.3, con el propósito de identificar variables descartables según su habilidad predictiva.

```
# Función para determinar el valor de informacion de las variables

def IV_calc(data,var):
    if data[var].dtypes == "object":
        dataf = data.groupby(['bin_var'], observed=False)['status_generado'].agg(['count', 'sum'])
        dataf.columns = ["Total","bad"]
        dataf["good"] = dataf["Total"] - dataf["bad"]
        dataf["bad_per"] = dataf["bad"]/dataf["bad"].sum()
        dataf["good_per"] = dataf["good"]/dataf["good"].sum()
        dataf["I_V"] = (dataf["good_per"] - dataf["bad_per"]) * np.log(dataf["good_per"]/dataf["bad_per"])
        return dataf
    else:
        data['bin_var'] = pd.qcut(data[var].rank(method='first'),10)
        dataf = data.groupby(['bin_var'], observed=False)['status_generado'].agg(['count', 'sum'])
        dataf.columns = ["Total","bad"]
        dataf["good"] = dataf["Total"] - dataf["bad"]
        dataf["bad_per"] = dataf["bad"]/dataf["bad"].sum()
        dataf["good_per"] = dataf["good"]/dataf["good"].sum()
        dataf["I_V"] = (dataf["good_per"] - dataf["bad_per"]) * np.log(dataf["good_per"]/dataf["bad_per"])
        return dataf
```

Figura 3.20: Algoritmo “Information Value”.

Si el tipo de datos es objeto, esto indica que es una variable categórica; por lo tanto,

cualquier otra variable, como int64, etc., se tratará como continua y se dividirá en 10 partes iguales (también conocidas como deciles).

Como resultado de los IV's descartamos las siguientes variables:

- creditosIntegranteCausanteDeMora 0.017
- creditosOtros 0.017
- creditosEnLocalizacion 0.015
- creditosCedidoAUnTercero 0.015
- creditosFraudulentos 0.015
- creditosConvenioDeFiniquito..... 0.014
- domicilios 0.014
- creditosIntegranteSubsidiadoParaEvitarMora.... 0.014

Utilizando la función train_test_split de la biblioteca $scikit-learn$,

$$X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 1),$$

donde X representa las variables independientes restantes en el dataframe, y y nuestra variable objetivo.

Se dividieron los datos (utilizando la semilla #1) en un conjunto de entrenamiento y un conjunto de prueba, en una proporción como se aprecia en la Figura 3.21.

	%	Observaciones
Entrenamiento	75%	7,402
Validación	25%	2,467
Total	100%	9,869

Figura 3.21: Segmentación de las consultas.

El conjunto de entrenamiento se utilizó para ajustar el modelo y aprender los patrones y relaciones presentes en los datos. Por otro lado, el conjunto de prueba fue utilizado exclusivamente para evaluar el rendimiento del modelo entrenado, para proporcionar una estimación realista de cómo el modelo generalizará a nuevos datos.

Tal segmentación ayuda a prevenir el sobreajuste y asegurar que la capacidad predictiva del modelo pueda considerarse sólida para aplicarse en futuras consultas.

Para observar los valores p obtenidos para las variables en la regresión, se ajustó el modelo con 14 variables predictoras, las restantes después del proceso de optimización y eliminación de variables poco influyentes o que pudieran provocar problemas al modelo.

Para hacerlo, se utilizó la función de regresión logística de la biblioteca $statsmodels$.

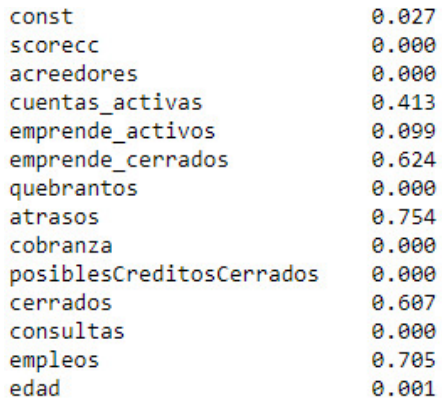


Figura 3.22: Valores p.

Podemos apreciar los valores p resultantes con las variables restantes, estamos tomando valores p mayores a 0.9 para descartar variables, en este caso, no fue necesario retirar variables mediante este filtro.

Además, comprobamos el VIF para identificar variables que hayan sido pasadas por alto como posibles fuentes de multicolinealidad. Para este proyecto tomamos un VIF $>$ 5 para descartar variables, por lo que la única descartada por este filtro fue la variable "scorecc" con un VIF de 22.63.

Una vez que se ha identificado la mejor combinación de variables usando el conjunto de datos de entrenamiento, el objetivo final es predecir la categoría basada en la probabilidad de aprobación. Existen diversas formas de definir el umbral que convierte la probabilidad predicha en una clase concreta. Para encontrar el umbral de probabilidad ideal, hemos utilizado un $grid\ search$.

