Tipo de documento: Tesis de maestría

Escuela de Negocios. Master in Management + Analytics

Predicción del riesgo crediticio en productos de factoring financiero utilizando modelos de aprendizaje automático

Autoría: Castaño, Franco Tomás
Año: 2024

UNIVERSIDAD
TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICCIÓN DEL RIESGO CREDITICIO EN PRODUCTOS DE FACTORING FINANCIERO UTILIZANDO MODELOS DE APRENDIZAJE AUTOMÁTICO

TESIS

Franco Tomás Castaño
Mayo 2024

Tutor: Andrés Gago

Resumen

La asignación eficiente de recursos financieros es un desafío global que impacta significativamente en todas las empresas. La incapacidad de las entidades crediticias para segmentar adecuadamente a sus clientes puede resultar en consecuencias adversas, como ofrecer condiciones de crédito menos favorables a los buenos pagadores. Este fenómeno se manifiesta en tasas de interés más altas o límites de crédito más bajos, con el objetivo de maximizar la rentabilidad de determinados segmentos de clientes.

En este contexto, los datos relacionados con la facturación y las cobranzas de las empresas están emergiendo como recursos valiosos, cada vez más accesibles y aplicables para los proveedores financieros. La presente tesis se propone poner a prueba la hipótesis de que estos datos pueden emplearse eficazmente para mejorar la segmentación de clientes por parte de las instituciones financieras.

Para alcanzar este objetivo, se emplean técnicas estadísticas avanzadas y de aprendizaje automático, incluyendo la ingeniería de atributos y análisis descriptivos univariados y multivariados. El problema de predicción se enmarca dentro de la naturaleza de la clasificación, con el objetivo de desarrollar un modelo predictivo robusto.

Los resultados obtenidos de esta investigación revelan la construcción de un modelo cuya performance, medida mediante el Área Bajo la Curva (AUC), alcanza un 90.99%. Este logro ha facilitado una clasificación más eficiente de las líneas de crédito, lo que tiene importantes implicaciones para la optimización de la asignación de recursos financieros y la mejora de la rentabilidad en el sector financiero.

Abstract

Efficient allocation of financial resources is a global challenge that significantly impacts all businesses. The inability of credit entities to properly segment their customers can result in adverse consequences, such as offering less favorable credit terms to good payers. This phenomenon manifests in higher interest rates or lower credit limits, aimed at maximizing the profitability of certain customer segments.

In this context, data related to billing and collections from companies are emerging as valuable resources, increasingly accessible and applicable for financial providers. This thesis aims to test the hypothesis that such data can be effectively used to enhance customer segmentation by financial institutions.

To achieve this goal, advanced statistical and machine learning techniques are employed, including feature engineering and univariate and multivariate descriptive analysis. The prediction problem is framed within the nature of classification, with the aim of developing a robust predictive model.

The results obtained from this research reveal the construction of a model whose performance, measured by the Area Under the Curve (AUC), reaches an 90.99%. This achievement has facilitated a more efficient classification of credit lines, which has significant implications for optimizing the allocation of financial resources and improving profitability in the financial sector.

Índice

1. Introducción

1.1. Contexto

El éxito de la industria crediticia se fundamenta en una gestión efectiva del riesgo crediticio. Al otorgar un crédito, una entidad asume el riesgo de incobrabilidad, es decir, la posibilidad de que el prestatario no cumpla con la obligación de reembolso. Las entidades financieras administran la asignación de recursos financieros basándose en los perfiles de riesgo de sus clientes. En otras palabras, al identificar clientes con menor riesgo crediticio, estas instituciones pueden ofrecer ofertas más competitivas, como la reducción de tasas de interés, para fomentar una mayor adopción.

En los últimos años, la industria financiera ha experimentado una evolución disruptiva. Emergieron bancos digitales que operan exclusivamente en el ámbito digital sin sucursales físicas, mientras que los bancos tradicionales, aunque de manera más gradual, también han adoptado prácticas y tecnologías orientadas al entorno digital. Adicionalmente, han surgido las *fintech*, empresas emergentes con un fuerte componente tecnológico cuyo propósito es ofrecer servicios financieros. Todas estas entidades desempeñan un papel crucial en el proceso de otorgamiento de créditos y se enfrentan al dilema de tomar decisiones en la gestión del riesgo crediticio: conceder el crédito si existe una probabilidad significativa de reembolso o abstenerse si el prestatario presenta altas probabilidades de incumplimiento.

1.2. Problema

La necesidad de una empresa *fintech*, especializada en otorgamiento de créditos para mejorar la clasificación de sus clientes, es evidente para optimizar el proceso de originaciones. En la actualidad, la falta de una clasificación adecuada de los clientes y la escasa validación de las políticas de riesgo obstaculizan la capacidad de la empresa para determinar con precisión el nivel de riesgo asociado a sus clientes.

Es importante que la empresa fortalezca su marco de clasificación de riesgo, implementando criterios más refinados y validando de manera rigurosa sus políticas. Una clasificación más precisa permitirá una evaluación más fundamentada de los perfiles de riesgo individuales, lo que facilitará la toma de decisiones informadas en el proceso de otorgamiento de créditos.

Esta mejora en la clasificación de riesgos también puede impulsar la eficiencia operativa y la rentabilidad al permitir una asignación más precisa de recursos y términos crediticios. Además, contribuirá a la construcción de un portafolio más sólido y equilibrado, mitigando posibles pérdidas y optimizando la gestión general de riesgos.

1.3. Objetivo

La exploración de la información interna de la compañía, incluyendo el historial de crédito de sus clientes, junto con datos obtenidos de terceros como agencias de crédito y entidades públicas, plantea la posibilidad de diseñar e implementar un modelo de aprendizaje automático con el objetivo de clasificar a los prestatarios según su nivel de riesgo. La respuesta a esta interrogante puede tener un impacto significativo en la gestión de riesgos y las políticas de crédito de la empresa.

La implementación de un modelo de aprendizaje automático aprovechando estos datos puede aportar numerosos beneficios. Primero, permitiría una evaluación más precisa y eficiente de los riesgos asociados a cada prestatario, al integrar una gama más amplia de información en tiempo real. Este enfoque podría aumentar la capacidad de la empresa para identificar patrones y señales de riesgo tempranas, mejorando así la toma de decisiones.

Además, un modelo de aprendizaje automático bien diseñado podría ofrecer una mayor capacidad de adaptación a los cambios en el comportamiento financiero de los prestatarios y a las tendencias del mercado. Esto proporcionaría una ventaja estratégica al ajustar dinámicamente las políticas de crédito según las condiciones cambiantes, optimizando la gestión de riesgos.

No obstante, es crucial abordar ciertos desafíos potenciales, como la interpretabilidad del modelo y la necesidad de asegurar la privacidad y confidencialidad de los datos. La empresa deberá implementar prácticas sólidas de gobernanza de datos y cumplir con las regulaciones pertinentes para garantizar la integridad y ética en el uso de la información.

En resumen, la implementación de un modelo de aprendizaje automático puede ser viable y, si se ejecuta adecuadamente, podría tener un impacto significativo en la gestión de riesgos y las políticas de crédito, brindando a la empresa una ventaja competitiva en el sector *fintech*.

1.4. Funcionamiento del factoring internacional

El factoring se configura como un modelo de negocio financiero que implica la cesión de cuentas por cobrar. Este proceso se fundamenta en la existencia de una relación comercial entre el comprador y el vendedor, o entre un proveedor de servicios y su cliente. En esta dinámica, el comprador o cliente adquiere un bien o servicio y se compromete a efectuar el pago en un plazo diferido.

En este contexto, la empresa de factoring juega un papel crucial al adquirir el derecho de cobro que la entidad vendedora o prestadora de servicios posee. Aunque la obligación de pago persiste en el comprador o cliente, se introduce una variación sustancial: el pago ahora debe ser efectuado en una cuenta corriente designada específicamente para este propósito, perteneciente al proveedor financiero.

Esta modalidad financiera, al facilitar la liberación anticipada de los fondos asociados a las cuentas por cobrar, no solo optimiza la liquidez de la empresa vendedora, sino que también transfiere la responsabilidad de la gestión del cobro al factoraje, permitiendo así una mayor concentración de recursos y esfuerzos por parte del vendedor en sus operaciones comerciales centrales.

La relación commercial más fuerte se da entre la empresa exportadora y la empresa *fintech*. La empresa exportadora si bien cede el crédito no cede el riesgo por en caso de *default* por parte de la empresa importadora el exportador debe hacer frente a dicha obligación. La empresa opera principalmente con exportadores basados méxico, y si bien el importador puede estar basados en cualquier parte del mundo, principalmente operan con importadores en Europa y Estados Unidos.

2. Fuente de los datos

La empresa *fintech* se dedica al otorgamiento de créditos a través de la compra de facturas de exportación. Una operación de comercio internacional suele tener un plazo de entrega de entre 30 y 180 días, dependiendo de la complejidad de la logística. Durante dicho período, las empresas importadoras no pueden afrontar el pago de la compra debido a que no recibieron el producto dentro de su flujo operacional, y la empresa exportadora, al no recibir el cobro en este período, genera un efecto negativo en su flujo operativo.

Dicha operación genera la necesidad de liquidez en el exportador, y este agente contacta a la empresa *fintech* para ceder su derecho de cobro. Es importante mencionar que la responsabilidad de pago recae sobre la empresa importadora; sin embargo, la empresa exportadora es garante de dicha operación, por lo cual, si el importador no afronta el pago, la empresa *fintech* puede ejecutar al exportador.

La empresa cuenta con datos internos como el historial crediticio de sus propios clientes, fecha de originación de cada crédito, monto del mismo, y plazo de originación que corresponde a cuántos días faltan para el vencimiento de las facturas desde el momento en que el exportador decide solicitar su adelanto. Cabe destacar que algunas empresas tienen habilitada la opción de pedir el adelanto sobre facturas vencidas.

Los importadores con los que trabaja la empresa están basados en diferentes países del mundo. Sin embargo, los exportadores están basados en México. Cuando los exportadores comienzan a operar a través de la *fintech*, deben otorgar el permiso de

acceso a la base de datos pertinente del Servicio de Administración Tributaria (SAT), que básicamente consta principalmente de un listado de facturas emitidas por dicha empresa, además de datos identificatorios de la empresa y el tipo de industria de la misma.

La base de datos externa que proviene del SAT, que consta de las facturas emitidas por la empresa exportadora, será modelada para obtener variables del tipo de volumen de ventas, frecuencia y variabilidad tanto para las ventas domésticas como exportaciones, y ventas totales, que representarán la suma de ventas domésticas y exportaciones.

La pregunta que buscamos resolver en la presente tesis es si existe una relación, o modelo, que ayude a entender si los datos que la empresa obtuvo a través del SAT tienen relación alguna con los datos del historial crediticio de la empresa, y si dicha relación o modelo es útil para generar una segmentación de clientes por tipo de riesgo.

Una observación en esta base de datos corresponde a una versión estática del perfil de ventas de una empresa exportadora al momento en que la misma cedió una facture a la empresa *fintech*, con información de si ese crédito fue repagado o no.

La base de datos seleccionada para este estudio comprende los créditos otorgados por la empresa *fintech* durante el periodo de enero a septiembre del año 2022. No obstante, con el propósito de llevar a cabo un análisis descriptivo, se limitará la consideración a la información correspondiente a los créditos originados hasta el mes de octubre. Los créditos originados en noviembre y diciembre se reservarán para la validación del modelo, siendo comúnmente denominado como el conjunto de datos "out of time". La base de créditos hasta julio consta de un total de 10,946 créditos originados, de los cuales aproximadamente 99 han registrado impago.

3. Estado del Arte y Revisión de Literatura :

El armado de modelos estadisticos para la predicción del riesgo crediticio en el ámbito del factoring financiero ha sido objeto de varios estudios, cada uno utilizando diversas técnicas y enfoques metodológicos. A continuación, se presentan algunos de los enfoques de métodos destacados en la literatura:

Modelos Tradicionales : Los modelos tradicionales de evaluación de riesgos han sido fundamentales en el ámbito financiero durante décadas. Autores como Smith et al. (2015) han investigado y aplicado técnicas como la regresión logística y los árboles de decisión para predecir el riesgo crediticio en el contexto del factoring financiero. Estos modelos se caracterizan por su relativa simplicidad y alta interpretabilidad, lo que los hace ampliamente aceptados en entornos donde la transparencia y la comprensión del proceso de toma de decisiones son cruciales.

Por ejemplo, Smith et al. (2015) demostraron cómo la regresión logística puede ser efectiva para modelar relaciones lineales entre variables predictoras y el riesgo crediticio, permitiendo a las instituciones financieras evaluar la solvencia de sus clientes de manera confiable.

Aprendizaje Automático Avanzado: En contraste, estudios más recientes han adoptado métodos avanzados de aprendizaje automático para mejorar la precisión predictiva y manejar la complejidad inherente en los datos financieros. Investigaciones como las de Chen et al. (2020) y Lee (2021) han destacado la eficacia de modelos como las máquinas de vectores de soporte (SVM) y las redes neuronales en la predicción del riesgo crediticio en productos de factoring financiero.

Chen et al. (2020) utilizaron SVM para capturar relaciones no lineales entre las características del cliente y el riesgo crediticio, demostrando una mejora significativa en la capacidad del modelo para generalizar sobre datos nuevos y no vistos previamente. Este enfoque permite una adaptación más dinámica a los cambios en el comportamiento financiero y las condiciones del mercado.

Por otro lado, Lee (2021) aplicó redes neuronales profundas para explorar patrones complejos y no lineales en los datos de factoring financiero, logrando una mayor precisión en la identificación de perfiles de riesgo crediticio que podrían no ser capturados por modelos tradicionales.

Comparación y Tendencias Futuras

La comparación entre modelos tradicionales y de aprendizaje automático revela un desplazamiento hacia métodos más complejos pero también más poderosos en la predicción del riesgo crediticio. Si bien los modelos tradicionales son valorados por su interpretabilidad, los avances en aprendizaje automático permiten capturar relaciones más sutiles y dinámicas entre las variables, adaptándose mejor a entornos financieros cambiantes y datos complejos.

En la presente tesis, abordaremos ambos tipos de modelos: inicialmente, un modelo tradicional y más simple, seguido de un modelo más complejo. Posteriormente, se realizará una comparación entre ambos.

Siguiendo con lo mencionado anteriormente, el propósito fundamental de esta tesis consiste en perfeccionar la gestión del riesgo crediticio mediante la utilización de datos relativos al comportamiento de pago de los deudores, las características empresariales y el comportamiento de ventas. En consonancia con este objetivo, se propone la elaboración y desarrollo de un modelo de aprendizaje automático.

Este enfoque busca capitalizar la riqueza de información disponible para lograr una evaluación más precisa y dinámica de los riesgos asociados a la concesión de créditos. El modelo de aprendizaje automático se erigirá como una herramienta avanzada que integrará diversas variables y patrones, permitiendo así una toma de decisiones más ágil y fundamentada en tiempo real.

La tesis se centrará en la formulación y optimización de este modelo, considerando cuidadosamente factores como la interpretabilidad del modelo, la robustez estadística

y la ética en el manejo de los datos. Al proponer y validar este modelo, se pretende contribuir significativamente a la eficiencia y eficacia de la gestión del riesgo crediticio en el contexto de la empresa, estableciendo un marco sólido para la toma de decisiones estratégicas en la concesión de créditos.

En el ámbito de la predicción de incumplimientos de crédito, la comparación de distintos modelos de minería de datos es esencial por varias razones técnicas fundamentales. En primer lugar, estos modelos están construidos sobre diferentes suposiciones y algoritmos, lo que influye en su capacidad para capturar la complejidad inherente de los datos financieros y de comportamiento del cliente. Al comparar múltiples modelos, se puede determinar cuál de ellos se ajusta mejor a los datos disponibles y proporciona predicciones más precisas y confiables (Yang & Zhang, 2018).

3.1 Aprendizaje supervisado

En el campo del aprendizaje automático, se resaltan las técnicas de aprendizaje supervisado. En este marco, se opera con un conjunto de atributos x, que pueden ser expresados como $x_{i,}\ i = 1, 2, \dots, n$, cada valor tiene asociado un resultado y_i (James et al., 2023).

$$Y = f(x) + \varepsilon \tag{1}$$

Este método de aprendizaje supervisado implica que el modelo se instruya utilizando un conjunto de datos en el cual se conoce la relación entre los atributos y los resultados correspondientes. El propósito es que el modelo aprenda a asignar eficientemente los atributos a los resultados, lo que posibilita hacer predicciones precisas en nuevos conjuntos de datos sin etiquetar. La estimación de la función f cobra una relevancia significativa por dos razones fundamentales: su capacidad predictiva y su capacidad para realizar inferencias (James et al., 2023).

Uno es la capacidad predictiva, que surge como un componente esencial en situaciones donde se dispone de un conjunto de variables de entrada X, pero la obtención de la variable de interés Y presenta un desafío considerable. En esta configuración particular, bajo la presunción de que el término de error tiene una media nula, se abre la posibilidad de anticipar Y mediante la siguiente expresión:

$$\hat{Y} = \hat{f}(x) \tag{2}$$

En donde $\hat{f}(x)$ representa nuestra estimación para $f(x)$ y $\hat{Y}$ representa la predicción resultante para Y. En este contexto, $\hat{f}(x)$ suele ser un modelo de caja negra dado que no es relevante entender la relación que hay entre X e Y en la medida que la predicción sea precisa (James et al., 2023).

Este enfoque de predicción, respaldado por la condición de que el término de error promedie a cero, fundamenta la confianza en la precisión y utilidad de las predicciones resultantes. La capacidad de anticipar Y basándose en las variables de entrada disponibles se convierte en un recurso invaluable para la toma de decisiones estratégicas.

La inferencia tiene el interés en comprender como es afectado Y ante una alteración en $x_{i,}\ i = 1, 2, \dots, n$. En esta situación deseamos estimar $\hat{f}(x)$, pero nuestro objetivo no es necesariamente para hacer predicciones para Y. Ahora $\hat{f}$ no puede ser tratado como una caja negra, porque necesitamos saber su forma exacta. En este contexto, uno puede estar interesado en responder las siguientes preguntas: : (James et al., 2023).
¿Qué variables predictoras están vinculadas con la respuesta? Frecuentemente, solo una pequeña parte de las variables predictoras disponibles están notablemente relacionadas con Y.