El $grid\ search$ (Figura 3.23) modifica el umbral de probabilidad por el cual se realiza la clasificación, obteniendo la precisión del modelo para cada valor del umbral, para así encontrar el umbral óptimo.

```
# Determinando el umbral de cambio para el status

both = pd.DataFrame()
both['y_pred'] = model.predict(X_train)
both['probs'] = model.predict_proba(X_train)[:,1]


for i in list(np.arange(0,1,0.1)):
    both["y_pred"] = 0
    both.loc[both["probs"] > i, 'y_pred'] = 1
    print ("Threshold",round(i, 3),"Train Accuracy:",
    round(accuracy_score(y_train, both['y_pred']),4))
```

```
Threshold 0.0 Train Accuracy: 0.527
Threshold 0.1 Train Accuracy: 0.7417
Threshold 0.2 Train Accuracy: 0.8052
Threshold 0.3 Train Accuracy: 0.8393
Threshold 0.4 Train Accuracy: 0.8543
Threshold 0.5 Train Accuracy: 0.8479
Threshold 0.6 Train Accuracy: 0.8316
Threshold 0.7 Train Accuracy: 0.8008
Threshold 0.8 Train Accuracy: 0.7379
Threshold 0.9 Train Accuracy: 0.623
```

Figura 3.23: *Grid search* para el Umbral Óptimo.

La ecuación resultante se muestra en la Figura 3.24; en esta podemos apreciar todas las variables que se utilizan como predictoras para la aprobación del crédito.

```
coeficientes = model.coef_
coeficientes = coeficientes[coeficientes != 0]
intercepto = model.intercept_
variables = X.columns

ecuacion = f"Log(odds) = {intercepto[0]:.3f}"

for i,coef in enumerate(coeficientes):
    ecuacion += f"+ {coef:.3f} * {variables[i]}"

print(ecuacion)
```

```
Log(odds) = 2.230+ -1.553 * acreedores+ 0.037 * cuentas_activas+ 0.322 * emprende_activos+ 0.005 * emprende_cerrados+ -0.531 *
quebrantos+ -0.105 * atrasos+ 0.058 * cobranza+ -0.851 * posiblesCreditosCerrados+ 0.000 * cerrados+ 0.058 * consultas+ 0.023 *
empleos+ 0.009 * edad
```

Figura 3.24: Ecuación Log(odds).

De los coeficientes de nuestros predictores se hacen ciertas inferencias;

- La covariable “acredores” con un coeficiente negativo de 1.553 disminuye la probabilidad de aprobación a medida que esta aumenta.
- La covariable “cerrados” con un coeficiente de cero, para el modelo final, no tiene ninguna influencia sobre las predicciones.
- La covariable “emprende_activos” con un coeficiente de 0.322 es la covariable que más ayuda en aumentar la probabilidad de aprobación.

3.3.3. Modelo de Bosques Aleatorios

Debido a que las pruebas de significancia no son válidas para los modelos de aprendizaje automático, no hemos eliminado cada una de las variables del análisis para este modelo. Para garantizar la robustez del modelo, no obstante, se ha llevado a cabo una validación cruzada de cinco pliegues en los datos de entrenamiento.

En realidad, la multicolinealidad no es un problema para el modelo base sobre el cual se ha construido el ensamble; es un árbol de decisión. En comparación con la regresión logística, el bosque aleatorio requiere mucho menos intervención y esfuerzo humano para entrenar el modelo.

Los datos se han dividido en un 70-30. La razón es que, durante el entrenamiento, se realiza una validación cruzada de cinco pliegues para el $grid\ search$; esto tiene un efecto similar a dividir los datos en un 50-25-25 para los conjuntos de datos de entrenamiento, validación y prueba, respectivamente.

Utilizando el clasificador por bosques aletorios de la biblioteca $sklearn$, con los siguientes valores de hiperparámetros asumidos, nuestro modelo queda definido de esta manera:

$$clf = RandomForestClassifier(n_estimators = 1000, max_depth = 100,$$

$$min_samples_split = 3, min_samples_leaf = 2, random_state = 1)$$

donde;

- n_estimators=1000: El número de árboles en el bosque es 1000. Más árboles pueden mejorar la precisión del modelo, pero también aumentan el tiempo de cómputo.
- max_depth=100: La profundidad máxima que puede alcanzar cada árboles 100.
- min_samples_split=3: El número mínimo de muestras requerido para dividir un nodo es 3. Esto ayuda a prevenir la creación de nodos con muy pocas observaciones, lo cual puede llevar a modelos menos robustos y más susceptibles al ruido.
- min_samples_leaf=2: El número mínimo de muestras que debe contener una hoja es 2. Esto asegura que no se creen nodos demasiado pequeños, lo cual también puede ayudar a prevenir el sobreajuste y mejorar la generalización del modelo.

- random_state=1: La semilla del generador de números aleatorios para asegurar la reproducibilidad de los resultados.

Las siguientes son algunas reglas prácticas también tomadas en cuenta al seleccionar las submuestras de observaciones dentro de nuestro clasificador:

1. Para cada árbol individual, aproximadamente 2/3 de las observaciones se encuentran en los datos de entrenamiento.
2. Para un problema de clasificación, si p es el número total de columnas en los datos de entrenamiento, se utilizan sqrt(p).

Por último, buscamos maximizar la precisión del bosque aleatorio utilizando el $grid\ search$ para encontrar los hiperparámetros ideales, los cuales podrían tener una precisión mucho mayor que los hiperparámetros seleccionados al azar.

Para esto utilizamos la función proporcionada por la biblioteca $sklearn$ "HalvingGridSearchCV". La estrategia de búsqueda comienza evaluando todos los candidatos con una pequeña cantidad de recursos y, de manera iterativa, selecciona los mejores candidatos utilizando cada vez más recursos. Esta es una versión experimental de la función, que tiene como propósito optimizar los recursos para el equipo con el que se utilice un $grid\ search$ [35].

Los siguientes fueron los candidatos:

- clf_n_estimators: (1000, 2000, 3000, 4000, 5000).
- clf_max_depth: (100, 200, 300, 400, 500).
- clf_min_samples_split: (2, 3, 4, 5).
- clf_min_samples_leaf: (1, 2, 3, 4, 5).

La combinación resultante óptima de hiperparámetros fue la siguiente:

$$clf = RandomForestClassifier(n_estimators = 1000, max_depth = 300,$$

$$min_samples_split = 3, min_samples_leaf = 3, random_state = 1)$$

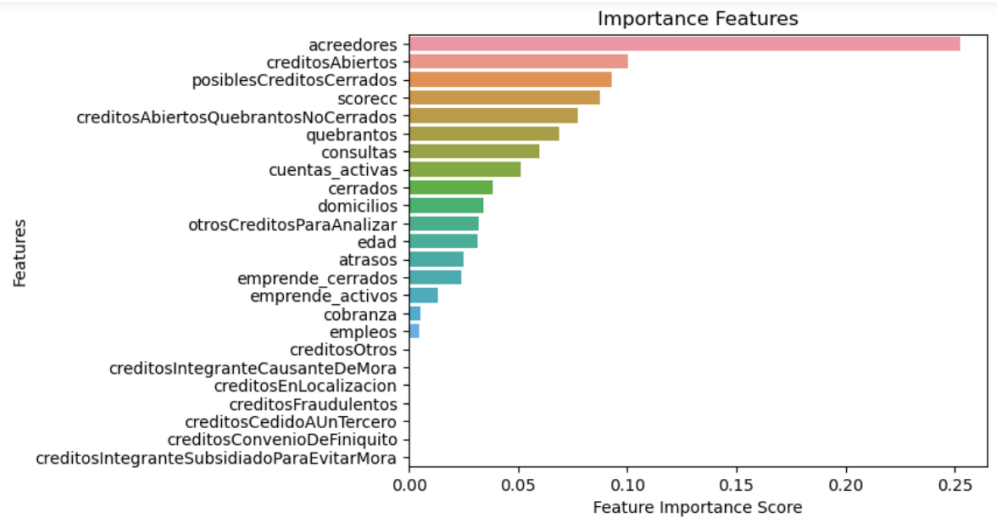


Figura 3.25: Importancia de las Variables.

De la Figura 3.25 observamos cuáles son las variables que mejor ayudan a dividir las muestras en los predictores (árboles) o lo que sería equivalente a conseguir un índice Gini de cero o cercano a cero, lo que demuestra su influencia a la hora de la clasificación.

Finalmente, en la Figura 3.26 se observa una ilustración de uno de los predictores o árboles del bosque, en el se aprecia la magnitud de los mismos.

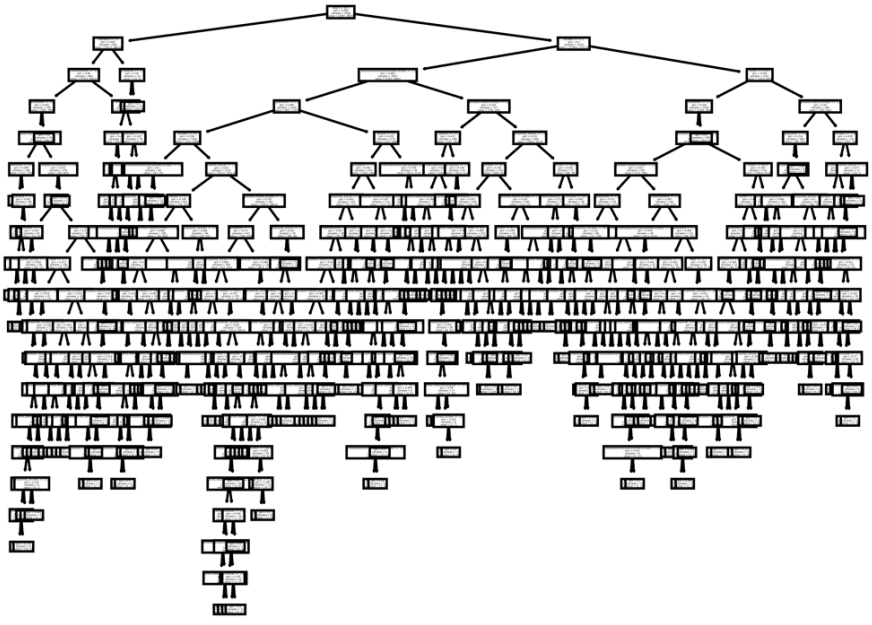


Figura 3.26: Árbol de Decisión del modelo entrenado.