¿Cuál es la relación entre la respuesta y cada predictor? Algunos predictores pueden tener una relación positiva con Y, en el sentido que valores más grandes del predictor

están asociados con valores más grandes de Y. Otros predictores pueden tener la relación opuesta.

¿Es posible resumir adecuadamente la relación entre Y y cada variable predictora mediante una ecuación lineal, o la relación es más compleja? En ciertas situaciones, esta suposición es razonable o incluso deseable. Sin embargo, en muchos casos, la verdadera relación es más complicada, en cuyo caso un modelo lineal puede no ofrecer una representación precisa de la relación entre las variables de entrada y salida (James et al., 2023).

En este contexto de inferencia, la transparencia y comprensión detallada de la forma de f son cruciales para extraer conocimiento sobre la relación entre las variables de entrada y la variable de interés. Estas interrogantes apuntan a profundizar en la interpretación y el entendimiento del modelo, proporcionando información valiosa más allá de su capacidad predictiva.

En el transcurso de la presente tesis, se avanzará en el desarrollo de modelos con un enfoque predominante en la optimización del rendimiento predictivo, priorizando esta consideración por encima de la capacidad de realizar inferencias detalladas.

Este enfoque estratégico se orienta a maximizar la exactitud y la eficacia en la predicción de la variable de interés, aunque pueda resultar en modelos más complejos y menos interpretables desde el punto de vista inferencial. La meta primordial es lograr una capacidad predictiva sobresaliente, especialmente cuando se enfrenta a situaciones en las cuales la variable de interés no está fácilmente accesible y se requiere realizar predicciones precisas basadas en un conjunto dado de variables de entrada.

Este planteamiento busca explorar y aprovechar las capacidades avanzadas de los modelos enfocados en la predicción, reconociendo que, en determinados contextos, la interpretación detallada del modelo puede no ser la prioridad principal. La evaluación y validación de estos modelos se llevarán a cabo con un énfasis particular en métricas de

rendimiento predictivo, proporcionando así una perspectiva clara sobre su eficacia en escenarios específicos.

3.2 "Trade-off" entre la precisión en predicción y la interpretabilidad del modelo

Se plantea un "trade-off" fundamental entre la precisión en predicción y la interpretabilidad del modelo en el marco de esta investigación. Los métodos que se emplearán pueden variar en su flexibilidad, lo que influye directamente en la diversidad de formas que $\hat{f}$ puede tomar.

Algunos métodos, como la regresión lineal, son considerados menos flexibles o más restrictivos, ya que generan únicamente una gama relativamente estrecha de formas para la función estimada $\hat{f}$. Este enfoque es limitado en su capacidad para capturar patrones complejos o no lineales en los datos, pero a menudo ofrece modelos más fácilmente interpretables.

Contrastando con esto, modelos como los *splines* permiten una mayor flexibilidad. Estos modelos, que definen funciones lineales por segmentos en el dominio, ofrecen la posibilidad de adaptarse a patrones más intrincados y no lineales en los datos. No obstante, esta mayor flexibilidad puede dar lugar a modelos más difíciles de interpretar, ya que la relación entre las variables de entrada y la variable de interés puede volverse más compleja (James et al., 2023).

La elección entre métodos más rígidos y modelos más flexibles dependerá de la importancia relativa de la precisión en predicción y la interpretabilidad en el contexto específico de la investigación. Esta consideración del "trade-off" entre precisión e interpretabilidad guiará la selección de las herramientas analíticas más apropiadas para los objetivos de la tesis.

3.3 Modelado de problemas de clasificación

Las variables en estudio pueden ser clasificadas en dos categorías principales: cuantitativas y cualitativas (también conocidas como categóricas). Las variables cuantitativas se caracterizan por tomar únicamente valores numéricos, proporcionando información de naturaleza numérica o de cantidad. En contraste, las variables cualitativas, también denominadas categóricas, toman valores que se distribuyen en distintas clases o categorías (James et al., 2023).

Ejemplificando, las variables cuantitativas pueden representar medidas como las ventas domésticas medidas en dólares de una empresa, las exportaciones medidas en dólares, las ventas domesticas o exportaciones promedio en un período determinado, el coeficiente de variación para determiner la volatidad de ventas, etc. Estas variables ofrecen información numérica que puede ser sometida a análisis matemático y estadístico.

En contraste, las variables cualitativas representan atributos que pertenecen a diferentes categorías discretas. Por ejemplo, si la empresa se dedica principalmente a la exportación de productos perecederos, o no. Estas variables son fundamentales para comprender la diversidad de características presentes en un conjunto de datos, y su análisis requiere enfoques diferentes a los empleados para variables cuantitativas (James et al., 2023).

Esta distinción entre variables cuantitativas y cualitativas proporciona una base fundamental para la comprensión y el análisis de datos en diversas disciplinas.

Como fue mencionado, el objetivo del modelo desarrollado en la presente tesis es clasificar el perfil de pago de empresas que defaultean, y empresas que repagan el crédito. Dicho problema puede ser modelado como una clasificación binaria, expresando con un “1” en caso de *default* y “0” en caso contrario.

3.4 Técnicas de modelado de variables (*Feature engineering*)

La creación del conjunto de datos para el entrenamiento del modelo merece una mención especial, ya que constituye una parte significativa del trabajo realizado. En este contexto, se parte de tablas con datos transaccionales, como la tabla de préstamos generados, que representa información interna de la compañía, y la tabla de facturación de las empresas sujetas a análisis. Esta última información fue adquirida por la empresa con el consentimiento del interesado a través de la API del Servicio de Administración Tributaria.

Se han empleado principalmente estos conjuntos de datos como base para aplicar métodos avanzados de *Feature Engineering*, los cuales se detallarán a continuación.

Es crucial subrayar en este punto que los datos han sido reagrupados a nivel de factura. A través de las técnicas de ingeniería de atributos, transformamos estas facturas en características pertinentes en el contexto de la calificación crediticia. Es decir, para la variable de ventas domésticas del último mes, utilizando la fecha de interés, por ejemplo, el 14 de abril de 2021, consideramos exclusivamente las facturas emitidas desde el 15 de marzo hasta el 14 de abril para el cálculo de dicha variable. Este enfoque garantiza una representación precisa y relevante de la información temporal en la construcción de características para el análisis crediticio.

En el proceso de construcción y computo de variables, adoptamos como unidad de observación la relación entre la empresa y la fecha en la que ésta tomó el crédito en cuestión. En este enfoque, para el cálculo de las variables, utilizamos esta relación considerando información retrospectiva y excluyendo todos los datos correspondientes a esa empresa con fecha posterior a la referencia temporal. Esta estrategia se implementa para prevenir el "*Data Leakage*" o fuga de datos, asegurando que las variables se calculen exclusivamente con información disponible hasta la fecha de referencia y evitando la inclusión de datos futuros que podrían influir en la predicción del modelo. De esta manera, se garantiza una correcta representación de la realidad temporal en la construcción de las características, mejorando la integridad y validez del modelo predictivo.

El modelado de variables financieras desempeña un papel crucial en la evaluación de la performance crediticia al proporcionar una comprensión detallada de la salud financiera de una empresa. Al analizar variables como el apalancamiento, la rentabilidad, la liquidez y la actividad comercial, junto con factores macroeconómicos relevantes, los modelos pueden predecir con mayor precisión la probabilidad de incumplimiento financiero. Este enfoque estructural permite a las instituciones financieras y otras partes interesadas evaluar y gestionar el riesgo crediticio de manera más efectiva, lo que resulta fundamental para la toma de decisiones informadas en la concesión de créditos y la gestión de carteras. Además, al centrarse en el contexto específico del mercado y las características de las empresas evaluadas, estos modelos pueden adaptarse mejor a las dinámicas crediticias cambiantes, mejorando así la capacidad de pronóstico y reduciendo la exposición al riesgo crediticio. (Rikkers & Thibeault, 2015).

Un *feature* o variable se define como una representación numérica de un aspecto específico de los datos en bruto. Estas variables, desempeñan un papel crucial en el proceso de aprendizaje automático, actuando como el puente entre los datos sin procesar y los modelos. La ingeniería de atributos, a menudo denominada *data* engineering, consiste en la extracción de características de los datos en bruto y su transformación en formatos apropiados para el modelo de aprendizaje automático (Zheng & Casari, 2018).

Este proceso es esencial, ya que las características correctas pueden mitigar la complejidad del modelado, permitiendo así que el proceso genere resultados de mayor calidad y mejore el rendimiento general del modelo. La ingeniería de atributos contribuye significativamente a la capacidad del modelo para comprender y aprender patrones relevantes en los datos, facilitando así la toma de decisiones más precisa y eficiente.

La cantidad de *features* o variables también es importante para un modelo de aprendizaje automático. Si no hay suficientes variables en el conjunto de datos,

entonces el modelo no podrá aprender el patrón o el comportamiento deseado(Zheng & Casari, 2018).

3.4.a Binarización

La binarización es una técnica que implica la conversión de variables a valores de 1 o 0, especialmente aplicada a variables ordinales o numéricas. Un ejemplo ilustrativo de binarización sería representar con un valor de 1 si la variable de ventas mensuales supera el monto de la cuota del crédito adeudado, y con un valor de 0 en caso contrario. Esta estrategia simplifica la representación de la información y facilita la interpretación de las relaciones.

3.4.b Cuantización

La cuantización es una técnica similar a la binarización, pero a diferencia de convertir las variables exclusivamente en 1 o 0, ofrece la posibilidad de asignar la variable a una gama más amplia de valores. Un ejemplo concreto para comprender este concepto sería la agrupación de personas por edades, como [0 a 12 años], [13 a 17 años], y así sucesivamente. En este tipo de cuantización, es esencial contar con un conocimiento profundo del dominio y del problema de negocio.

Otra forma de cuantización implica agrupar los datos en cantidades iguales por cada grupo, creando cuantiles, sextiles, deciles, entre otros. En términos generales, la cuantización se utiliza para transformar variables numéricas en variables discretas ordinales, proporcionando una representación más manejable y estructurada de la información.

3.4.c Escalado Min-Max

En el contexto de una variable numérica con valores en el dominio de los números reales, la técnica de escalado Min-Max se aplica considerando los valores mínimo y máximo que la variable puede tomar en el conjunto de datos. La transformación Min-Max implica tomar el valor de cada observación, restarle el valor mínimo, y luego dividirlo por el rango, que se define como la diferencia entre el valor máximo y el valor mínimo que la variable puede abarcar.

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

3.4.d Estandarización

De manera similar, la estandarización de una variable implica restarle el valor promedio de la variable y dividir por la desviación estándar. Este proceso transforma los valores de la variable para que tengan una media de cero y una desviación estándar de uno. La estandarización es útil para comparar y analizar variables que pueden tener diferentes unidades o escalas, ya que coloca todas las variables en una escala común, facilitando la interpretación y comparación de sus efectos en el conjunto de datos.

$$\tilde{x} = \frac{x - \text{promedio}(x)}{\sigma(x)} \tag{4}$$

Escalarización: La escalarización representa un paso crucial en el proceso de *feature engineering*, donde se busca homogeneizar las escalas de las variables para garantizar comparaciones significativas y coherentes. Esta técnica implica la normalización de datos mediante escalares, lo que permite transformar las variables sin perder el orden numérico, pero alterando la relación mensurable entre ellas. Por ejemplo, en una escalarización lineal, se aplican multiplicadores a los datos para ajustarlos, mientras que en la escalarización logarítmica se emplean transformaciones logarítmicas. Otras técnicas comunes incluyen la estandarización, que ajusta los datos para que tengan una media de cero y una desviación estándar de uno. Esta fase de escalarización es esencial

para preparar los datos antes de aplicar algoritmos de *machine learning*, mejorando así la convergencia y el rendimiento del modelo resultante.

3.4.e Índice de Herfindahl-Hirschman (HHI)

El Índice de Herfindahl-Hirschman (HHI) es una medida de concentración principalmente utilizada para medir concentración de mercado y, por ende, el grado de competencia entre las empresas de un sector. En el cual se suma la potencia de la proporción en cada agente del mercado. (Rhoades, 1993).

$$\text{HHI} = \ 10{,}000 \ \times \ \sum_{i=1}^{N} s_i^2 \qquad (5)$$

En donde N representa el número total de entidades, y s_i representa la proporción sobre que representa dicha entidad sobre el total.

En la presente tesis, el indicador HHI será utilizado como una métrica para medir la concentración de clientes y proveedores de las empresas exportadoras.

3.4.f Espacios de características

Como concepto general, los espacios de características se refieren al conjunto de variables utilizadas para describir un conjunto de datos. En el *feature engineering*, explorar y diseñar espacios de características adecuados es esencial para capturar la información relevante y discriminativa para el problema en cuestión. Esto puede implicar la combinación de características existentes, la creación de nuevas características a partir de las disponibles o la selección de las características más importantes mediante técnicas como la eliminación de características irrelevantes o la reducción de dimensionalidad. Al diseñar cuidadosamente el espacio de características, se puede mejorar la capacidad predictiva de los modelos de machine learning y facilitar una mejor comprensión del problema subyacente.

Un caso de modelado de variable basado en espacios de características es el algoritmo de *K-means* que nos interesa desarrollar a continuación.

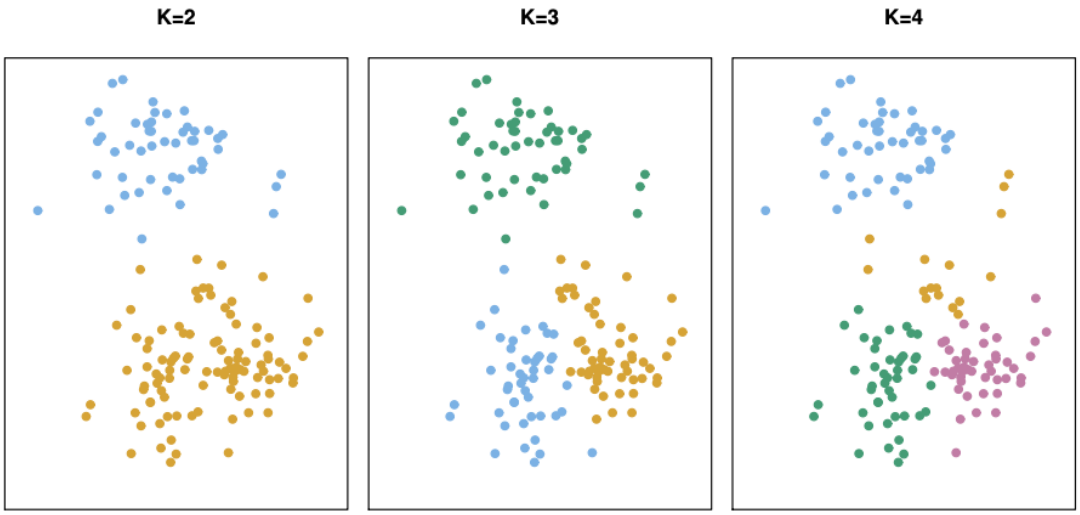
3.4.f.1 *K-means*

El algoritmo de *K-means* pertenece al ámbito del aprendizaje automático, específicamente al aprendizaje no supervisado. En este enfoque, se busca analizar datos donde se tienen las características del conjunto $x_i = x_1, \dots, x_n$ pero estas no están vinculadas a una variable de respuesta y_i. En este contexto, de alguna manera, estamos trabajando a ciegas; la situación se denomina no supervisada porque carecemos de una variable de respuesta que pueda guiar nuestro análisis. El algoritmo de K-means se utilizará para generar una nueva variable en los datos, la cual será empleada en el proceso de aprendizaje del modelo (James et al., 2023).

El algoritmo *K-means* es reconocido por su capacidad para generar *clusters* a partir de conjuntos de datos de entrenamiento. El término "*cluster*" se refiere a un conjunto amplio de técnicas que buscan identificar subgrupos o conglomerados dentro de un conjunto de datos. Al realizar el agrupamiento de observaciones en un conjunto de datos, el objetivo es dividirlas en grupos de tal manera que las observaciones dentro de cada grupo sean bastante similares entre sí, mientras que las observaciones en grupos diferentes sean notablemente diferentes entre sí. En este contexto, el concepto de "cluster" se utilizará para describir el proceso de agrupamiento de datos mediante técnicas de aprendizaje no supervisado (James et al., 2023).

K-means es una metodología sencilla para dividir un conjunto de datos en K grupos distintos y no superpuestos. Para llevar a cabo la agrupación mediante *K-means*, es necesario definir previamente el número deseado de agrupaciones, K. Posteriormente, el algoritmo de *K-means* asigna cada observación de manera exclusiva a uno de los K grupos.

Figura 1) Ejemplo de K-Means Clustering.



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 12.

La figura 1) muestra un ejemplo de cómo el algoritmo *K-Means Clustering* agrupa datos en clusters basados en sus características. Cada color representa un cluster identificado por el algoritmo *K-Means*. El ejemplo ilustra cómo *K-Means* puede ser aplicado para segmentar datos en un conjunto de datos simulados de ejemplo.

En este ejemplo, se presentan los resultados derivados de la aplicación del algoritmo de agrupamiento K-medias en un escenario simulado que comprende 150 observaciones bidimensionales. Se exploraron tres valores distintos de K en el proceso de agrupamiento con el fin de examinar la variación en la estructura de los grupos resultantes.

El procedimiento de agrupamiento de K-medias resulta de un problema matemático simple e intuitivo. Comenzamos definiendo alguna notación. Sea $c_1, \dots, c_k$ denota conjuntos que contienen los índices de las observaciones en cada grupo. Estos conjuntos satisfacen dos propiedades:

1) $c_1 \cup c_2 \cup \dots \cup c_k = \{1, \dots, n\}$ Cada observación pertenece a al menos una clase.
2) $c_k \cap c_{k'} = \emptyset \; \forall \; k \neq k'$ Ninguna observación pertenece a más de un *cluster*.