Capítulo 4

RESULTADOS

Las matrices de confusión, los porcentajes de precisión del modelo tanto para el conjunto de entrenamiento como para el de validación, así como la gráfica de la curva ROC (Receiver Operating Characteristic) nos brindan una interpretación eficaz del comportamiento del modelo.

La habilidad del modelo para clasificar adecuadamente las instancias de las clases positivas y negativas se muestra detalladamente en las matrices de confusión. Las matrices de confusión para los conjuntos de validación y entrenamiento se muestran a continuación en la Figura 4.1:

- Matriz de Confusión del Conjunto de Entrenamiento:

  Esta matriz muestra el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos en el conjunto de datos de entrenamiento.

- Matriz de Confusión del Conjunto de Validación:

  Esta matriz representa los resultados de las clasificaciones del modelo del conjunto de datos de validación, permitiendo así, evaluar la capacidad del modelo para generalizar datos nuevos.

En el caso de la regresión logística, la matriz de confusión del conjunto de entrenamiento muestra una alta tasa de precisión en la clasificación, la cual predomina más para el conjunto de validación que para el conjunto de entrenamiento.

		Predecido	
		0	1
Real	0	2686	813
	1	264	3635

(a) Entrenamiento

		Predecido	
		0	1
Real	0	922	246
	1	91	1207

(b) Prueba

Figura 4.1: Matrices de Confusión - Regresión Logística.

Para el caso del modelo de bosques aleatorios, podemos apreciar una cantidad aún menos de resultados erróneos. Estos resultados se deben a la capacidad de los bosques aleatorios para manejar mejor la complejidad de los datos, y al reducir el sobreajuste mediante un $grid\ search$.

		Predecido	
		0	1
Real	0	1277	125
	1	106	1452

Figura 4.2: Matriz de Confusión para la Prueba - Bosques Aleatorios.

Después de haber realizado las optimizaciones necesarias y eliminar las variables conflictivas que podrían haber inflado los resultados, se obtuvieron los siguientes resultados en las pruebas de significancia para la regresión logística:

- El valor p para la razón de verosimilitud de cero nos indica que el modelo es altamente significativo.
- El valor de la pseudo R^2 de 0.4601 sugiere que el modelo tiene un ajuste muy fuerte.

La significancia del modelo también se ve reflejada en los valores p para los coeficientes de los estimadores, cuyos valores en su mayoría son cero o cercanos a cero.

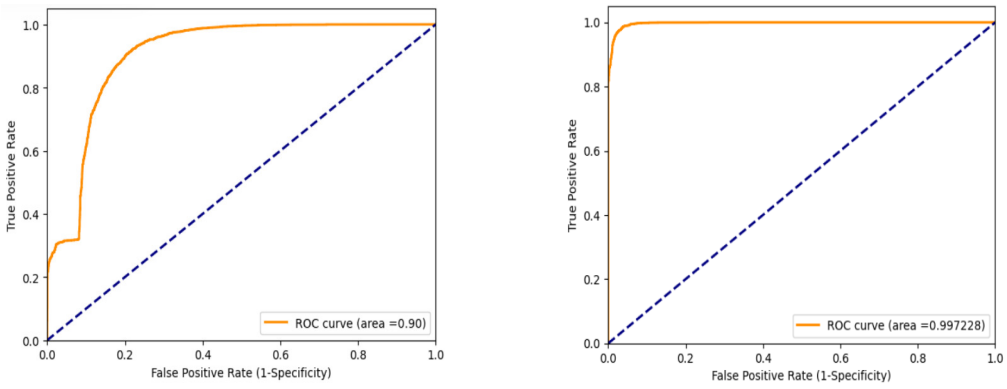
Logit Regression Results

Dep. Variable:	status_generado	No. Observations:	7398
Model:	Logit	Df Residuals:	7385
Method:	MLE	Df Model:	12
Date:	Sun, 22 Sep 2024	Pseudo R-squ.:	0.4601
Time:	13:10:06	Log-Likelihood:	-2762.8
converged:	True	LL-Null:	-5117.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.2331	0.128	17.483	0.000	1.983	2.483
acreedores	-1.5570	0.058	-26.921	0.000	-1.670	-1.444
cuentas_activas	0.0377	0.028	1.331	0.183	-0.018	0.093
emprende_activos	0.3224	0.059	5.444	0.000	0.206	0.439
emprende_cerrados	0.0049	0.004	1.184	0.236	-0.003	0.013
quebrantos	-0.5312	0.061	-8.777	0.000	-0.650	-0.413
atrasos	-0.1055	0.036	-2.929	0.003	-0.176	-0.035
cobranza	0.0588	0.082	0.721	0.471	-0.101	0.219
posiblesCreditosCerrados	-0.8528	0.040	-21.326	0.000	-0.931	-0.774
cerrados	0.0002	0.001	0.182	0.856	-0.002	0.003
consultas	0.0577	0.006	9.795	0.000	0.046	0.069
empleos	0.0241	0.069	0.350	0.726	-0.111	0.159
edad	0.0086	0.003	3.004	0.003	0.003	0.014

Figura 4.3: Resultados de la regresión logística con $statsmodels$.

Las siguientes gráficas ROC presentan el área bajo la curva (AUC), que es una medida del poder discriminatorio del modelo. Estas presentan un AUC cercano a 1 (0.9 para la regresión logística y 0.997 para los bosques aleatorios), indicando un excelente rendimiento de los modelos en la separación de las clases positivas y negativas, pero mostrando una gran diferencia.



(a) Regresión logística (b) Bosques aleatorios

Figura 4.4: Curvas de ROC-AUC.

Además, las cifras resultantes para la precisión, sensibilidad y puntaje f1 de los modelos, y la exactitud en ambos conjuntos de datos muestran que efectivamente los modelos tienen un buen desempeño. Los siguientes son los resultados:

1. Regresión Logística

   - Precisión del modelo: 0.8306951135581555
   - Sensibilidad del modelo 0.9298921417565486
   - F1-Score del modelo 0.8774990912395493
   - Exactitud en el Conjunto de Entrenamiento: El modelo ha alcanzado un porcentaje de precisión del 85 %, lo que indica su capacidad para clasificar correctamente las instancias durante el proceso de entrenamiento.
   - Exactitud en el Conjunto de Validación: En el conjunto de validación, el modelo ha logrado una precisión del 86 %, lo que demuestra su efectividad y robustez al aplicarse a datos nuevos y no vistos.

2. Bosques Aleatorios

   - Precisión del modelo 0.9088069956277327
   - Sensibilidad del modelo 0.9338896020539152
   - F1-Score del modelo 0.9211775878442545
   - Exactitud en el Conjunto de Entrenamiento: El modelo mostró una precisión del 91.35 % en el conjunto de entrenamiento.
   - Exactitud en el Conjunto de Validación: En el conjunto de validación, el modelo ha logrado una precisión del 91.58 %, lo que demuestra su efectividad y robustez al controlar la varianza individual de cada árbol.

     La mejora en la precisión se debe a la combinación de múltiples árboles de decisión que capturan diversas relaciones y patrones en los datos.

Nuevamente en el puntaje f1 podemos apreciar la superioridad del modelo de bosques aleatorios sobre el de regresión logística.

Capítulo 5

DISCUSIÓN

5.1. Interpretación de Resultados

Haciendo una comparación de los resultados sobre la regresión logística y los bosques aleatorios, se puede notar una diferencia significativa en el desempeño. Según la regresión logística, el conjunto de entrenamiento tiene una precisión del 85 % y el conjunto de validación del 86 %. Con esto se puede confirmar la solidez de los hallazgos, mostrando que el modelo es capaz de clasificar las solicitudes de crédito de forma adecuada, además tenemos un 92.12 % de precisión en cuanto al conjunto de entrenamiento y un porcentaje del 92.19 con respecto al conjunto de validación, lo que nos indica por su parte, una mayor capacidad de generalización y una menor tendencia al sobreajuste.

Por otro lado, las matrices de confusión nos demuestran que los bosques aleatorios cometen menos errores en comparación con la regresión logística. ¿La razón? Sencillo, al combinar múltiples árboles de decisión, se capturan patrones más complejos en los datos y existe más fluidez inherente en las solicitudes de crédito.

5.2. Implicaciones Prácticas

Para poder aprobar una solicitud de crédito, debemos considerar las implicaciones prácticas de estos hallazgos, pues son de carácter primordial. Un modelo con mayor precisión y poder discriminatorio reduce el riesgo de conceder créditos a solicitantes no solventes, mejorando así la estabilidad financiera de la institución crediticia. Se considera que los bosques aleatorios son la mejor opción para esta tarea porque tienen una mayor precisión y una tasa de error más baja.

Los bosques aleatorios brindan información valiosa para modificar las políticas crediticias, centrándose en factores que afectan directamente la solvencia de los solicitantes, lo cual permite identificar variables significativas que influyen en la decisión de aprobar o rechazar los créditos.

5.3. Limitaciones del Estudio

Debemos considerar que, incluso teniendo resultados favorables y prometedores, este trabajo nos mostró algunas limitaciones que deben ser consideradas. Primero, en comparación con la regresión logística, la complejidad de los bosques aleatorios puede resultar en un entrenamiento más prolongado y en una mayor necesidad de recursos computacionales.

Además, podemos ver que por la naturaleza de caja negra que tienen los bosques aleatorios, puede generar una dificultad en cuando a las decisiones individuales y la interpretación de sus resultados, incluso si se muestra una alta precisión. Por el contrario, la regresión logística llega a ser considerada como "más adecuada" en situaciones donde la interpretación y la transparencia son esenciales, ya que es un modelo menos complejo entender y más amigable para quienes no sepan mucho del tema.

Otra limitación que se puede captar es, que los resultados de este estudio se basan en un conjunto de datos particular. La generación de estos resultados a otros conjuntos de datos o contextos debe ser realizada con precaución. Es bastante importante validar los modelos en distintos escenarios y con datos adicionales para asegurar su robustez y efectividad.

Es decir, aunque los bosques aleatorios presentan ventajas evidentemente claras en términos de precisión y capacidad, la elección del modelo debe considerar también factores como la interpretabilidad, los recursos computacionales y la aplicación específica en el contexto de la toma de decisiones de crédito.