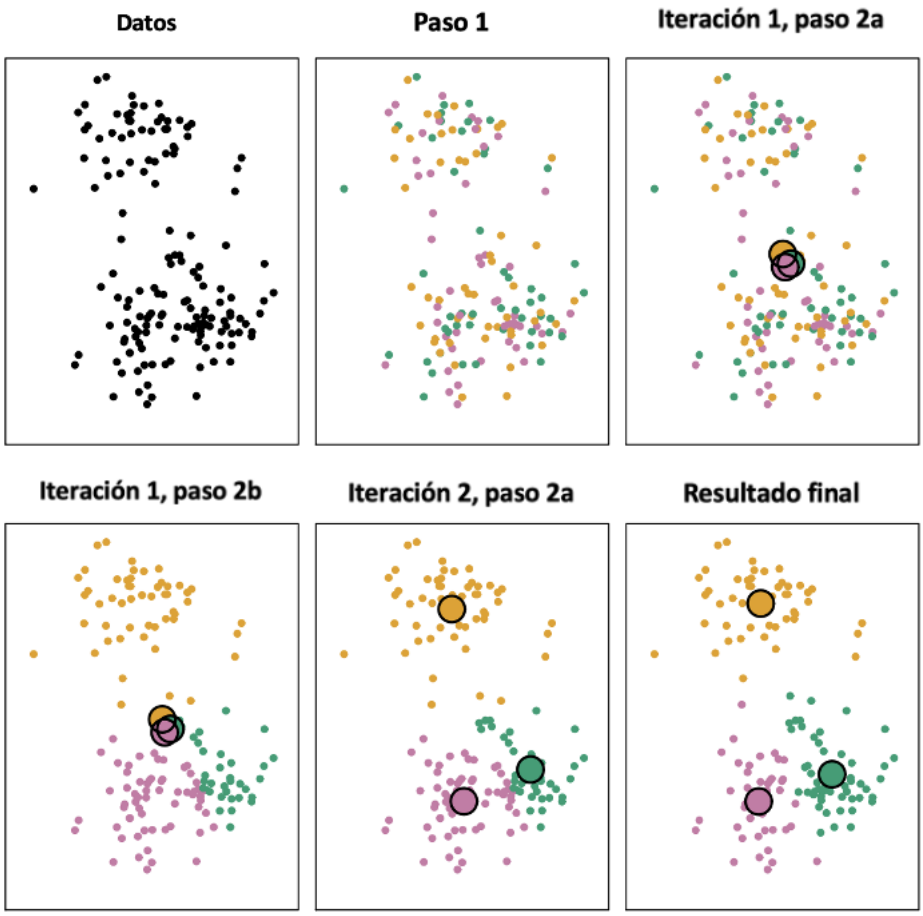
La premisa subyacente en el *clustering* de K-medias es que la efectividad de un agrupamiento se define por una variación mínima dentro de cada cluster (James et al., 2023).

$$\frac{minimize}{c_1,\ldots,c_k}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\} \quad (6)$$

El procedimiento de K-medias opera de la siguiente manera (James et al., 2023).

1. Asigna aleatoriamente un número, del 1 al K, a cada observación como una asignación inicial de grupos.
2. Repite iterativamente hasta que las asignaciones de los *clusters* dejen de cambiar:
    a. Para cada uno de los K grupos, calcula el centroide del grupo. El centroide del k-ésimo grupo es el vector que representa las medias de las características para las observaciones en el k-ésimo grupo.
    b. Asigna cada observación al grupo cuyo centroide esté más cercano, donde la proximidad se define mediante la distancia euclidiana.

Figura 2 Progreso del algoritmo K-Means en el ejemplo de la Figura 1 con K=3



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 12.

Figura 2) Muestra como el K-means funciona con K=3. En la parte superior izquierda se presentan las observaciones, mientras que en la parte superior derecha, en el paso 1 del algoritmo, cada observación se asigna aleatoriamente a un grupo. Posteriormente, en el paso 2(a), se calculan los centroides del grupo, representados como grandes discos de colores. Inicialmente, los centroides se superponen casi por completo debido a que las asignaciones iniciales de conglomerados se eligen al azar. En la parte inferior izquierda, en el paso 2(b), cada observación se asigna al centroide más cercano. En el centro inferior, se repite el paso 2(a), dando lugar a nuevos centroides de grupo. Finalmente, en la parte inferior derecha, se muestran los resultados después de diez iteraciones del algoritmo (James et al., 2023).

3.4.f.2 Selección de k óptimo a través del método del codo

Se emplea ampliamente en el análisis de agrupamiento para determinar el número más apropiado de conjuntos en un conjunto de datos. El término deriva de la forma de la curva generada al graficar la suma de las distancias al cuadrado de cada punto respecto al centroide más cercano en relación con el número de agrupaciones (k). Esta técnica desempeña un papel fundamental en el proceso de agrupamiento al permitir una identificación objetiva del número óptimo de grupos para representar la estructura inherente de los datos (James et al., 2023).

La noción básica detrás del enfoque del codo radica en la observación de que, conforme se incrementa el número de conjuntos, la suma de las distancias al cuadrado tiende a decrecer, ya que los centroides se ajustan mejor a los puntos individuales. No obstante, llega un punto en el que agregar más conjuntos no conlleva una mejora sustancial en la estructura de agrupamiento, y la disminución en la suma de las distancias al cuadrado se vuelve menos marcada. Este punto de inflexión se asemeja a un "codo" en la curva, y se considera como la ubicación del número óptimo de conjuntos (James et al., 2023).

La identificación del punto de codo puede llevarse a cabo visualmente al examinar la curva resultante, aunque también existen métodos más formales para su determinación, como el análisis de la tasa de cambio de la suma de las distancias al cuadrado en función de k (James et al., 2023).

3.5 Modelos de aprendizaje supervisado para el problema de clasificación

3.5.a Regresión Logística

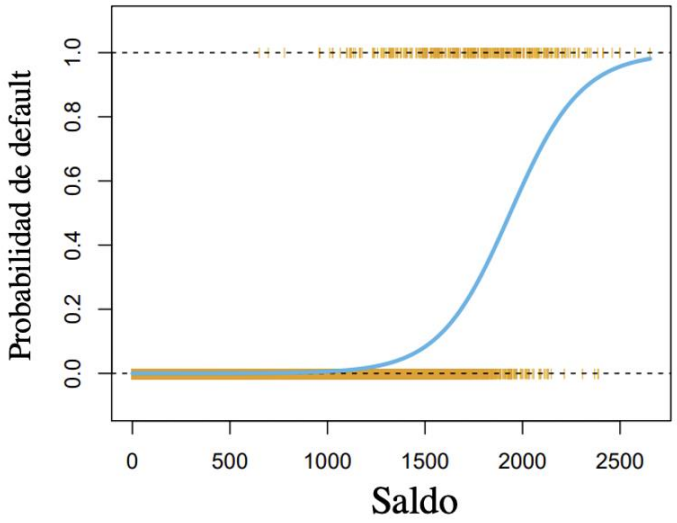
La regresión logística se encarga de modelar la probabilidad que Y pertenezca a una categoría en particular. Dicha función tiene fomra sigmoidal o de sigmoide, lo que significa que tiene una curva en forma de "S". Esto la hace útil para modelar procesos que tienen un crecimiento inicial acelerado seguido de una estabilización. Para los datos

de incumplimiento de pago, la regresión logística modela la probabilidad de dicho incumplimiento. Por ejemplo, la probabilidad de incumplimiento dado el saldo se puede escribir como para abordar esta cuestión. En este ejemplo existe una una variable X, entre las metodologías que proponemos se incluyen:

$$\Pr(default = Si \mid saldo) \tag{7}$$

Los valores de la probabilidad de que ocurra un incumplimiento, representados como $\Pr(default \mid saldo)$, varían entre 0 y 1. Esto significa que para cualquier cantidad de saldo, podemos hacer una predicción sobre si habrá un incumplimiento. Por ejemplo, si la probabilidad de incumplimiento es mayor que 0.5, podríamos predecir que el incumplimiento será sí para cualquier individuo. Por otro lado, si una empresa quiere ser más cautelosa al predecir quiénes podrían incumplir, puede elegir un umbral más bajo, como una probabilidad de incumplimiento mayor que 0.1. (James et al., 2023).

Figura 3) Representación gráfica de la función logísitca.



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4.

La figura 3: Es una representación gráfica de la regresión logística todas las valores de probabilidades se encuentran dentro del rango 0 a 1, representada con una forma de sigmoide.

La regresión logística se percibe comúnmente como un modelo robusto para llevar a cabo inferencia estadística detallada. Su capacidad para proporcionar interpretaciones claras de los efectos de las variables explicativas en la probabilidad de un resultado binario lo hace valioso en contextos donde la interpretabilidad es fundamental. No obstante, algunos estudios sugieren que la regresión logística puede tener limitaciones en términos de poder predictivo en comparación con otros modelos.

Este modelo puede no capturar relaciones complejas o no lineales de manera tan efectiva. Modelos más flexibles, como *support vector machine*, *random forest* o *neuronal network*, podrían superar a la regresión logística en términos de precisión predictiva en ciertos escenarios.

La elección entre la regresión logística y modelos más avanzados dependerá de los objetivos específicos de la tarea. Si se prioriza la interpretabilidad y la capacidad de realizar inferencias detalladas, la regresión logística puede ser preferida. Sin embargo, si se busca un mayor rendimiento predictivo, la exploración de modelos más complejos podría ser beneficiosa, considerando siempre la compensación entre precisión y capacidad de interpretación.

Entonces la regresión logística se basa en la función logística, que dé genera como *output* valores Y entre 0 y 1 para todos los valores de x. (a continuación la función analítica para el caso de una sola variable predictora)

$$p(x) = \frac{e^{\beta_0 + e^{\beta_1}}}{1 + e^{\beta_0 + e^{\beta_1}}} \qquad o \qquad \frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} \tag{8}$$

La consideración de x como la variable de saldo en cuenta corriente para predecir el *default* o repago de un crédito es un escenario típico en el análisis de riesgo crediticio. En este contexto, la función logística se revela como una elección prudente, ya que aborda de manera efectiva los desafíos asociados con la predicción de probabilidades en un rango acotado.

Al emplear la función logística, se garantiza que las predicciones estén confinadas en el intervalo [0, 1], lo cual es crucial al modelar probabilidades. Para saldos bajos, la función logística predice una probabilidad de incumplimiento cercana a cero, pero nunca por debajo, evitando así problemas de interpretación que surgirían con una regresión lineal. Del mismo modo, para saldos altos, se predice una probabilidad de incumplimiento cercana a uno, pero nunca por encima de uno, mitigando el riesgo de predicciones poco realistas (James et al., 2023).

Podemos deducir además que $\frac{p(X)}{1-p(X)}$ o $\beta_0 + \beta_1 x$ puede tomar cualquier valor de cuota entre 0 e infinito. Los valores de las probabilidades cercanos a 0 y infinito indican muy bajos y muy altas probabilidades de incumplimiento, respectivamente. Si aplicamos el logaritmo en ambos lados de la segunda función análitica en la referencia (8) podemos deducir (James et al., 2023).

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \tag{9}$$

Esto es importante para realizar inferencia, sin embargo en un modelo de regresión logística, aumentar X en una unidad cambia el log probabilidades en β_1. De manera equivalente, multiplica las probabilidades por $e^{\beta 1}$. Sin embargo, debido a que la relación entre $p(x)$ y X en no es una línea recta, $\beta_1$no corresponde al cambio en $p(x)$ asociado con una unidad aumento en X. La cantidad que $p(x)$ cambia debido a un cambio de una unidad en X depende del valor actual de X. Pero independientemente del valor de X, si β_1 es positivo, entonces el aumento de X estará asociado con el aumento de $p(x)$, y si β_1 es negativo, entonces aumentar X se asociará con disminuir $p(x)$. El hecho de que no existe una relación lineal entre $p(x)$ y X, y el hecho de que la tasa de cambio en $p(x)$ por unidad de cambio en X depende del valor actual de X (James et al., 2023).

La forma en S de la función logística, independientemente del valor de X, proporciona una predicción sensible y lógica. Este comportamiento característico de la función logística la convierte en una elección sólida para abordar problemas de clasificación

binaria, como el pronóstico de *default* o repago en la evaluación crediticia, al tiempo que preserva la coherencia y la interpretabilidad en todo el rango de valores de X.

3.5.a.1 Estimación de los coeficientes de regresión

Los coeficientes β_0 y β_1 en el modelo de regresión logística son parámetros desconocidos que deben ser estimados a partir de los datos de entrenamiento disponibles. En el contexto de modelos de clasificación, se emplea el método de máxima verosimilitud para llevar a cabo esta estimación, ya que este método posee propiedades estadísticas favorables.

La máxima verosimilitud busca encontrar los valores de los parámetros que maximizan la probabilidad de observar los datos que realmente se han observado, dada una distribución asumida. En el caso de la regresión logística, esto implica encontrar los β_0y $\beta_1$que maximizan la probabilidad conjunta de observar las respuestas binarias reales dadas las variables predictoras X.

Para ajustar un modelo de regresión logística, buscamos estimaciones para $\hat{\beta}_0$ y $\hat{\beta}_1$ tal que la probabilidad prevista $\hat{p}(x_i)$ de incumplimiento para cada individuo. Tratamos de encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que reemplazando estas estimaciones en el modelo para $p(x)$, 1 un número cercano a uno para todos los individuos que incumplieron el pago, y un número cercano a cero para todas las personas que no lo hicieron. Esta intuición se puede formalizar mediante una expresión matemática llamada función de verosimilitud (James et al., 2023).

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \quad (10)$$

Las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ se eligen para maximizar esta función de verosimilitud. Dado que los valores $\hat{\beta}_i$ se encuentran dentro de $p(x_i)$.

En efecto, la extensión del modelo de regresión logística para abordar el problema de predecir una respuesta binaria utilizando múltiples predictores se realiza de manera análoga a la extensión del modelo de regresión lineal simple a múltiple.

La forma general del modelo de regresión logística con múltiples predictores se expresa como (James et al., 2023).

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad \text{o} \qquad p(x) = \frac{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}{1+\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} \quad (11)$$

3.5.b Modelos de Regresión basados en árboles.

Los modelos de ensambles de árboles han tomado una mayor relevancia en los últimos años debido al potencial que estos han demostrado con su gran poder predictivo. Por ejemplo en China se ha experimentado un gran crecimiento en la industria financiera a partir de modelos de créditos entre pares (P2P). El mismo propone la utilización de ensamble de modelos de árboles como una estrategia efectiva y prometedora para mejorar la precisión de estas predicciones (Chen, Ding, Li, & Yang, 2018).

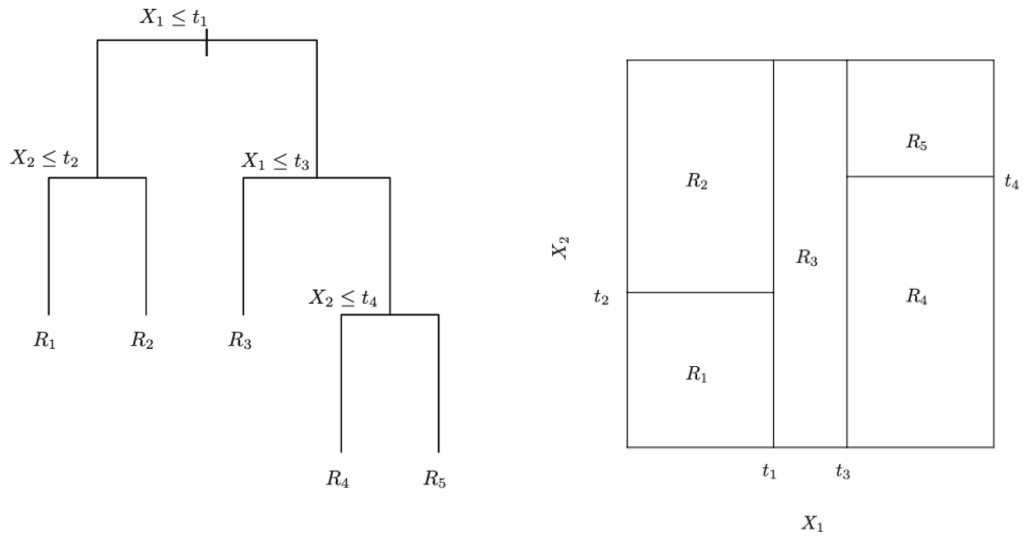
Estos métodos implican la segmentación del espacio de predictores en regiones más simples a través de decisiones basadas en las características de los datos. La construcción de un árbol de decisión implica dividir recursivamente el conjunto de datos en subconjuntos más pequeños, de modo que las observaciones dentro de cada subconjunto compartan características similares (James et al., 2023)

En un caso de clasificación binaria, el árbol de decisión se utiliza para predecir la probabilidad de pertenencia a una clase dada para una observación específica. Para hacer esto, se evalúa el camino a lo largo del árbol que la observación sigue, y la probabilidad se calcula tomando el promedio de las clases de las observaciones de entrenamiento dentro de la región a la que pertenece la observación.

Las decisiones en el árbol de decisión están determinadas por reglas de división que se aplican a las características de los datos en cada nodo del árbol. La construcción del árbol continúa hasta que se alcanza un criterio de parada, como una profundidad máxima predefinida o cuando el número de observaciones en un nodo es inferior a un umbral.

Debido a su naturaleza de representación en forma de árbol, estos métodos se conocen como métodos de árbol de decisión. Son ampliamente utilizados en el aprendizaje automático para problemas de clasificación y regresión debido a su capacidad para manejar relaciones no lineales y su relativa facilidad de interpretación.

Figura 4) Funcionamiento del modelo de árbol de decision con dos predictores.



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 8.

La figura 4) Muestra como gráficamente podemos visualizar un árbol. Las *variables* x_1*y* x_2 *representan las variables predictoras, mientras que* t_1, t_2, t_3 *y* t_4 *son los puntos de corte del árbol. En el ejemplo podemos ver 5 distintas regiones representadas por* R_i*. En modo de ejemplo, si en entrenamiento encontramos cien observaciones quedan en* R_1 *de las cuales una sola observación corresponde a la categoría de default. Entonces a la hora de utilizar este modelo todas las observaciones que recaigan sobre* R_1*tendrá una probabilidad de default del uno sobre cien* (James et al., 2023).

El procedimiento delineado previamente tiene el potencial de generar predicciones precisas en el conjunto de entrenamiento; sin embargo, existe la probabilidad de que se produzca un sobreajuste a los datos, lo cual resultaría en un rendimiento subóptimo en el conjunto de validación. Este fenómeno se traduce en la propensión del árbol a generar un número significativo de regiones R_i. La elección de un árbol más pequeño, con menos divisiones y, por ende, menos regiones R_i, puede mitigar la varianza, mejorar la interpretabilidad. No obstante, es importante señalar que esta alternativa tiende a ser demasiado genérica y poseer un poder predictivo limitado (James et al., 2023).

La estrategia para generar los puntos de corte en árboles de decisión puede variar según la implementación, pero en general, los modelos suelen utilizar el Índice de Gini.

El Índice de Gini es una medida de impureza que evalúa qué tan mezcladas están las clases en un nodo del árbol. Cuanto menor sea el Índice de Gini, más puros y homogéneos son los grupos resultantes después de una división. En el proceso de construcción del árbol, se busca minimizar el Índice de Gini en cada paso, eligiendo la división que proporciona la mayor pureza en los nodos hijos.

Esta técnica, basada en el Índice de Gini, es una aproximación común para la selección de puntos de corte en árboles de decisión, y su aplicación proporciona un método efectivo para construir divisiones que optimizan la homogeneidad dentro de cada subconjunto.

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \, (1 - \hat{p}_{mk}) \tag{12}$$

La lógica subyacente en la utilización del Índice de Gini reside en la búsqueda de un punto de corte que logre una separación efectiva entre las clases. El objetivo es asegurar que, después de la división, la mayor proporción de la clase 0 se encuentre a un lado del corte, mientras que, en el otro lado, se ubique la mayor proporción posible de las demás clases. En esencia, se busca minimizar la impureza en cada nodo del árbol, fomentando la formación de subconjuntos más homogéneos y facilitando la toma de decisiones

efectiva en el proceso de clasificación. Este enfoque, basado en el Índice de Gini, contribuye a construir árboles de decisión que optimizan la pureza y la eficacia en la separación de clases (James et al., 2023).

3.5.b.1 Técnica de bagging en modelos de árboles

El *bagging*, que es una abreviatura de "bootstrap aggregating", es un procedimiento de propósito general diseñado para reducir la varianza de un método de aprendizaje estadístico. Se aplica en conjuntos de n observaciones independientes, $z_1, \dots, z_n$, cada una con varianza σ^2, la varianza de la media $\bar{z}$ de las observaciones está dada por $\frac{\sigma^2}{n}$.

En términos más simples, la esencia del *bagging* radica en el promedio de un conjunto de observaciones para reducir la varianza. En consecuencia, una estrategia natural para disminuir la varianza y mejorar la precisión en el conjunto de pruebas de un método de aprendizaje estadístico es utilizar el *bagging*. Este procedimiento implica la realización de muestras aleatorias del conjunto de entrenamiento para generar múltiples conjuntos de entrenamiento. Posteriormente, se construye un modelo de predicción utilizando cada conjunto de entrenamiento de manera independiente, y finalmente, se promedian las predicciones resultantes. Este enfoque contribuye a obtener modelos más robustos y generalizables al reducir la variabilidad asociada con un único conjunto de entrenamiento. En un sentido más claro, podríamos calcular $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ utilizando B conjuntos de entrenamiento distintos y luego promediarlos. Este enfoque nos proporcionaría un único modelo de aprendizaje estadístico con baja varianza, representado por la siguiente expresión (James et al., 2023).

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \tag{13}$$

Donde $\hat{f}_{avg}(x)$ es el modelo combinado que beneficia de la diversidad introducida por el bagging, y B es la cantidad de conjuntos de entrenamiento utilizados. Este proceso de

promedio contribuye a suavizar las predicciones y reducir la variabilidad inherente a un único modelo, mejorando así la capacidad de generalización del modelo resultante.

3.5.b.2 El modelo de Random Forest

Random forest representan una mejora con respecto a los árboles de *bagging* al introducir un ajuste sutil que des-correlaciona los árboles. Aunque la construcción de árboles de decisión a partir de muestras de entrenamiento es un proceso compartido con el *bagging*, *random forest* incorpora una dimensión adicional de aleatoriedad. En este método, cada vez que se obtiene una muestra aleatoria de observaciones para la construcción de un árbol, también se realiza una selección aleatoria de un subconjunto de predictores o variables X. Esta elección aleatoria de características en cada iteración contribuye a reducir la correlación entre los árboles, mejorando así la diversidad del conjunto y, en consecuencia, la capacidad predictiva del modelo resultante (James et al., 2023).

Al construir un *random forest*, en cada división del árbol, generalmente no se utilizan la mayoría de los predictores disponibles. Aunque esta práctica pueda parecer peculiar, se sustenta en una lógica inteligente. Supongamos que existe un predictor extremadamente fuerte en el conjunto de datos, junto con varios otros predictores moderadamente fuertes. En la construcción de la colección de árboles mediante bagging, es probable que la mayoría o todos los árboles empleen este predictor fuerte en la división inicial. Como consecuencia, los árboles en la colección *bagging* tenderán a ser bastante similares entre sí. Esta similitud lleva a que las predicciones de los árboles *bagging* estén altamente correlacionadas (James et al., 2023).

Lamentablemente, promediar muchas cantidades altamente correlacionadas no resulta en una reducción significativa de la varianza, a diferencia de promediar cantidades no correlacionadas. En particular, esto implica que el *bagging* no conduce a una reducción sustancial de la variabilidad en comparación con un solo árbol en este escenario. Este fenómeno destaca la importancia de la estrategia de Random Forest de introducir más

aleatoriedad en la selección de predictores, descorrelacionando así los árboles y mejorando la eficacia del conjunto (James et al., 2023).

El modelo de Random Forest aborda este problema al imponer que cada nuevo árbol considere únicamente un subconjunto de predictores en cada división. En consecuencia, la tasa de predictores a utilizar generalmente se convierte en un hiperparámetro que requiere identificación. Este proceso puede entenderse como una estrategia de descorrelación entre los árboles, lo que se traduce en que el promedio de los árboles resultantes sea menos variable y, por ende, más confiable (James et al., 2023).

3.5.b.3 El modelo de Árboles basados en boosting

Boosting opera de manera análoga a *bagging*, aunque con la distinción de que los árboles son entrenados de manera secuencial. Esto implica que cada árbol utiliza información proveniente de árboles desarrollados previamente. A diferencia del muestreo de arranque utilizado en *bagging*, el *boosting* no recurre a esta técnica; en cambio, cada árbol se ajusta a una versión modificada del conjunto de datos original. La lógica subyacente en esta aproximación radica en evitar el sobreajuste inherente a ajustar un único árbol de decisión grande y, en su lugar, adoptar un enfoque de aprendizaje más gradual (James et al., 2023).

La metodología de *boosting* aprende de manera iterativa. Dado el modelo actual, ajustamos un árbol de decisión a los residuos del modelo existente. Este enfoque tiene la ventaja de focalizarse en las áreas donde el modelo actual presenta deficiencias, permitiendo así que cada nuevo árbol se centre en corregir las deficiencias identificadas por los modelos previos.

En otras palabras, el procedimiento consiste en ajustar un árbol utilizando los residuos actuales en lugar del resultado Y como respuesta. Posteriormente, se incorpora este nuevo árbol de decisión a la función ajustada para actualizar los residuos. Cada uno de estos árboles puede ser de tamaño reducido, con un número limitado de nodos

terminales, determinado por los parámetros en el algoritmo. Al ajustar árboles pequeños a los residuos, se mejora gradualmente la estimación $\hat{f}$ en las regiones donde no funcionaba adecuadamente.

Esta variante de modelo de Árboles incluye el parámetro conocido como λ (tasa de aprendizaje) que ralentiza aún más el proceso de aprendizaje, permitiendo que más árboles, y de formas diversas, aborden los residuos. En términos generales, los enfoques de aprendizaje estadístico que adoptan un ritmo lento de aprendizaje tienden a exhibir un buen desempeño (James et al., 2023).

3.5.b.3.a Parámetros de un modelo de boosting

Los hiperparámetros son parámetros externos al modelo de *machine learning* que no se aprenden durante el entrenamiento del modelo, pero que influyen en su comportamiento y rendimiento. A diferencia de los parámetros del modelo, que se estiman a partir de los datos durante el proceso de entrenamiento, los hiperparámetros se establecen antes de iniciar el proceso de entrenamiento y afectan cómo se ajusta y generaliza el modelo.

Los hiperparámetros son esenciales para optimizar el rendimiento del modelo y pueden tener un impacto significativo en la capacidad del modelo para aprender patrones útiles en los datos y evitar el sobreajuste. Algunos ejemplos comunes de hiperparámetros incluyen la tasa de aprendizaje en algoritmos de aprendizaje supervisado, el número de árboles en un bosque aleatorio, la profundidad máxima de un árbol de decisión, el número de vecinos en el algoritmo k-NN, entre otros.

A continuación serán mencionados los hiperparámetros de un modelo de boosting.

Máxima profundidad: Este parámetro controla la profundidad máxima posible de cada árbol de decisión en el modelo. A mayor profundidad el modelo captura relaciones más

complejas en los datos de entrenamiento, pero también puede aumentar el riesgo de sobreajuste.

Tasa de aprendizaje: Es la tasa de aprendizaje que controla la contribución de cada árbol al modelo. Una tasa de aprendizaje más baja significa que el modelo se ajustará más lentamente a los datos, lo que a menudo conduce a un mejor rendimiento del modelo, pero a costa de un mayor tiempo de entrenamiento.

Cantidad de estimadores: Es el número de árboles de decisión que se van a construir en el modelo. Un mayor número de estimadores generalmente mejora el rendimiento del modelo, pero también puede aumentar el tiempo de entrenamiento.

Peso mínimo de nodo: Este parámetro controla la cantidad mínima de instancias que deben existir en cada nodo hoja de un árbol. Aumentar este valor puede ayudar a prevenir el sobreajuste al evitar la partición de nodos que contienen muy pocas instancias.

Gama: Es un parámetro de regularización que controla cuánto se necesita reducir la función de pérdida para hacer una nueva partición en el árbol. Un valor más alto de gamma implica una mayor regularización.

Submuestra: Es la fracción de muestras que se utilizará para entrenar cada árbol. Un valor menor de subsample puede ayudar a prevenir el sobreajuste al introducir más variabilidad en el proceso de entrenamiento.

Muestreo de columnas por arbol: Es la fracción de columnas (características) que se utilizarán para entrenar cada árbol. Similar a subsample, esto introduce más variabilidad en el entrenamiento y puede ayudar a prevenir el sobreajuste.

Regularizacion alfa: Es un parámetro de regularización que penaliza los valores extremos de los pesos de las características. Ayuda a prevenir el sobreajuste al forzar que los pesos de las características sean más dispersos.

Regularizacion lambda: Es un parámetro de regularización que penaliza los valores extremos de los pesos de las características. Al igual que regularizacion alfa, ayuda a prevenir el sobreajuste al penalizar los pesos de las características más grandes.

3.5.b.3.b Optimización de hiperparámetros

La optimización de hiperparámetros juega un papel fundamental en el desarrollo de modelos de aprendizaje automático, ya que estas configuraciones impactan directamente en el rendimiento y generalización de los modelos. Sin embargo, la búsqueda exhaustiva de los mejores hiperparámetros puede ser computacionalmente costosa, especialmente en conjuntos de datos grandes o en modelos complejos. En este contexto, surge la pregunta de si métodos más simples, como la búsqueda aleatoria, pueden ofrecer resultados comparables en términos de rendimiento del modelo con un menor costo computacional.

En la literatura, se han propuesto diversos enfoques para la optimización de hiperparámetros, incluyendo técnicas avanzadas como la optimización bayesiana. Sin embargo, vamos a aplicar las técnicas de "Random Search for Hyper-Parameter Optimization" de Bergstra y Bengio (2012). Estos autores destacan la eficacia de la búsqueda aleatoria como una alternativa simple pero efectiva. Demuestran experimentalmente que la búsqueda aleatoria puede encontrar configuraciones de hiperparámetros competitivas con métodos más sofisticados, pero a un costo computacional mucho menor (Bengio & Bergstra, 2012).

3.6 Optimización y Evaluación de Modelos de Clasificación Binaria

Los modelos diseñados para predecir una clase binaria en un problema de este tipo generan un valor de salida situado en el intervalo entre 0 y 1. Esto se aplica tanto a la regresión logística como a los modelos que exploraremos más adelante. Una técnica

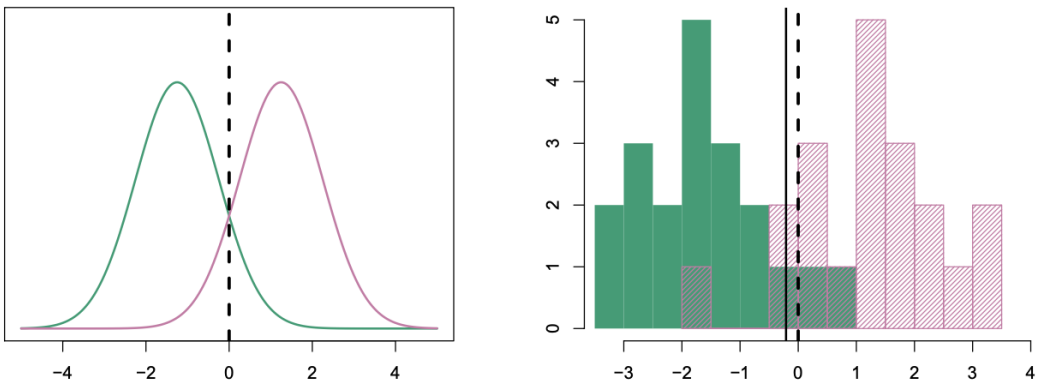
para evaluar el rendimiento de estos modelos es mediante el análisis discriminante lineal, considerando un solo predictor, que en este caso sería la variable de salida del modelo.

El análisis discriminante lineal se centra en examinar la distribución de una variable cuantitativa, y puede emplearse para evaluar el desempeño de un modelo de clasificación. Se supone que $f_k(x)$ sigue una distribución normal o gaussiana, siendo k la cantidad de clases a predecir. En el entorno unidimensional, la función de densidad normal se expresa de la siguiente manera (James et al., 2023).

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right) \quad (14)$$

donde μ_k y σ_k^2 son los parámetros de media y varianza para la k-ésima clase. Por ahora, supongamos además que $\sigma_1^2 = \cdots = \sigma_k^2$: es decir, existe un término de varianza compartido entre todas las K clases, que por simplicidad podemos denotar como σ^2 (James et al., 2023).

Figura 5) : Comparación de límites de decisión en clasificación con funciones de densidad normales y LDA.



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4

En la figura 5), el gráfico de la izquierda se muestran dos funciones de densidad normal unidimensionales, una para cada clase. Estas funciones representan cómo están distribuidos los datos de cada clase en términos de una única variable. La línea vertical punteada representa el límite de decisión. Este límite separa las dos clases basándose en la máxima probabilidad de clasificación correcta según las distribuciones de las clases y la matriz de costos asociada. El gráfico de la derecha, es conceptualmente lo mismo, con la diferencia que muestra la distribución en forma de histograma.

En el contexto de un problema de clasificación binaria, se busca predecir dos sucesos diferentes, y cada observación puede pertenecer a una de estas dos clases.
Figura 6) matriz de confusión.

		Clase real	
		Negativos	Positivos
Clase Predicha	Negativos	Verdadero Negativo (VN)	Falso Negativo (FN)
	Positivos	Falso Positivo (FP)	Verdadero Positivo (VP)

Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4

La figura 6) muestra conceptualmente que es una matriz de confusion. Esta herramienta es utilizada en el campo de la clasificación de datos para evaluar el rendimiento de un modelo predictivo. Esta matriz organiza las predicciones de un modelo en relación con los resultados reales en forma de una tabla, facilitando la comprensión de cómo el modelo está realizando sus predicciones. Esto da lugar a cuatro posibles situaciones:
A. Verdadero positivo (VP): Se predice correctamente el evento de interés (por ejemplo, *default*) y la observación efectivamente pertenece a esa clase.
B. Falso positivo (FP): Se predice incorrectamente el evento de interés (por ejemplo, *default*) cuando la observación pertenece a la otra clase (por ejemplo, repago).
C. Verdadero negativo (VN): Se predice correctamente que la observación pertenece a la clase opuesta al evento de interés (por ejemplo, repago).
D. Falso negativo (FN): Se predice incorrectamente que la observación pertenece a la clase opuesta al evento de interés cuando en realidad pertenece a la clase de interés.

Tabla 1) Simulación de 10,000 casos para clasificación en matriz de confusion.

		Clase real		
		Repagado	*Default*	Total
Clase predicha	Repagado	9,644	252	9,896
	Default	23	81	104
	Total	9,667	333	10,000

Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4.

Tabla 1) para las 10.000 observaciones de entrenamiento en el conjunto de datos predeterminado. Los elementos en la diagonal de la matriz representan individuos cuyos estados predeterminados se predijeron correctamente, mientras que los elementos fuera de la diagonal representan individuos que fueron clasificados erróneamente. análisis discriminante lineal hizo predicciones incorrectas para 23 personas que no incumplieron y para 252 personas que sí lo hicieron.

Los datos de la matriz de confusión ofrecen una forma clara de comprender esta información. La tabla revela que, según el análisis de discriminación lineal, se predijo que un total de 104 personas incumplirían sus pagos. De estas personas, 81 realmente incumplieron y 23 no lo hicieron. Por otro lado, se asignaron 252 observaciones al grupo de "no incumplimiento", de un total de 9,896 casos, lo que equivale a un 2.5% de incumplimiento (James et al., 2023).

Tabla 2) Simulción de 10,000 con diferente límite en la clasificación.

		Clase real		
		Repagado	*Default*	Total
Clase predicha	Repagado	9,432	138	9,570
	Default	235	195	430
	Total	9,667	333	10,000

Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4.

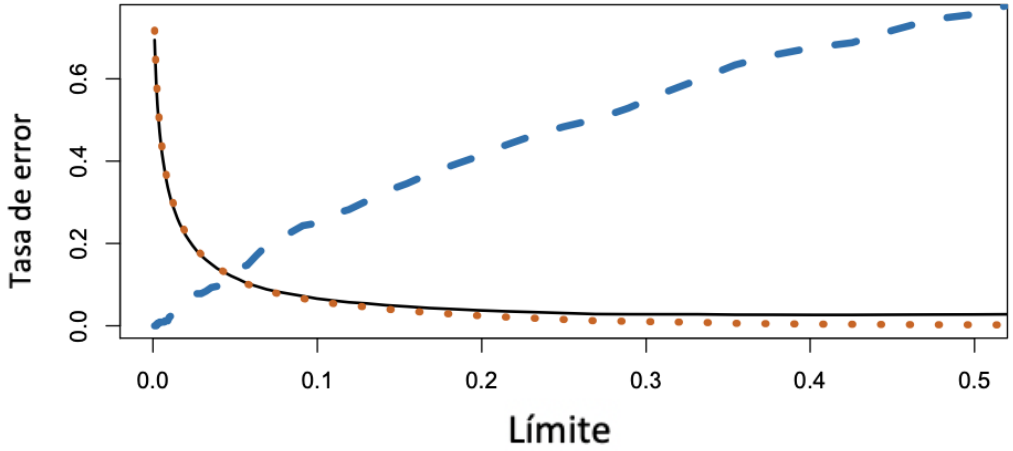
En la tabla 2) siguiendo el mismo ejemplo de la tabla 1) si modificamos la definición de predicción de incumplimiento (es decir, considerar como incumplimiento aquellos casos con una probabilidad del 20% o más en lugar del 50% o más), obtendríamos una matriz de confusión diferente para el mismo modelo sin que su rendimiento se vea alterado.