Capítulo 6

CONCLUSIONES

El presente trabajo encontró el modelo más apto para clasificar las solicitudes de crédito de un banco. Los modelos estudiados fueron la regresión logística y los bosques aleatorios, considerando que la distinción de la clasificación era entre solicitudes aprobadas y rechazadas únicamente. Como se puede apreciar a lo largo del trabajo, obtuvimos resultados aceptables en ambos modelos, además de que tenían varias similitudes entre sí. Esto complicó la elección del mejor clasificador, puesto que además de analizar rigurosamente las métricas del rendimiento obtenidas, fue necesario comprender y considerar la implementación de cada uno.

Al final, la regresión logística fue elegida como el modelo preferido por dos motivos sustanciales. Primero, su sencillez resulta cómoda, ya que solo requiere de conocimientos básicos de estadística para su entendimiento y manejo. Segundo, los resultados son fácilmente comprendidos y pueden presentarse a un público más amplio, ya que las predicciones y las decisiones son fáciles de digerir aún si las personas no tienen una sólida formación en la rama de la estadística. Este modelo es popular en las instituciones financieras no solo por la sencillez en su manejo, sino también porque proporciona resultados y predicciones aceptables y entendibles.

Los modelos de clasificación crediticia permiten diferenciar a los clientes buenos de los malos, sin embargo, estos modelos presentan un gran desafío, ya que su precisión depende completamente de la recolección de los datos de los clientes, puede haber errores y tampoco es posible garantizar al cien por cierto su autenticidad; además, las circunstancias personales y económicas de los mismos pueden evolucionar, no son constantes. Esto puede afectar su comportamiento crediticio para bien o para mal, entonces es necesario y sumamente importante mantener el clasificador en constante monitoreo.
El perfil de un actuario en este trabajo fue oportuno, ya que me fue posible combinar perspectivas financieras y estadísticas, así como transmitir conocimientos técnicos a un

público no especializado en el tema. A su vez, el manejo de los lenguajes de programación modernos resultó ser una alternativa que facilitó el proceso de clasificación de los créditos. Aún así, cabe destacar que la aplicación y comprensión de estos nuevos métodos requirieron un gran esfuerzo y dedicación.

En resumen, tanto los modelos de regresión logística como los de bosques aleatorios resultaron ser útiles para la clasificación de crédito. La elección de uno u otro no depende de otra cosa más que de las métricas de rendimiento, la facilidad de implementación y las políticas de riesgo de cada institución; por lo que un modelo no es mejor que el otro, simplemente uno se ajusta más a las necesidades que cada institución presente. Sin embargo, es importante que éstos se actualicen y adapten constantemente para mantener su efectividad y precisión.

En la Figura 6.1 podemos observar el flujo de trabajo efectuado para el procesamiento y análisis de los modelos de clasificación de solicitudes de crédito que se emplearon en el presente proyecto. Cabe destacar que busca ser de ayuda para otras investigaciones que se desarrollen en el mismo ámbito, que a su vez se basen en los modelos de machine learning, por lo que se destacan los procedimientos clave para garantizar su efectividad.

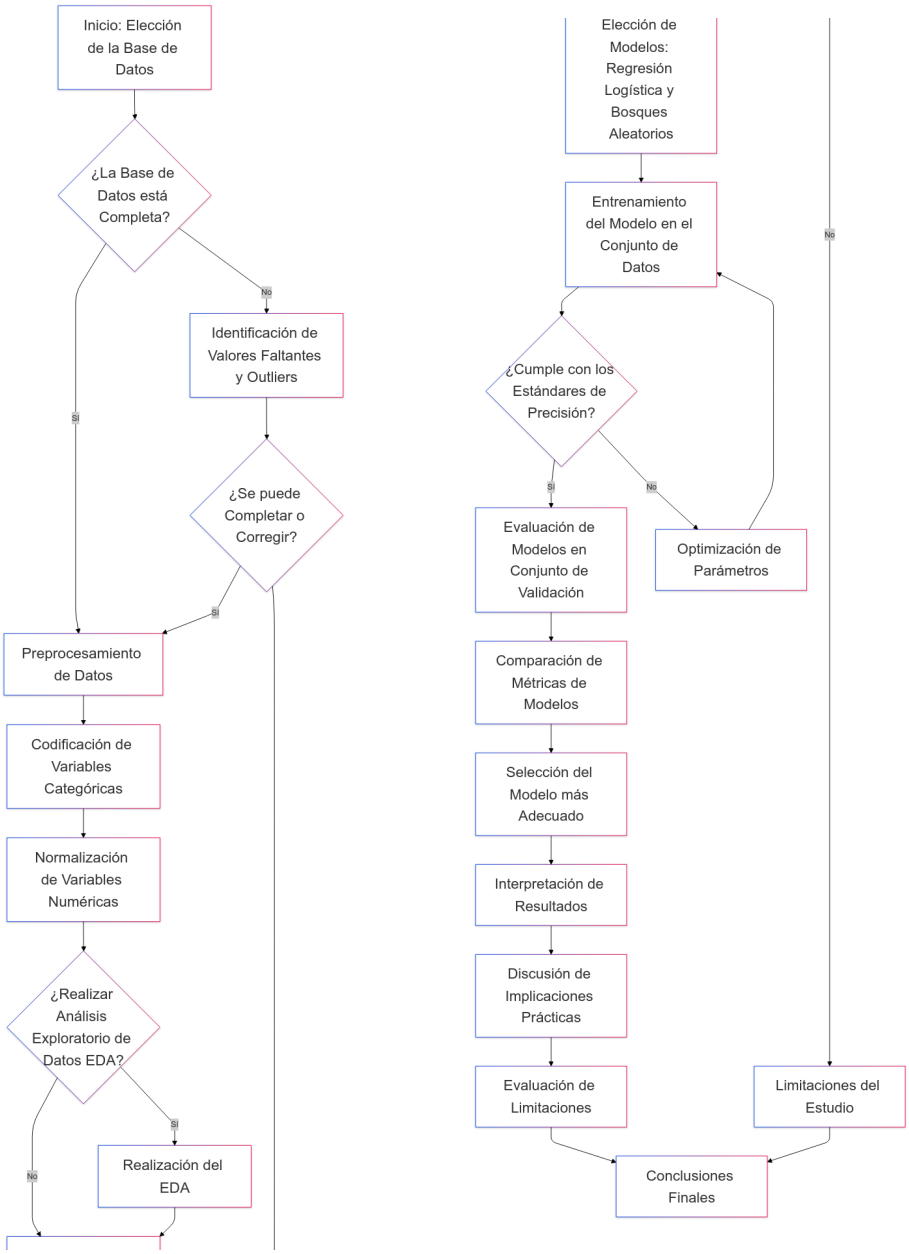


Figura 6.1: Diagrama de Flujo Metodológico.

BIBLIOGRAFIA

[1] V. Ledenyov y D. Ledenyov. *Shaping the international financial system in century of globalization*. arXiv. 2012. URL: https://EconPapers.repec.org/RePEc:arx:papers:1206.2022.

[2] *OPERACIONES FINANCIERAS FUNDAMENTALES*. 2010. URL: https://www.uv.mx/personal/joacosta/files/2010/08/MODULO-II-OPERACIONES-FINANCIERAS-FUNDAMENTALES.pdf.

[3] Diego Cardeña Sosa, Bulmaro Juárez Hernández y Víctor H. Vázquez Guevara. "Modelos de machine learning de clasificación aplicados a un caso de estudio de microcrédito". En: *XVII Semana Internacional de la Estadística y la Probabilidad, FCFM-BUAP*. Puebla, México, 2024.

[4] Scotiabank. *Los 7 tipos de créditos que debes conocer*. n.d. URL: https://www.scotiabankcolpatria.com/educacion-financiera/finanzas-personales/tipos-de-credito.

[5] *Riesgo de crédito o crediticio: qué son y cuáles hay*. 2023. URL: https://www.santanderconsumer.es/blog/post/riesgo-de-credito-o-crediticio-que-son-y-cuales-hay.

[6] Somnath Chatterjee. *Modelos del riesgo de crédito*. 2016. URL: https://www.cemla.org/PDF/boletin/PUB_BOL_LXII-03-03.pdf.

[7] Bonàs Piella A. et al. *Riesgo de crédito: Amenaza u oportunidad*. 2007. URL: https://www.bsm.upf.edu/documents/mmf/06_09_riesgo_credito.pdf.

[8] *Comptroller's Handbook: Rating Credit Risk*. 2001. URL: https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/rating-credit-risk/index-rating-credit-risk.html.

[9] J. E. Baiden. *The 5 C's of Credit in the Lending Industry*. Available at SSRN. 2011. URL: https://ssrn.com/abstract=1872804.

[10] *Principles for the Management of Credit Risk*. 2000. URL: https://www.bis.org/publ/bcbs75.htm.

[11] D. Utami y H. R. Kawulur. "Analysis of Bankruptcy Prediction by Altman Z-Score Modification (Study at PT BPR Primaesa Sejahtera Kota Manado)". En: *International Journal of Applied Business and International Management* 31 (ago. de 2020), págs. 8-15. URL: https://doi.org/10.32535/ijabim.v0i0.871.

[12] W. Mu y M. Liu. "Research on Financial Early Warning of Private Listed Companies Based on Z-score Model". En: *Proceedings of the 4th International Conference on Economy, Judicature, Administration and Humanitarian Projects (JAHP 2019)*. 2019, págs. 360-363. URL: https://doi.org/10.2991/jahp-19.2019.76.

[13] Wang Tao y Li Mengqi. "Analysis of the Business of the Regional Enterprises in the Belt and Road". En: *International Journal of Economics and Financial Research* 5.2 (feb. de 2019), págs. 43-48. DOI: 10.32861/ijefr.52.43.48.

[14] M. P. Pantoja-Aguilar et al. "Testing Altman's Z"-Score to assess the level of accuracy of the model in Mexican companies". En: *Nova Scientia* 13.27 (2021), pág. 00019. URL: https://doi.org/10.21640/ns.v13i27.2881.