En la matriz de confusión, al utilizar un umbral más exigente, observamos una disminución en los casos del grupo predicho como buenos pagadores que terminaron en default, es decir, los falsos negativos, que ahora son 138 casos. Mientras tanto, el total de casos del grupo de buenos pagadores es de 9,570. Por lo tanto, la tasa de default ha disminuido al 1.4% para el mismo resultado del modelo.

Podemos deducir que, dado el resultado de un modelo de clasificación binaria con estas características, existen diferentes puntos de corte y una tasa de error o tasa de *default* asociada a cada uno. La elección del umbral y la tasa de error más apropiados dependerá del problema específico y de la estrategia de negocio.

Podemos entonces definir la métrica de "tasa de precisión" como el ratio entre los positivos predichos que resultan ser positivos (verdaderos positivos), y los positivos totales (independientemente de la predicción). Esta métrica proporciona la proporción de predicciones positivas que son correctas. Y podemos en el siguiente gráfico identificarlo con la linea continua negra con puntos naranjas. Y al *recall* como la tasa de vedaderos positivo sobre todos los positivos predichos, a continuación en el gráfico representado con una linea discontinuada azul.

Figura 7) Tasas de error en función del umbral de probabilidad posterior en el conjunto de datos Default



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4.

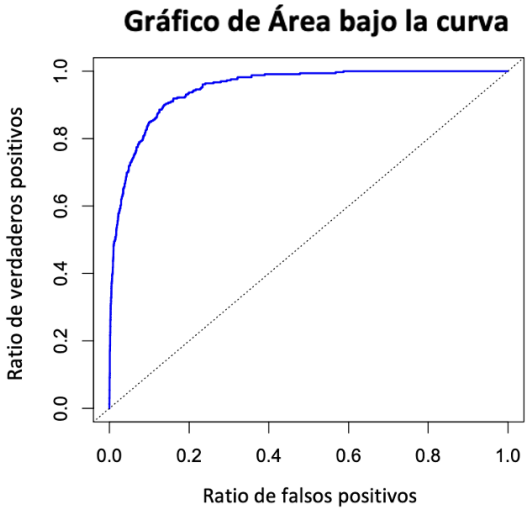
Figura 7) Para el conjunto de datos, las tasas de error se muestran como una función del valor umbral para la probabilidad posterior que se utiliza para realizar la asignación. La línea continua negra con puntos naranjas muestra la tasa de error general. La línea discontinua azul representa la fracción de clientes morosos que están clasificados incorrectamente, y la línea punteada naranja indica la fracción de errores entre los clientes no morosos (James et al., 2023).

Existe un *trade-off* inherente entre precisión y recuperación. A menudo, mejorar la precisión de un modelo puede resultar en una disminución de la recuperación y viceversa. Por ejemplo, si aumentamos el umbral de decisión para clasificar una instancia como positiva, es probable que mejoremos la precisión del modelo, ya que estaremos más seguros de que las instancias clasificadas como positivas son verdaderamente positivas. Sin embargo, esto podría resultar en una disminución de la recuperación, ya que el modelo podría perder algunos casos positivos al ser más selectivo. Por otro lado, si reducimos el umbral de decisión para clasificar más instancias como positivas, es probable que mejoremos la recuperación del modelo, ya que capturaremos más casos positivos. Sin embargo, esto podría resultar en una disminución de la precisión, ya que también aumentaría el número de falsos positivos (Davis J, Goadrich M. 2006).

3.6.a Curva ROC o AUC.

La curva *ROC* (*Receiver Operating Characteristic*), también conocida como AUC (Área Bajo la Curva), es un indicador utilizado para representar simultáneamente los dos tipos de errores para todos los umbrales posibles en un modelo de clasificación binaria (James et al., 2023).

Figura 8) Curva ROC para el clasificador análisis discriminante lineal.



Fuente: (James et al., 2023). *An Introduction to Statistical Learning with Applications in python.* New York: Springer. Capítulo 4.

La figura 8) muestra la curva ROC para el clasificador del análisis lineal discriminante. La curva muestra dos tipos de error a medida que se varía el valor umbral para la probabilidad *default*.

El rendimiento global de un clasificador, resumido sobre todos los umbrales posibles, se evalúa mediante el Área Bajo la Curva (AUC) de la curva ROC. Una curva ROC ideal se ubicaría en la esquina superior izquierda, y, por lo tanto, un AUC más cercano a 1 indica un mejor rendimiento del clasificador. En este caso, el AUC es de 0.95, lo cual se considera muy bueno y está cerca del máximo teórico de 1. Un clasificador que funcione al azar tendría un AUC de 0.5 al ser evaluado en un conjunto de pruebas independientes que no se utilizaron en el entrenamiento del modelo. Las curvas ROC son valiosas para comparar diferentes clasificadores, ya que consideran todos los umbrales posibles (James et al., 2023).

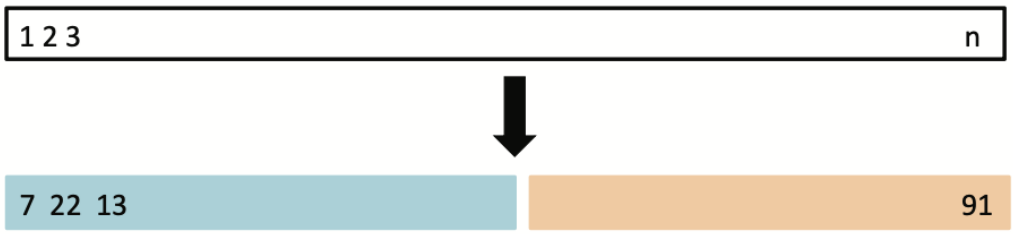
El Área Bajo la Curva (AUC) será empleada como una métrica fundamental en esta tesis para comparar el rendimiento de las técnicas desarrolladas y los modelos presentados.

3.7 *Cross Validation*

En lo que respecta a la evaluación del rendimiento de un modelo, es crucial tener en cuenta que los datos utilizados para el entrenamiento difieren de los datos futuros que el modelo encontrará en su uso prospectivo. Es fundamental destacar que calcular el error del modelo basándose únicamente en los datos de entrenamiento puede ser engañoso y no proporcionar una medida precisa de la eficacia del modelo en situaciones del mundo real.

La estrategia de *cross-validation* implica dividir de manera aleatoria el conjunto de datos disponible en dos partes: un conjunto de entrenamiento y un conjunto de validación o prueba. Inicialmente, el modelo se ajusta utilizando el conjunto de entrenamiento y luego se emplea para predecir las respuestas de las observaciones en el conjunto de validación. La tasa de error resultante en el conjunto de validación se evalúa comúnmente mediante métricas como el Área Bajo la Curva (AUC), especialmente aplicable a problemas de la naturaleza que estamos abordando. Este enfoque ayuda a evaluar el rendimiento del modelo en datos no utilizados durante el entrenamiento, proporcionando una medida más realista de su capacidad predictiva.

Figura 9) Esquema del enfoque de conjunto de validación en *cross validation*



Fuente: (James et al., 2023). An Introduction to Statistical Learning with Applications in python. New York: Springer. Capítulo 5.

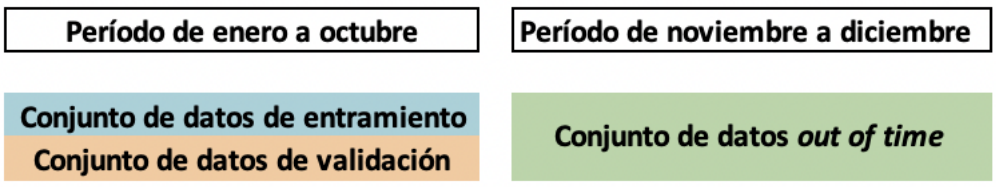
La figura 9) Muestra la estrategia de conjunto de validación implica dividir un conjunto de n observaciones en dos conjuntos aleatorios: un conjunto de entrenamiento (destacado en azul, que incluye las observaciones 7, 22 y 13, entre otras) y un conjunto de validación (resaltado en beige, que contiene la observación 91, entre otras).

El modelo de aprendizaje estadístico se ajusta utilizando el conjunto de entrenamiento, y su rendimiento se evalúa posteriormente en el conjunto de validación. Este enfoque proporciona una evaluación más precisa del rendimiento del modelo en datos no utilizados durante el proceso de entrenamiento (James et al., 2023).

3.7.a Validación *Out-of-time*

La distinción entre el conjunto de datos de entrenamiento y el conjunto de prueba se realiza de manera aleatoria, como se mencionó anteriormente. No obstante, también es crucial evaluar el rendimiento del modelo en datos que no solo son diferentes aleatoriamente, sino que también pertenecen a un momento temporal posterior. Esto se debe a que, a pesar de la estrategia de validación cruzada, pueden capturarse y validar patrones específicos del momento en que se entrena y valida el modelo. Por esta razón, es fundamental crear un tercer conjunto de datos que represente un espacio temporal diferente, asegurando así una evaluación más completa y fiable del rendimiento del modelo (Libro Vivo de Ciencia de Datos, n.d.).

Figura 10) Esquema del enfoque de conjunto de datos de validación en *out of time*



Fuente: (Libro Vivo de Ciencia de Datos, n.d.).

La figura 10) muestra como pueden convivir las estrategias de *cross validation* con *out of time*. Mientras que *out of* time es útil para validar que el modelo puede ser utilizado en datos desconocidos.

3.8 Métodos de re-sampleo

El re-sampleo es una técnica que se puede emplear para abordar el problema del desbalance de clases. Su objetivo es generar un conjunto de datos que exhiba una distribución de clases relativamente equilibrada. Esto permite que los modelos de clasificación capturen de manera más efectiva el límite de decisión entre las clases mayoritarias y minoritarias. Aunque los métodos de muestreo son útiles para facilitar la clasificación de los casos de clases minoritarias, es esencial que el conjunto de datos muestreado represente una aproximación "razonable" del conjunto de datos original. Entre los métodos desarrollados se encuentran el *undersampling* y el *oversampling*.

3.8.a *Undersampling*

En el *undersampling*, se descartan de manera aleatoria algunas observaciones que pertenecen a la clase mayoritaria con el fin de lograr una distribución más equilibrada. Como ejemplo, consideremos un conjunto de datos que consta de 100 observaciones de la clase minoritaria y 1000 observaciones de la clase mayoritaria. En el proceso de *undersampling*, se buscaría crear una distribución de clases equilibrada seleccionando aleatoriamente 900 observaciones de la clase mayoritaria para eliminarlas. El conjunto de datos resultante constará entonces de 200 observaciones, de las cuales 100 pertenecen a la clase que originalmente era minoritaria y otras 100 son observaciones aleatorias de la clase que originalmente era mayoritaria (Haibo & Ma, 2013).

3.8.b *Oversampling*

De manera alternativa, el *oversampling* implica duplicar y repetir de manera aleatoria las observaciones de las clases minoritarias en el conjunto de datos hasta lograr una distribución más equilibrada. Considerando el ejemplo anterior con un conjunto de

datos que consta de 100 observaciones de la clase minoritaria y 1000 observaciones de la clase mayoritaria, el *oversampling* tradicional replicaría de manera aleatoria las observaciones de la clase minoritaria unas 900 veces, resultando en un nuevo conjunto de datos con 1000 observaciones de la clase que originalmente era minoritaria y manteniendo las 1000 observaciones de la clase que originalmente era mayoritaria (Haibo & Ma, 2013).

3.8.c Apreciaciones sobre los métodos de re-sampleo

Aunque los métodos de re-muestreo generan distribuciones más equilibradas para facilitar la captura efectiva del comportamiento de la clase minoritaria durante el entrenamiento del algoritmo, ambos métodos mencionados presentan desventajas. Por ejemplo, en el *undersampling*, existe el riesgo potencial de descartar grandes cantidades de datos. En el ejemplo presentado, se eliminó el 90% de los datos, lo que podría dificultar el aprendizaje del límite de decisión entre las clases minoritarias y mayoritarias, resultando en una disminución del rendimiento de la clasificación (Haibo & Ma, 2013).

El inconveniente asociado con el *oversampling* radica en la repetición excesiva de observaciones, a veces en grados significativos. Tome, por ejemplo, el escenario de *oversampling* aleatorio previamente mencionado, donde cada instancia tuvo que replicarse 9 veces para equilibrar la distribución de clases. Al duplicar instancias de esta manera, existe el riesgo de inducir un sobreajuste en los datos de entrenamiento, lo que puede resultar en una disminución del rendimiento del modelo clasificador (Haibo & Ma, 2013).

Una estrategia que busca mitigar los inconvenientes asociados con el re-muestreo, mientras aprovecha los beneficios de ambas técnicas, consiste en utilizar una aproximación híbrida que combine tanto *undersampling* como *oversampling* simultáneamente. En el ejemplo mencionado anteriormente, de las 1000 observaciones, podríamos optar por descartar aleatoriamente 450 observaciones de la

clase mayoritaria y, al mismo tiempo, replicar 450 observaciones de la clase minoritaria. Este enfoque resultaría en un conjunto de datos equilibrado de manera óptima.

4. Análisis descriptivo

Como fue mencionado anteriormente, en la presente conjunto de datos una observación representa el estado de las variables que serán explicadas a continuación de una empresa exportadora, al momento en que la misma cedió una factura a la empresa *fintech*, con información de si ese crédito fue repagado o no.

4.a. Definición de variables predictoras

NO_SALES_DAYS: Cantidad de días que pasaron desde la última vez que la empresa exportadora facturó, al momento de ceder a la empresa fintech. Esta última no cuenta en dicha definición.

DAYS_TO_DUE: Cantidad de días restantes hasta el vencimiento de la factura que la empresa exportadora adelanta. En caso de operar facturas vencidas, este campo toma valores negativos.

ANOS_ACTIVIDAD_EXP: Antigüedad en años trabajando con exportaciones. Por cuestiones operativas de costo a la hora de extraer datos, la empresa tomó la decisión de no extraer datos más antiguos que 5 años. Por lo tanto, esta variable no toma valores mayores a 5 años.

ANOS_ACTIVIDAD_DOM: Antigüedad en años trabajando en el mercado local. Por cuestiones operativas de costo a la hora de extraer datos, la empresa tomó la decisión de no extraer datos más antiguos que 5 años. Por lo tanto, esta variable no toma valores mayores a 5 años.

FIRST_CREDIT_FLAG: Variable binarizada, vale 1 si es la primera vez que la empresa pide el crédito, y 0 en caso contrario.

VENTAS_DOM_MX: Donde X puede variar de 1 a 18. Ventas domésticas de la empresa exportadora del mes X anterior considerando la fecha del adelanto. Ejemplo: para el mes "1" si la empresa pide el adelanto el 15 de abril, se consideran las ventas desde el período del 15 de marzo hasta el 14 de abril.

VENTAS_DOM_MIN_LAST12: Ventas mínimas del período de los últimos 12 meses.

VENTAS_DOM_MAX_LAST12: Ventas máximas del período de los últimos 12 meses.

VENTAS_DOM_AVG_LAST12: Ventas promedio del período de los últimos 12 meses.

VENTAS_DOM_AVG_SMOOTHED_LAST12: Ventas promedio del período de los últimos 12 meses, quitando los valores mínimos y máximos, se promedian los 10 meses restantes.

DESV_STAND_VENTAS_DOM: Desvío estándar del período de los últimos 12 meses.

COEF_VAR_VENTAS_DOM: Coeficiente de variación de las ventas domésticas mensuales del período de los últimos 12 meses.

FLAG_VENTAS_DOM_MX: Donde X puede variar de 1 a 18. Binarización de la variable VENTAS_DOM_MX.

Q_MONTHS_DOM_SALES_LAST3: Cantidad de meses que la empresa exportadora tuvo ventas domésticas en el período de los últimos 3 meses.

Q_MONTHS_DOM_SALES_LAST6: Cantidad de meses que la empresa exportadora tuvo ventas domésticas en el período de los últimos 6 meses.

Q_MONTHS_DOM_SALES_LAST12: Cantidad de meses que la empresa exportadora tuvo ventas domésticas en el período de los últimos 12 meses.

RATIO_CANCELACIONES_DOM: Ratio del monto de facturas de ventas domésticas canceladas sobre facturas de ventas domésticas canceladas y válidas.

EXPORTACIONES_MX: Donde X puede variar de 1 a 18. Exportaciones de la empresa exportadora del mes X anterior considerando la fecha del adelanto.

EXPORTACIONES_MIN_LAST12: Exportaciones mínimas del período de los últimos 12 meses.

EXPORTACIONES_MAX_LAST12: Exportaciones máximas del período de los últimos 12 meses.

EXPORTACIONES_AVG_LAST12: Exportaciones promedio del período de los últimos 12 meses.

EXPORTACIONES_AVG_SMOOTHED_LAST12: Exportaciones promedio del período de los últimos 12 meses, quitando los valores mínimos y máximos, se promedian los 10 meses restantes.

COEF_VAR_EXPORTACIONES: Coeficiente de variación de las exportaciones mensuales del período de los últimos 12 meses, quitando los valores mínimos y máximos, se promedian los 10 meses restantes.

DESV_STAND_EXPORTACIONES: Desvío estándar de exportaciones del período de los últimos 12 meses.

FLAG_EXPORTACIONES_ MX: Donde X puede variar de 1 a 18. Binarización de la variable EXPORTACIONES_MX.

Q_MONTHS_EXPORTS_LAST3: Cantidad de meses que la empresa exportadora tuvo exportaciones en el período de los últimos 3 meses.

Q_MONTHS_EXPORTS_LAST6: Cantidad de meses que la empresa exportadora tuvo exportaciones en el período de los últimos 6 meses.

Q_MONTHS_EXPORTS_LAST12: Cantidad de meses que la empresa exportadora tuvo exportaciones en el período de los últimos 12 meses.