[15] IBM. *¿Qué es machine learning (ML)?* Recuperado de IBM. IBM. 2024. URL: https://www.ibm.com/mx-es/topics/machine-learning.

[16] Irving Ramirez. *Riesgo de crédito, modelos de inteligencia artificial IA*. Figura 2. Aproximaciones al riesgo de crédito. Miralta Bank. 2024. URL: https://www.miraltabank.com/riesgo-de-credito-modelos-de-inteligencia-artificial/.

[17] D. J. Hand y W. E. Henley. "Statistical classification methods in consumer credit scoring: a review". En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160.3 (1997), págs. 523-541.

[18] V. Dheepa y R. Dhanapal. "Behavior based credit card fraud detection using support vector machines". En: *ICTACT Journal on Soft Computing* 2.4 (2012), págs. 391-397.

[19] Luis Camarero Rioja, Antonio Almazán Llorente y Beatriz Mañas Ramírez. *Regresión Logística: Fundamentos y aplicación a la investigación sociológica*. 2013. URL: https://www2.uned.es/socioestadistica/Multivariante/Odd_Ratio_LogitV2.pdf.

[20] David G. Kleinbaum y Mitchel Klein. *Logistic Regression: A Self-Learning Text*. 2nd. Springer, 2002.

[21] Pratap Dangeti. *Statistics for Machine Learning*. Packt, 2017.

[22] M. E. Ferre Jaén. *FEIR 45: Regresión logística*. Apuntes del curso FEIR3, curso 2014/15 actualizados. 2019. URL: https://gauss.inf.um.es/feir/45/.

[23] Gareth James et al. *An Introduction to Statistical Learning*. Springer, 2013.

[24] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[25] C. Y. J. Peng, K. L. Lee y G. M. Ingersoll. "An Introduction to Logistic Regression Analysis and Reporting". En: *The Journal of Educational Research* 96.1 (oct. de 2002), págs. 3-14. URL: https://doi.org/10.1080/00220670209598786.

[26] V. Plevris et al. "Investigation of performance metrics in regression analysis and machine learning-based prediction models". En: *ECCOMAS 2022*. 2022. URL: https://www.scipedia.com/public/Plevris_et_al_2022a.

[27] Vimukthini Pinto y Roshini Sooriyarachchi. "Comparison of methods of estimation for a goodness of fit test – an analytical and simulation study". En: *Journal of Statistical Computation and Simulation* 91 (ene. de 2021), págs. 1-21. DOI: 10.1080/00949655.2021.1872078.

[28] WK Wong. "Segmented R2: A Proposed Variant for Unbalanced and Skewed Dataset". En: ene. de 2024, págs. 202-205. DOI: 10.1109/GECOST60902.2024.10474788.

[29] X. Zhang y D. Hedeker. "Defining R-squared measures for mixed-effects location scale models". En: *Statistics in Medicine* 41.22 (2022), págs. 4467-4483. DOI: 10.1002/sim.9521.

[30] Barry H. Nathanson y Thomas L. Higgins. "An Introduction to Statistical Methods Used in Binary Outcome Modeling". En: *American Journal of Medical Quality* 23.3 (2008), págs. 206-215. URL: https://doi.org/10.1177/1089253208323415.

[31] V. Bewick, L. Cheek y J. Ball. "Statistics review 14: Logistic regression". En: *Critical Care* 9.112 (2005), págs. 1-8. URL: https://doi.org/10.1186/cc3045.

[32] Ke Xu. *How Has the Literature on Gini's Index Evolved in the Past 80 Years?* Dalhousie University, Economics Working Paper. Abr. de 2003. URL: https://ssrn.com/abstract=423200%20or%20http://dx.doi.org/10.2139/ssrn.423200.

[33] Frauke Degenhardt, Stephan Seifert y Silke Szymczak. "Evaluation of variable selection methods for random forests and omics data sets". En: *Briefings in Bioinformatics* 20.2 (mar. de 2019), págs. 492-503. DOI: 10.1093/bib/bbx124.

[34] A. Irawan y E. B. Santoso. "Analytical tools beyond Gini index to study inequality: a case of City of Blitar, East Java, Indonesia". En: *IOP Conference Series: Earth and Environmental Science* 340 (2019), pág. 012004. URL: https://doi.org/10.1088/1755-1315/340/1/012004.

[35] scikit-learn developers. *sklearn.model_selection.HalvingGridSearchCV*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html. BSD License. 2007–2024.

Apéndice A

Repositorio de Algoritmos

En el desarrollo de esta tesis, se implementaron diversos algoritmos para la evaluación del riesgo crediticio, incluyendo modelos de regresión logística y de bosques aleatorios. Estos algoritmos han sido fundamentales para el análisis de los datos y la obtención de los resultados presentados.

Para facilitar la revisión y replicabilidad de estos modelos, se ha creado un repositorio en línea con el código utilizado, el cual está disponible en el siguiente enlace:

Repositorio de Algoritmos - Credit Classification Models

En el repositorio, encontrarás:

- Código de los modelos: Implementación de los modelos de regresión logística y bosques aleatorios.
- Scripts de procesamiento: Código empleado para la preparación y limpieza de los datos.