VENTAS_TOT_MX: Donde X puede variar de 1 a 18. Ventas domésticas y exportaciones de la empresa exportadora del mes X anterior, considerando la fecha del adelanto. Por ejemplo para el mes "1", si la empresa solicita el adelanto el 15 de abril, se consideran las ventas desde el período del 15 de marzo hasta el 14 de abril.

VENTAS_TOTALES_MIN_LAST12: Ventas totales mínimas del período de los últimos 12 meses.

VENTAS_TOTALES_MAX_LAST12: Ventas totales máximas del período de los últimos 12 meses.

VENTAS_TOT_AVG_LAST12: Ventas totales promedio del período de los últimos 12 meses.

VENTAS_TOT_AVG_SMOOTHED_LAST12: Ventas totales promedio del período de los últimos 12 meses, excluyendo los valores mínimos y máximos; se promedian los 10 meses restantes.

COEF_VAR_VENTAS_TOT: Coeficiente de variación de las ventas totales mensuales del período de los últimos 12 meses, excluyendo los valores mínimos y máximos; se promedian los 10 meses restantes.

DESV_STAND_VENTAS_TOT: Desviación estándar de las ventas totales del período de los últimos 12 meses.

FLAG_VENTAS_TOTAL_ MX: Donde X puede variar de 1 a 18. Binarización de la variable VENTAS_TOT_MX.

Q_MONTHS_VENTAS_TOTAL_LAST3: Cantidad de meses en los cuales la empresa exportadora ha tenido ventas domésticas o exportaciones en los últimos 3 meses.

Q_MONTHS_VENTAS_TOTAL_LAST6: Cantidad de meses en los cuales la empresa exportadora ha tenido ventas domésticas o exportaciones en los últimos 6 meses.

Q_MONTHS_VENTAS_TOTAL_LAST12: Cantidad de meses en los cuales la empresa exportadora ha tenido ventas domésticas o exportaciones en los últimos 12 meses.

SHARE_EXPORTACIONES_MX: Donde X puede variar de 1 a 18. Porcentaje de representación de las exportaciones del mes anterior sobre las ventas totales en dicho período.

RATIO_CANCELACIONES_EXP: Ratio del monto de facturas de exportación canceladas sobre facturas de exportación canceladas y válidas.

RATIO_EXPORTACIONES_SOBRE_VENTAS: Ratio de exportaciones sobre ventas domésticas en el período de 12 meses.

SHARE_VENTAS_DOM_MX: Donde X puede variar de 1 a 18. Porcentaje de representación de las ventas domésticas del mes anterior sobre las ventas totales en dicho período.

HHI_PROVIDERS_L3M: Índice de Herfindahl-Hirschman de las compras, medido en los últimos 3 meses.

HHI_PROVIDERS_L6M: Índice de Herfindahl-Hirschman de las compras, medido en los últimos 6 meses.

HHI_PROVIDERS_L9M: Índice de Herfindahl-Hirschman de las compras, medido en los últimos 9 meses.

HHI_PROVIDERS_L12M: Índice de Herfindahl-Hirschman de las compras, medido en los últimos 12 meses.

HHI_CLIENTES_L3M: Índice de Herfindahl-Hirschman de las ventas, medido en los últimos 3 meses.

HHI_CLIENTES_L6M: Índice de Herfindahl-Hirschman de las ventas, medido en los últimos 6 meses.

HHI_CLIENTES_L9M: Índice de Herfindahl-Hirschman de las ventas, medido en los últimos 9 meses.

HHI_CLIENTES_L12M: Índice de Herfindahl-Hirschman de las ventas, medido en los últimos 12 meses.

SEGMENTO_INDUSTRIA: Variable categórica que indica si la empresa se dedica principalmente a la comercialización de productos perecederos o no perecederos.

Categoría *k-means*: Variable categórica del cluster no ordinal a la que pertenece la observación. Toma valores discretos en el dominio {1, 5}.

EXPORTER_SIZE: Variable categórica que clasifica a las empresas exportadoras en {CORP, LARGE, MEDIUM, SMALL, OCASIONAL} según el volumen y la consecutividad de exportaciones.

Tabla 3) Descripción de las variables númericas.

Descripción	Promedio	Desvío estandar	Mínimo	Primer cuantil	Mediana	Tercer cuantil	Máximo
DIAS SIN FACTURACIÓN	6	54	1	1	1	5	574
DAYS_TO_DUE	75	70	-153	28	57	112	349
ANOS_ACTIVIDAD_EXP	4	1	0	4	4	4	5
ANOS_ACTIVIDAD_DOM	4	1	0	4	4	4	5
FIRST_CREDIT_FLAG	0	0	0	0	0	0	1
VENTAS_DOM_M1	9,596,793	15,601,220	0	88,057	1,577,472	5,908,216	43,101,431
VENTAS_DOM_M2	9,438,704	15,479,842	0	99,934	1,445,490	5,661,707	44,299,365
VENTAS_DOM_M3	9,449,068	15,576,369	0	100,626	1,470,218	5,349,097	45,016,251
VENTAS_DOM_M4	9,627,781	15,772,221	0	95,061	1,490,529	5,797,866	77,462,717
VENTAS_DOM_M5	9,156,093	15,178,810	0	93,407	1,420,567	5,168,695	44,428,569
VENTAS_DOM_M6	8,730,015	14,262,333	0	91,065	1,429,983	5,399,742	40,203,857
VENTAS_DOM_M7	8,976,810	14,670,598	0	92,751	1,433,484	5,153,226	40,629,798
VENTAS_DOM_M8	8,673,720	14,196,716	0	81,415	1,512,009	5,497,706	40,476,719
VENTAS_DOM_M9	8,543,915	14,023,411	0	85,744	1,421,968	4,878,288	38,890,208
VENTAS_DOM_M10	8,592,903	14,195,213	0	83,463	1,491,148	4,912,374	39,841,717
VENTAS_DOM_M11	8,632,281	14,131,282	0	85,438	1,523,502	4,907,680	40,635,134
VENTAS_DOM_M12	8,543,151	14,049,359	0	73,665	1,370,325	4,885,303	38,755,798
VENTAS_DOM_M13	8,226,132	13,500,224	0	79,912	1,492,283	4,588,757	38,756,507
VENTAS_DOM_M14	8,362,040	13,723,774	0	81,940	1,507,936	4,758,259	37,717,850
VENTAS_DOM_M15	8,414,076	13,957,847	0	77,074	1,275,867	4,581,668	38,151,306
VENTAS_DOM_M16	8,025,883	13,353,281	0	71,613	1,213,712	4,338,301	36,333,073
VENTAS_DOM_M17	7,895,633	13,134,539	0	75,452	1,196,791	4,326,340	44,996,222
VENTAS_DOM_M18	7,982,508	13,326,140	0	89,804	1,201,912	4,174,326	45,653,234
VENTAS_DOM_MIN_LAST12	7,108,953	12,072,356	0	35,852	914,300	3,821,721	33,241,673

VENTAS_DOM_MAX_LAST12	11,012,941	16,956,988	0	194,070	2,603,490	6,447,549	77,462,717
VENTAS_DOM_AVG_LAST12	8,996,770	14,669,264	0	89,713	1,730,339	5,218,201	37,989,108
VENTAS_DOM_AVG_SMOOTHED_LAST12	8,983,934	14,733,672	0	87,152	1,711,397	5,140,188	38,576,648
DESV_STAND_VENTAS_DOM	3,457,080	5,044,046	0	144,715	1,123,207	4,748,113	68,308,531
COEF_VAR_VENTAS_DOM	1.75	4.68	0.00	0.39	0.85	1.43	289
FLAG_VENTAS_DOM_M1	0.9925	0.0862	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M2	0.9914	0.0923	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M3	0.9911	0.0937	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M4	0.9907	0.0961	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M5	0.9910	0.0942	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M6	0.9910	0.0942	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M7	0.9878	0.1100	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M8	0.9838	0.1261	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M9	0.9779	0.1470	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M10	0.9758	0.1537	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M11	0.9759	0.1534	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M12	0.9767	0.1508	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M13	0.9754	0.1548	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M14	0.9738	0.1598	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M15	0.9706	0.1690	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M16	0.9676	0.1772	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M17	0.9662	0.1807	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_DOM_M18	0.9635	0.1876	0.0000	1.0000	1.0000	1.0000	1.0000
Q_MONTHS_DOM_SALES_LAST3	2.8800	0.5624	0.0000	3.0000	3.0000	3.0000	3.0000
Q_MONTHS_DOM_SALES_LAST6	5.7489	1.1144	0.0000	6.0000	6.0000	6.0000	6.0000
Q_MONTHS_DOM_SALES_LAST12	11.4001	2.2075	0.0000	12.0000	12.0000	12.0000	12.0000
RATIO_CANCELACIONES_DOM	0.0937	0.1070	0.0000	0.0300	0.0500	0.1400	0.0937

EXPORTACIONES_M1	1,004,356	1,652,059	0	155,243	250,354	1,421,724	14,709,300
EXPORTACIONES_M2	958,490	1,649,911	0	135,800	244,887	1,305,800	15,054,608
EXPORTACIONES_M3	918,674	1,479,797	0	145,531	230,789	1,201,630	9,544,281
EXPORTACIONES_M4	921,396	1,466,723	0	135,859	252,198	1,222,477	9,669,921
EXPORTACIONES_M5	920,558	1,548,224	0	135,800	258,221	1,194,903	10,090,273
EXPORTACIONES_M6	912,975	1,557,773	0	130,425	236,526	1,049,140	9,544,281
EXPORTACIONES_M7	877,176	1,533,523	0	126,002	234,315	988,212	10,090,273
EXPORTACIONES_M8	820,571	1,442,034	0	127,192	241,693	864,653	9,544,281
EXPORTACIONES_M9	758,367	1,288,565	0	105,633	250,150	771,950	9,386,070
EXPORTACIONES_M10	737,827	1,311,050	0	102,780	279,261	720,941	10,142,608
EXPORTACIONES_M11	696,886	1,195,887	0	99,584	268,130	779,701	9,568,612
EXPORTACIONES_M12	650,509	1,052,197	0	88,173	276,814	777,170	9,568,612
EXPORTACIONES_M13	589,844	974,498	0	90,147	236,808	757,562	7,863,437
EXPORTACIONES_M14	547,181	860,440	0	68,781	220,593	767,383	6,752,846
EXPORTACIONES_M15	488,946	769,006	0	50,901	205,862	725,412	6,211,166
EXPORTACIONES_M16	453,092	683,158	0	51,428	205,470	670,661	6,155,934
EXPORTACIONES_M17	469,209	725,308	0	43,949	228,993	667,120	6,155,934
EXPORTACIONES_M18	455,954	729,267	0	45,586	197,720	641,278	6,254,592
EXPORTACIONES_MIN_LAST12	383,272	686,063	0	20,534	131,355	349,345	4,432,483
EXPORTACIONES_MAX_LAST12	1,657,242	2,546,124	0	328,357	593,432	2,005,234	15,054,608
EXPORTACIONES_AVG_LAST12	848,149	1,281,861	0	173,506	247,821	1,085,485	7,228,648
EXPORTACIONES_AVG_SMOOTHED_LAST12	813,727	1,237,315	0	167,958	231,547	1,026,463	6,842,369
COEF_VAR_EXPORTACIONES	2.11	3.81	0.00	1.00	1.70	2.32	151
DESV_STAND_EXPORTACIONES	1,165,244	1,860,925	0	226,083	434,758	906,940	10,441,462
FLAG_EXPORTACIONES_M1	0.0004	0.0191	0.0000	0.0000	0.0000	0.0000	1.0000
FLAG_EXPORTACIONES_M2	0.9804	0.1388	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M3	0.9814	0.1352	0.0000	1.0000	1.0000	1.0000	1.0000

FLAG_EXPORTACIONES_M4	0.9712	0.1672	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M5	0.9724	0.1638	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M6	0.9699	0.1708	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M7	0.9721	0.1646	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M8	0.9630	0.1888	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M9	0.9560	0.2052	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M10	0.9365	0.2439	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M11	0.9406	0.2363	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M12	0.9365	0.2439	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M13	0.9217	0.2686	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M14	0.9100	0.2862	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M15	0.8939	0.3079	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M16	0.8799	0.3251	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M17	0.8707	0.3355	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_EXPORTACIONES_M18	0.8523	0.3548	0.0000	1.0000	1.0000	1.0000	1.0000
Q_MONTHS_EXPORTS_LAST3	2.9329	0.3547	0.0000	3.0000	3.0000	3.0000	3.0000
Q_MONTHS_EXPORTS_LAST6	5.8474	0.6929	0.0000	6.0000	6.0000	6.0000	6.0000
Q_MONTHS_EXPORTS_LAST12	11.5017	1.5548	0.0000	12.0000	12.0000	12.0000	12.0000
VENTAS_TOT_M1	10,601,150	15,325,604	0	436,505	2,962,331	9,169,228	43,261,770
VENTAS_TOT_M2	10,397,194	15,212,740	0	415,842	2,971,988	8,808,897	44,418,239
VENTAS_TOT_M3	10,367,742	15,316,869	0	367,305	2,834,793	8,358,208	45,147,606
VENTAS_TOT_M4	10,549,177	15,518,214	0	370,234	2,896,081	8,607,685	77,462,717
VENTAS_TOT_M5	10,076,652	14,931,409	0	349,759	2,585,597	7,758,409	44,561,299
VENTAS_TOT_M6	9,642,990	14,037,817	0	334,528	2,682,524	8,405,216	40,371,710
VENTAS_TOT_M7	9,853,986	14,457,060	0	331,358	2,834,803	7,969,033	40,864,113
VENTAS_TOT_M8	9,494,291	14,010,663	0	327,584	2,782,520	7,830,011	40,715,932
VENTAS_TOT_M9	9,302,282	13,885,229	0	318,158	2,737,869	6,921,285	39,111,893
VENTAS_TOT_M10	9,330,730	14,073,446	0	283,588	2,659,426	6,636,703	40,167,928

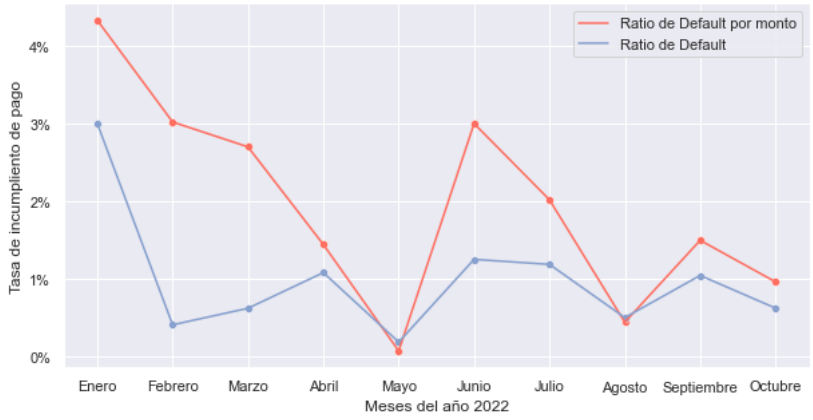
VENTAS_TOT_M11	9,329,167	14,026,157	0	312,040	2,784,473	6,344,765	41,005,599
VENTAS_TOT_M12	9,193,660	13,958,601	0	286,716	2,625,040	6,121,415	39,295,237
VENTAS_TOT_M13	8,815,976	13,383,482	0	273,182	2,399,753	5,917,530	39,076,219
VENTAS_TOT_M14	8,909,221	13,624,841	0	290,130	2,182,250	5,965,745	38,007,860
VENTAS_TOT_M15	8,903,022	13,868,037	0	270,185	2,050,848	5,605,280	38,318,168
VENTAS_TOT_M16	8,478,975	13,270,736	0	245,352	1,884,642	5,356,720	36,595,511
VENTAS_TOT_M17	8,364,842	13,061,757	0	232,380	1,884,773	5,538,220	45,313,014
VENTAS_TOT_M18	8,438,462	13,256,478	0	232,335	1,884,642	5,313,947	45,922,212
VENTAS_TOTALES_MIN_LAST12	7,651,187	11,963,947	0	201,230	1,733,202	4,651,239	33,740,934
VENTAS_TOTALES_MAX_LAST12	12,160,298	16,646,718	0	599,548	4,522,742	11,425,480	77,462,717
VENTAS_TOT_AVG_LAST12	9,844,918	14,456,670	0	356,961	2,650,868	7,635,751	38,228,446
VENTAS_TOT_AVG_SMOOTHED_LAST 12	9,832,754	14,523,457	0	346,302	2,605,598	7,422,219	38,835,624
COEF_VAR_VENTAS_TOT	1.05	5.14	0.00	0.41	0.71	1.19	526
DESV_STAND_VENTAS_TOT	4,029,934	5,017,720	0	347,260	1,732,485	7,156,953	68,308,531
FLAG_VENTAS_TOTAL_M1	0.9611	0.1934	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M2	0.9594	0.1973	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M3	0.9594	0.1973	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M4	0.9564	0.2042	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M5	0.9583	0.1998	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M6	0.9542	0.2090	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M7	0.9480	0.2220	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M8	0.9443	0.2294	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M9	0.9397	0.2380	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M10	0.9366	0.2437	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M11	0.9374	0.2422	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M12	0.9452	0.2276	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M13	0.9487	0.2205	0.0000	1.0000	1.0000	1.0000	1.0000

FLAG_VENTAS_TOTAL_M14	0.9453	0.2274	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M15	0.9457	0.2266	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M16	0.9491	0.2198	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M17	0.9461	0.2258	0.0000	1.0000	1.0000	1.0000	1.0000
FLAG_VENTAS_TOTAL_M18	0.9403	0.2370	0.0000	1.0000	1.0000	1.0000	1.0000
Q_MONTHS_VENTAS_TOTAL_LAST3	2.9751	0.2500	0.0000	3.0000	3.0000	3.0000	3.0000
Q_MONTHS_VENTAS_TOTAL_LAST6	5.9478	0.4767	0.0000	6.0000	6.0000	6.0000	6.0000
Q_MONTHS_VENTAS_TOTAL_LAST12	11.8257	1.0174	0.0000	12.0000	12.0000	12.0000	12.0000
SHARE_EXPORTACIONES_M1	0.3987	0.3482	0.0000	0.0446	0.3445	0.7406	1.0000
SHARE_EXPORTACIONES_M2	0.3909	0.3491	0.0000	0.0313	0.3197	0.7194	1.0000
SHARE_EXPORTACIONES_M3	0.3900	0.3471	0.0000	0.0281	0.3338	0.6876	1.0000
SHARE_EXPORTACIONES_M4	0.3894	0.3474	0.0000	0.0213	0.3278	0.7443	1.0000
SHARE_EXPORTACIONES_M5	0.3921	0.3468	0.0000	0.0170	0.3336	0.7286	1.0000
SHARE_EXPORTACIONES_M6	0.3901	0.3464	0.0000	0.0186	0.3266	0.7214	1.0000
SHARE_EXPORTACIONES_M7	0.3811	0.3450	0.0000	0.0156	0.3161	0.7224	1.0000
SHARE_EXPORTACIONES_M8	0.3750	0.3460	0.0000	0.0156	0.3080	0.7294	1.0000
SHARE_EXPORTACIONES_M9	0.3617	0.3460	0.0000	0.0174	0.3015	0.7102	1.0000
SHARE_EXPORTACIONES_M10	0.3541	0.3417	0.0000	0.0151	0.2753	0.6706	1.0000
SHARE_EXPORTACIONES_M11	0.3444	0.3389	0.0000	0.0152	0.2656	0.6025	1.0000
SHARE_EXPORTACIONES_M12	0.3419	0.3407	0.0000	0.0126	0.2508	0.6537	1.0000
SHARE_EXPORTACIONES_M13	0.3299	0.3342	0.0000	0.0088	0.2278	0.5978	1.0000
SHARE_EXPORTACIONES_M14	0.3198	0.3339	0.0000	0.0080	0.2205	0.5949	1.0000
SHARE_EXPORTACIONES_M15	0.3096	0.3308	0.0000	0.0073	0.1965	0.5492	1.0000
SHARE_EXPORTACIONES_M16	0.3026	0.3228	0.0000	0.0071	0.2121	0.5242	1.0000
SHARE_EXPORTACIONES_M17	0.2973	0.3196	0.0000	0.0077	0.1905	0.5241	1.0000
SHARE_EXPORTACIONES_M18	0.2904	0.3192	0.0000	0.0071	0.1721	0.5125	1.0000
RATIO_CANCELACIONES_EXP	0.0739	0.0952	0.0000	0.0000	0.0600	0.1100	0.9300
EXPORTACIONES SOBRE VENTAS DOM	1.3177	1.4302	0.0000	0.0582	0.4433	2.6855	2,135.83

SHARE_VENTAS_DOM_M1	0.6013	0.3482	0.0000	0.2594	0.6555	0.9554	1.0000
SHARE_VENTAS_DOM_M2	0.6091	0.3491	0.0000	0.2806	0.6803	0.9687	1.0000
SHARE_VENTAS_DOM_M3	0.6100	0.3471	0.0000	0.3124	0.6662	0.9719	1.0000
SHARE_VENTAS_DOM_M4	0.6106	0.3474	0.0000	0.2557	0.6722	0.9787	1.0000
SHARE_VENTAS_DOM_M5	0.6079	0.3468	0.0000	0.2714	0.6664	0.9830	1.0000
SHARE_VENTAS_DOM_M6	0.6099	0.3464	0.0000	0.2786	0.6734	0.9814	1.0000
SHARE_VENTAS_DOM_M7	0.6189	0.3450	0.0000	0.2776	0.6839	0.9844	1.0000
SHARE_VENTAS_DOM_M8	0.6250	0.3460	0.0000	0.2706	0.6920	0.9844	1.0000
SHARE_VENTAS_DOM_M9	0.6383	0.3460	0.0000	0.2898	0.6985	0.9826	1.0000
SHARE_VENTAS_DOM_M10	0.6459	0.3417	0.0000	0.3294	0.7247	0.9849	1.0000
SHARE_VENTAS_DOM_M11	0.6556	0.3389	0.0000	0.3975	0.7344	0.9848	1.0000
SHARE_VENTAS_DOM_M12	0.6581	0.3407	0.0000	0.3463	0.7492	0.9874	1.0000
SHARE_VENTAS_DOM_M13	0.6701	0.3342	0.0000	0.4022	0.7722	0.9912	1.0000
SHARE_VENTAS_DOM_M14	0.6802	0.3339	0.0000	0.4051	0.7795	0.9920	1.0000
SHARE_VENTAS_DOM_M15	0.6904	0.3308	0.0000	0.4508	0.8035	0.9927	1.0000
SHARE_VENTAS_DOM_M16	0.6974	0.3228	0.0000	0.4758	0.7879	0.9929	1.0000
SHARE_VENTAS_DOM_M17	0.7027	0.3196	0.0000	0.4759	0.8095	0.9923	1.0000
SHARE_VENTAS_DOM_M18	0.7096	0.3192	0.0000	0.4875	0.8279	0.9929	1.0000
HHHI_PROVIDERS_L3M	1,601	1,406	179	855	1,143	1,712	10,000
HHHI_PROVIDERS_L6M	1,457	1,341	210	781	1,017	1,619	10,000
HHHI_PROVIDERS_L9M	1,414	1,335	184	716	966	1,543	10,000
HHHI_PROVIDERS_L12M	1,355	1,251	189	687	936	1,445	10,000
HHHI_CLIENTES_L3M	2,949	2,674	78	850	2,006	5,011	10,000
HHHI_CLIENTES_L6M	2,898	2,606	80	772	2,009	4,956	10,000
HHHI_CLIENTES_L9M	2,833	2,562	81	834	1,962	4,762	10,000
HHHI_CLIENTES_L12M	2,745	2,521	83	839	1,848	4,743	10,000

A continuación, se presenta una representación gráfica de la evolución de la tasa de incumplimiento. Esta tasa se ha calculado considerando la ponderación del monto del crédito en color rojo, así como la ponderación de todos los créditos por igual en color azul. En el eje horizontal, se disponen los períodos mensuales correspondientes al origen de estos créditos, mientras que en el eje vertical se exhibe la tasa de incumplimiento asociada a cada uno de dichos períodos. Esta visualización se posiciona como una herramienta valiosa que nos permite adentrarnos en una comprensión más profunda del perfil de los deudores a lo largo del tiempo.

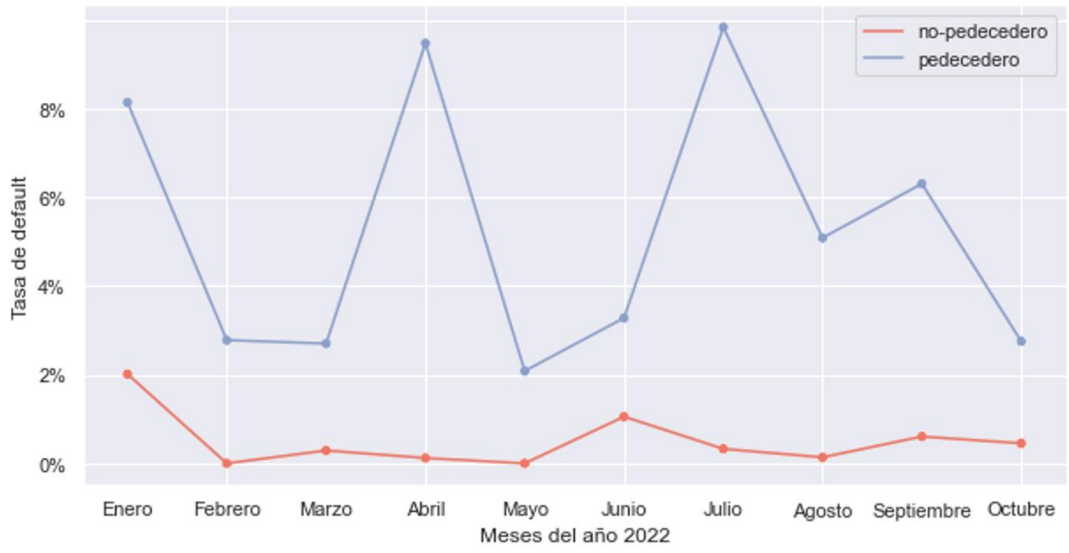
Figura 11) Evolucón de tasa de *default* ponderada por monto y no ponderada en el período de enero a octubre del año 2022.



El gráfico revela que, en la mayoría de los meses, con la excepción de mayo, la tasa de incumplimiento ponderada por el monto (representada por la línea roja) se sitúa por encima de la tasa de incumplimiento ponderando todos los créditos por igual (representada por la línea azul). Este patrón sugiere que los créditos que entraron en incumplimiento tuvieron un monto superior al promedio. La métrica de la tasa de incumplimiento ponderada por el monto se posiciona como una medida financiera y de negocios fundamental para evaluar la rentabilidad del producto. En contraste, la métrica sin esta ponderación ofrece una perspectiva más estadística y puede ser utilizada como una variable dicotómica a predecir. Por lo tanto, en el análisis descriptivo, se le dará prioridad a esta última métrica debido a su utilidad en el contexto de predicción.

En esta sección, nos centraremos en el análisis descriptivo de las variables independientes y su relación con la variable dependiente. Una variable de significativa importancia para este negocio es el tipo del segmento de los productos principales del exportador. Las empresas que están categorizadas como “Perecedero”, refiere a a empresas que principalmente vuelcan su negocio a la comercialización de productos perecederos, con un tiempo de vida muy limitado antes de que el producto se deteriore, como es el caso de la lechuga. Por otro lado, empresas que están categorizadas como “no-perecedero”en caso contrario hace referencia a empresas que principalmente comercializar productos que no tienen dicho riesgo inherente al producto.

Figura 11) Evolucón de tasa de *default* para los segmentos de industria perecedero y no perecedero en el período de enero a octubre del año 2022.

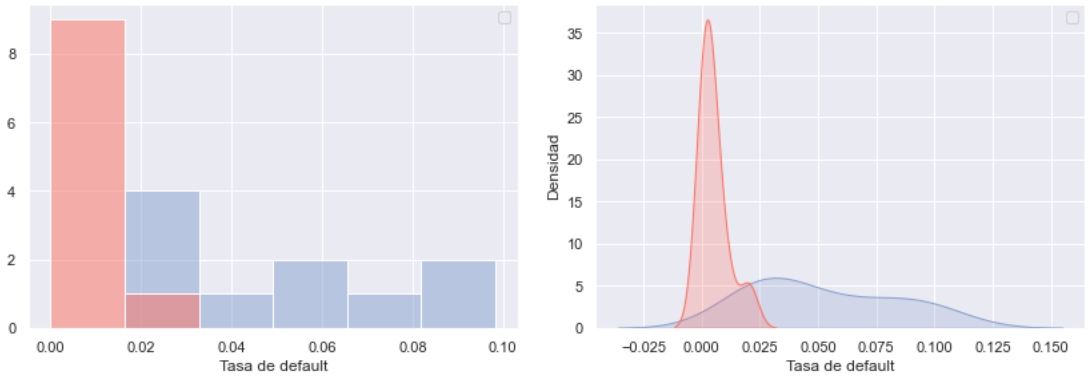


A figura 11) se presenta la evolución de la tasa de incumplimiento para ambos segmentos perecedero y no perecedero.

Como se puede apreciar, la tasa de incumplimiento del segmento perecedero no solamente es más elevada, sino también más volátil a lo largo del período analizado. Esta variabilidad se atribuye a la complejidad inherente al negocio del segmento

perecedero. Por ejemplo, la posibilidad de que los productos se deterioren durante el transporte prolongado puede generar disputas en el pago de las facturas.

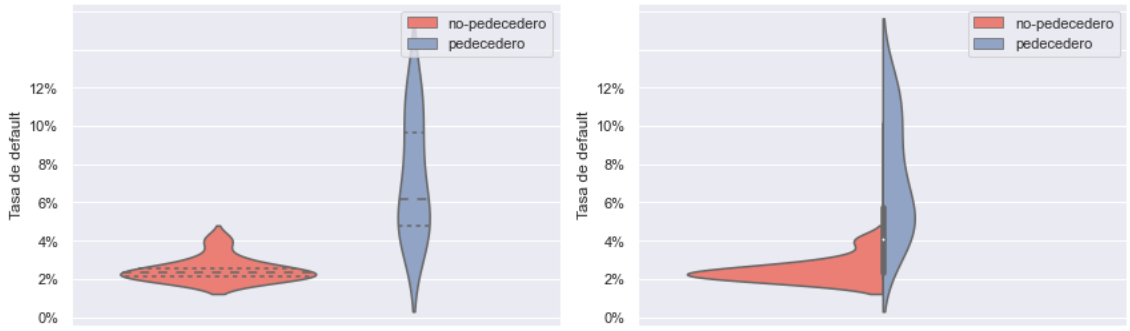
Figura 12) Histograma y distribución estimada de la tasa de *default* para los segmentos de industria perecedero y no perecedero.



La figura 12) muestra el histograma de arriba de la izquierda como la función de densidad de la derecha, ofrecen una representación visual de la distribución de la tasa de incumplimiento (default rate). En el eje X de ambos gráficos se representa la tasa de incumplimiento. Por otro lado, en el eje Y del histograma se muestra el recuento de observaciones dentro de cada intervalo de tasa de incumplimiento, mientras que en el gráfico de densidad se visualiza una función suavizada de dichos recuentos, denominada función de densidad estimada.

Como se puede observar, la distribución del segmento no perecedero está fuertemente concentrada en niveles de incumplimiento cercanos, especialmente en perfiles menos riesgosos. Por otro lado, el segmento perecedero se distribuye en un rango de valores más riesgosos.

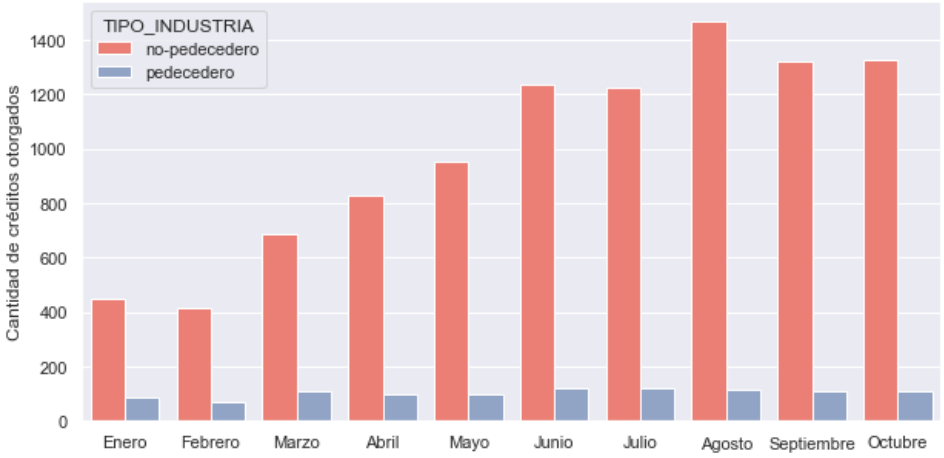
Figura 13) *Violin plot* de la tasa de *default* para los segmentos de industria perecedero y no perecedero.



La figura 13) muestra dos gráficos de el *violin plot* el primero ubicado a la izquierda, se muestra la distribución individual de cada conjunto de datos, lo que permite una evaluación detallada de la forma y la dispersión de cada distribución. Del lado derecho Podemos apreciar otro *violin plot*, en donde las distribuciones individuales se superponen, lo que facilita la comparación directa entre ellas y resalta las diferencias o similitudes entre los conjuntos de datos.

La figura 13) muestra de una forma más clara el mismo efecto que fue mostrado en la figura 12) en el cual segmento perecedero se distribuye en un rango de valores más riesgosos.

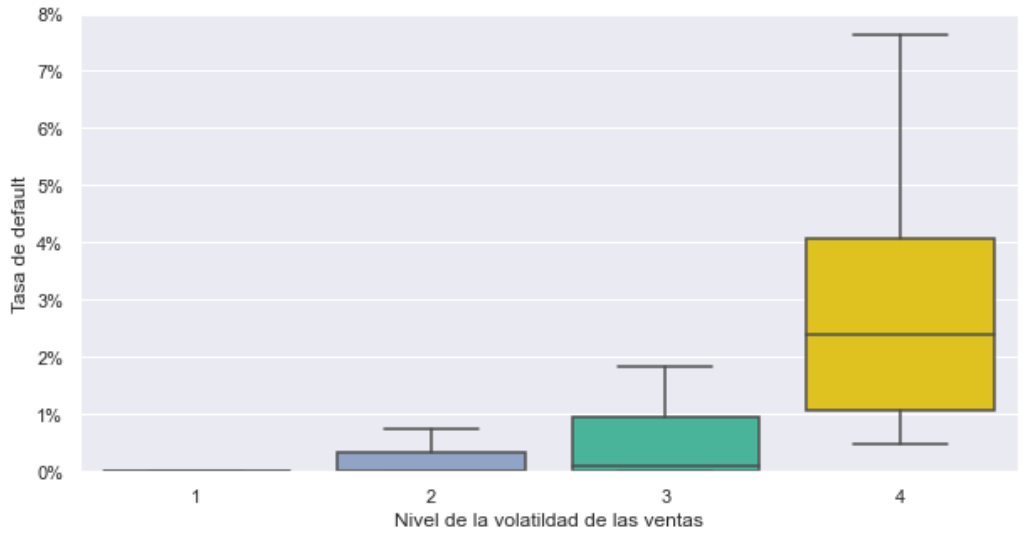
Figura 14) Cantidad de créditos otorgados en 2022, por segmento tipo de industria.



La figura 14) es un gráfico de barras que muestra la cantidad de créditos otorgados para los segmentos perecedero y no perecedero durante el año 2022. Cada barra refleja la cantidad respectiva de créditos en cada segmento.

El segmento no perecedero es el más relevante en cuanto a volume de créditos otorgados, además se puede apreciar un crecimiento durante el período analizado. En cambio, el segmento perecederos es menos importante en cuanto a volume y se aprecia una estabilidad de créditos otorgados a lo largo del análisis.

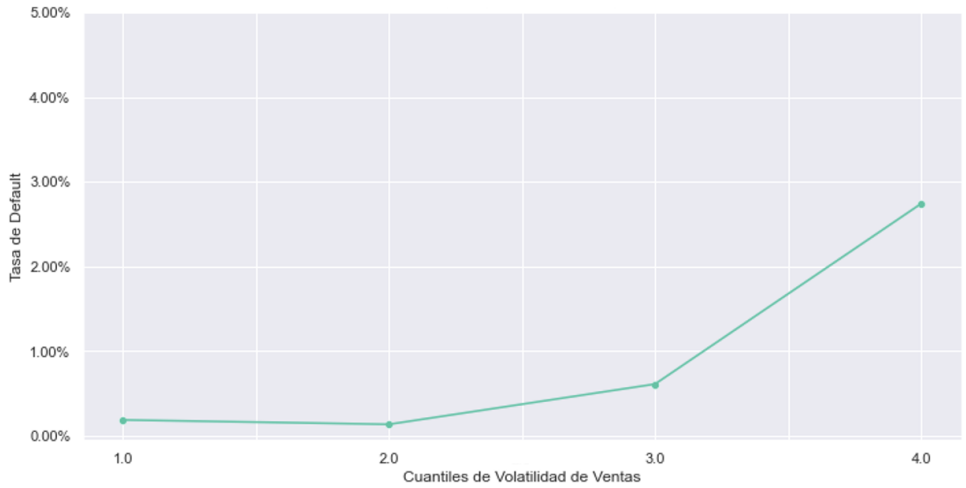
Figura 15) *boxplot* de la tasa de *default* por nivel de volatilidad de las ventas.



La figura 15) muestra como se distribuye la tasa de *default* con un *boxplot* para cada uno de los distintos niveles de volatilidad. Dicha variable representa los los cuantiles generados a partir de COEF_VAR_VENTAS_TOT.

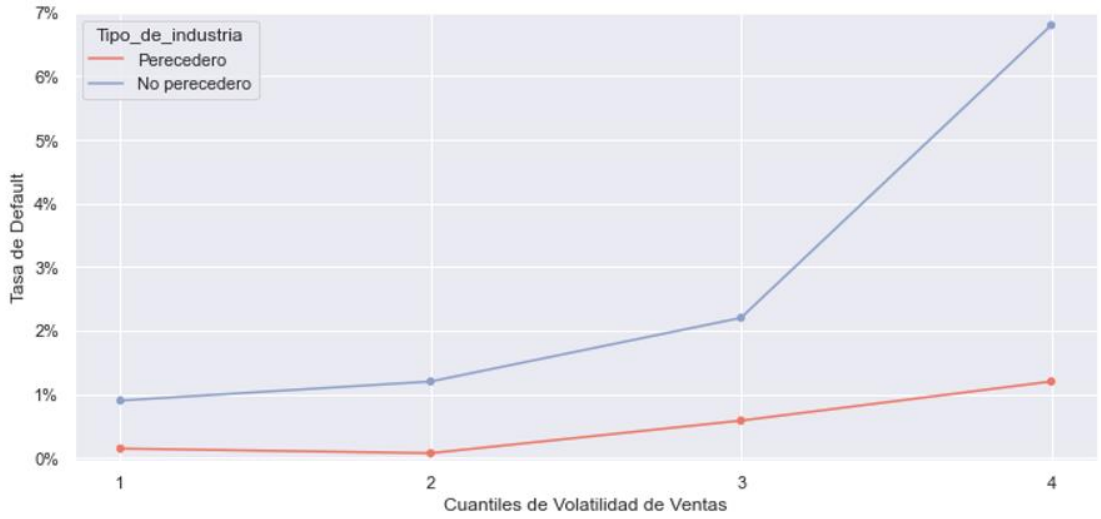
Si bien el objetivo de este gráfico no es utilizar la variable de cuantiles, es entender si existe alguna relación entre la tasa de default y el coeficiente de variación de las ventas mensuales. Como se puede apreciar, el primer cuantil (nivel estable), se conside con niveles de default más bajos, en consecutivo, niveles de volatilidad más altos se condicen con rangos de default más altos. Si bien los intervalos se solapan, el default mediano, el tercer cuartil y el máximo default es mayor en niveles de volatilidad más altos.

Figura 16) *lineplot* de la tasa de *default* promedio por nivel de volatilidad de las ventas.



La figura 16) muestra como evoluciona la tasa de *default* promedio a lo largo de los diferentes niveles de volatilidad, complementando a la figura 15). Como Podemos ver, a mayor nivel de volatilidad la tasa de *default* promedio del grupo aumenta.

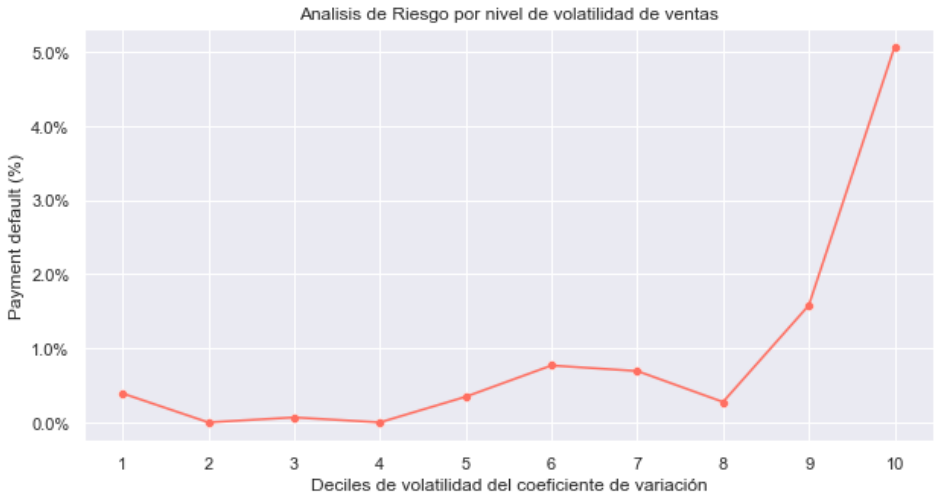
Figura 17) *lineplot* de la tasa de *default* promedio por nivel de volatilidad de las ventas para cada segmento de tipo de industria perecedero y no perecedero.



La figura 17) muestra la tasa promedio de *default* desglosada por tipo de segmento para los distintos niveles volatilidad de ventas. Este análisis proporciona una visión más detallada de la variabilidad en las tasas de incumplimiento en cada segmento, permitiendo una comprensión más profunda de los patrones y comportamientos asociados con cada categoría.

Como se puede apreciar en el gráfico, para el mismo nivel de volatilidad, el segmento perecedero sigue siendo más riesgoso que el no perecedero. Y visto por nivel de volatilidad, los niveles más volatiles son más riesgosos que los no volátiles.

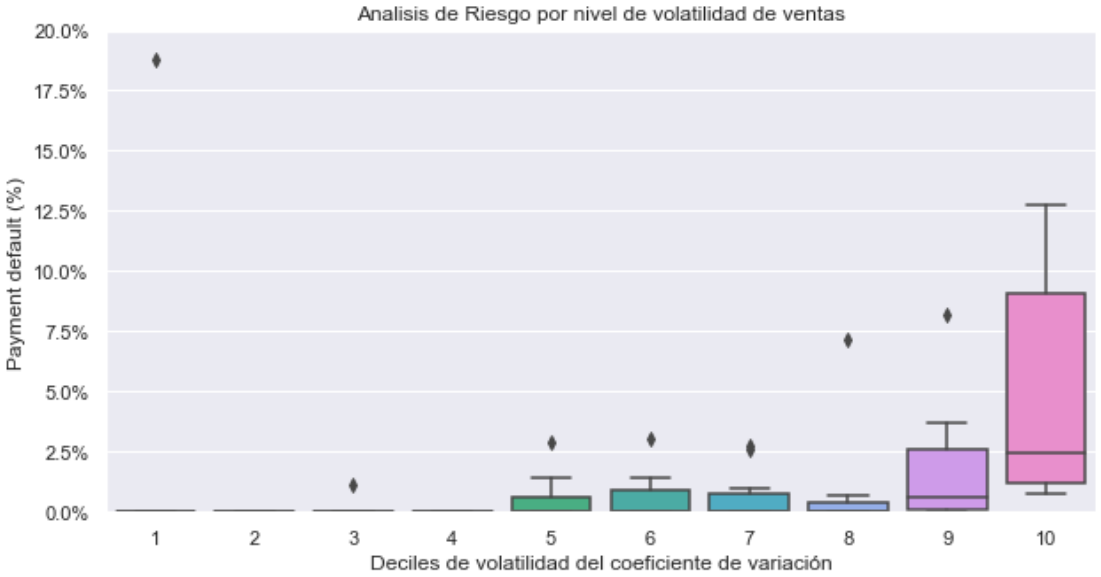
Figura 18) *lineplot* de la tasa de *default* promedio por nivel de volatilidad de las ventas en deciles.



La figura 18) ayuda a entender extender la técnica de cuantización vista en la figura 16) a deciles, la relación entre la volatilidad en ventas y el *default* persiste en una escala más detallada y finamente segmentada.

Notablemente, del primer al cuarto decil se muestra una tasa de *default* relativamente baja comparando a los siguientes deciles. Además el noveno y decimo decil muestran una tasa de default mayor al resto de deciles.

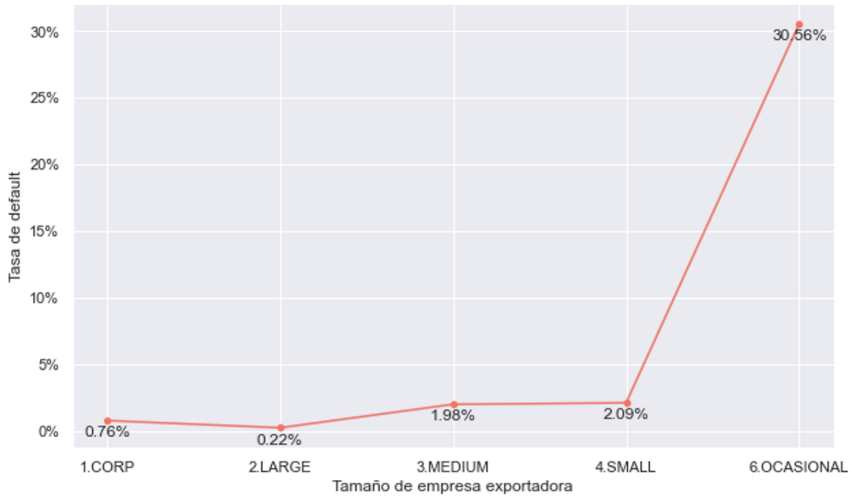
Figura 19) *boxplot* de la tasa de *default* promedio por nivel de volatilidad de las ventas en deciles.



La figura 19) muestra como se distribuye la tasa de *default* con un *boxplot* para cada uno de los distintos deciles de volatilidad.

Los primeros cuatro deciles tienen un nivel de *default* menor que los deciles consecutivos del quinto al octavo, y además, los deciles noveno y decimo tienen una tasa de default mayor que los niveles de volatilidad más estables. Dicho gráfico complementa la figura 18, de manera tal que el primer decil de default en la figura 18 muestra una tasa de default promedio por encima de los niveles de volatilidad segundo, tecer y cuarto decil. En la figura 19, podemos ver que el boxplot está representado como una línea en 0%, lo cual significa que, en general, el primer decil de volatilidad no tiene casos de default, excepto un romboide negro que representa un mes de originación puntual que tuvo un mal resultado. Sin embargo, esto no implica que el primer decil de volatilidad tenga un rendimiento peor que, por ejemplo, el segundo decil.

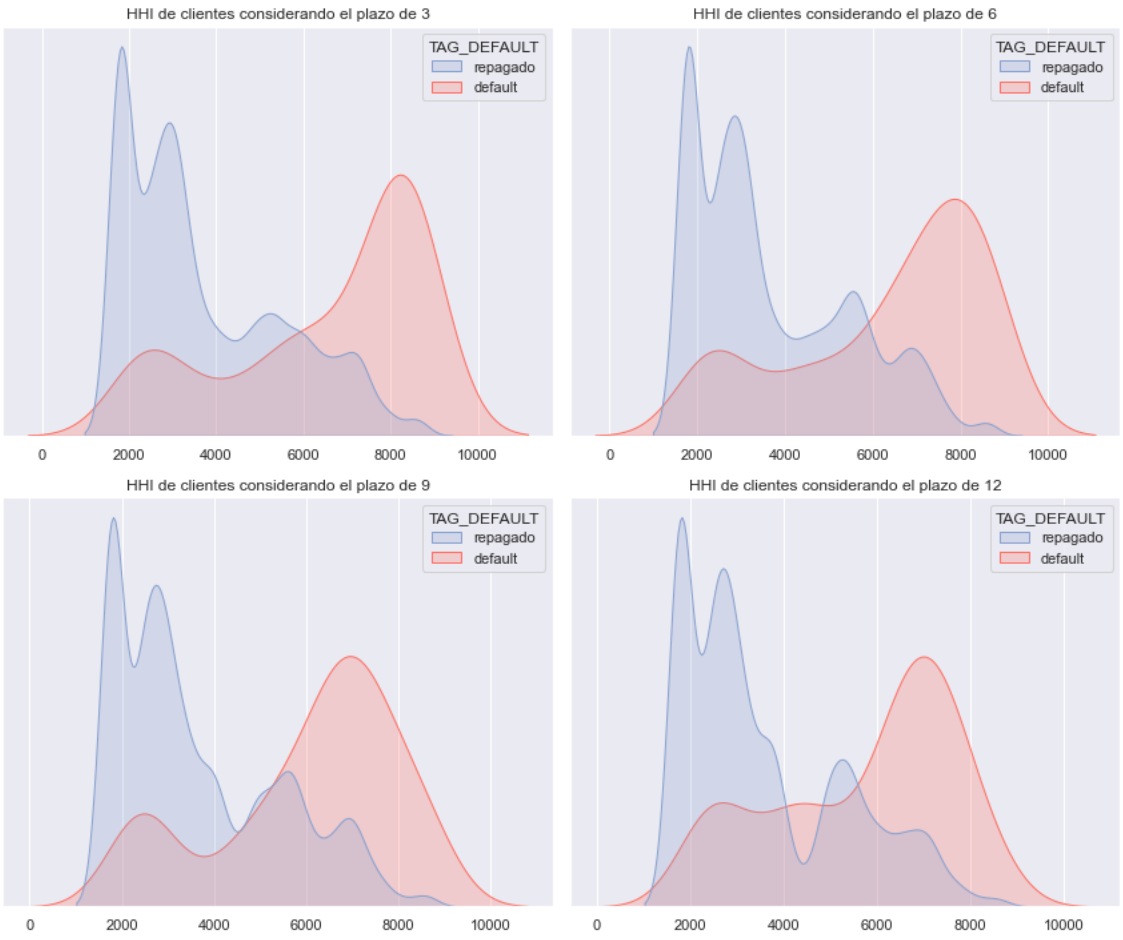
Figura 20) Tasa de *default* por tamaño de la empresa exportadora.



La figura 20) muestra la evolución de tasa de *default* por tamaño de la empresa exportadora. Las mismas se encuentran ordenadas de manera tal que el nivel de empresa Corp, que representa en tipo de tamaño más grande, hasta el nivel Ocasional que representa el tipo de tamaño más chico.

Como se observa en los resultados, existe una relación evidente entre el tamaño de la empresa exportadora y el nivel de *default*, mostrando que a medida que el tamaño de la empresa aumenta, la tasa de *default* tiende a disminuir, con la excepción del segmento Large/Corporate. En este último caso, las empresas Large exhiben una tasa de default inferior, a pesar de tener un volumen de exportaciones menor en comparación con las empresas Corporate. Es importante destacar que el segmento *Occasional* se refiere a empresas que exportan de manera eventual.

Figura 21) función de distribución estimada del HHI de las ventas de los 3, 6, 9 y 12 meses, por segmento de *default* y re pagado.



La figura 21) presenta la distribución estimada en función de la concentración de las ventas durante los últimos 3, 6, 9 y 12 meses anteriores a la calificación crediticia, tanto para las empresas que incumplieron como para aquellas que no lo hicieron. Este enfoque permite evaluar cómo la concentración de las ventas afecta la probabilidad de incumplimiento en diferentes periodos previos a la evaluación crediticia.

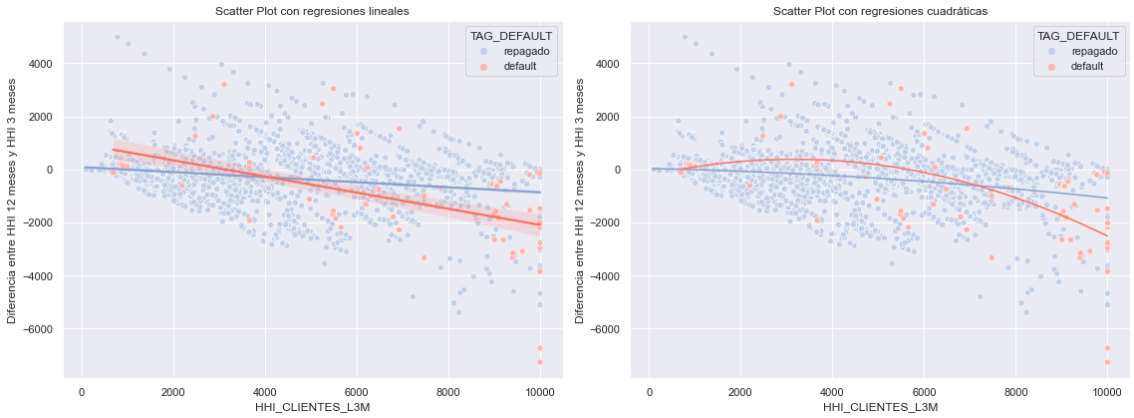
En el análisis de las cuatro distintas variantes, se observa una tendencia consistente en las empresas que incurrieron en incumplimiento, mostrando un Índice de Herfindahl-Hirschman (HHI) más elevado. Esto sugiere que las empresas dentro del grupo de *default* tienden a tener una mayor concentración de ventas en pocos clientes, mientras que las empresas con un HHI más diversificado exhiben un mejor comportamiento de pago.

Figura 21) *scatterplot* y regresión lineal del HHI de ventas de 12 y 3 meses por segmento de *default* y repagado.

La figura 21) cuenta con dos gráficos, el primero es un *scatterplot* en donde cada observación representa el HHI de la empresa en los últimos 3 meses (eje horizontal) y 12 meses (eje vertical), por segmento de *default* y repago. El segundo gráfico agrega además una regresión logistica para cada grupo basada en estos datos.

Observamos que, aunque ambos indicadores mantienen una alta correlación, el segmento que incurrió en incumplimiento presenta una pendiente levemente más horizontal. Este comportamiento indica un leve deterioro en el indicador de los últimos 3 meses en comparación con el comportamiento de los últimos 12 meses.

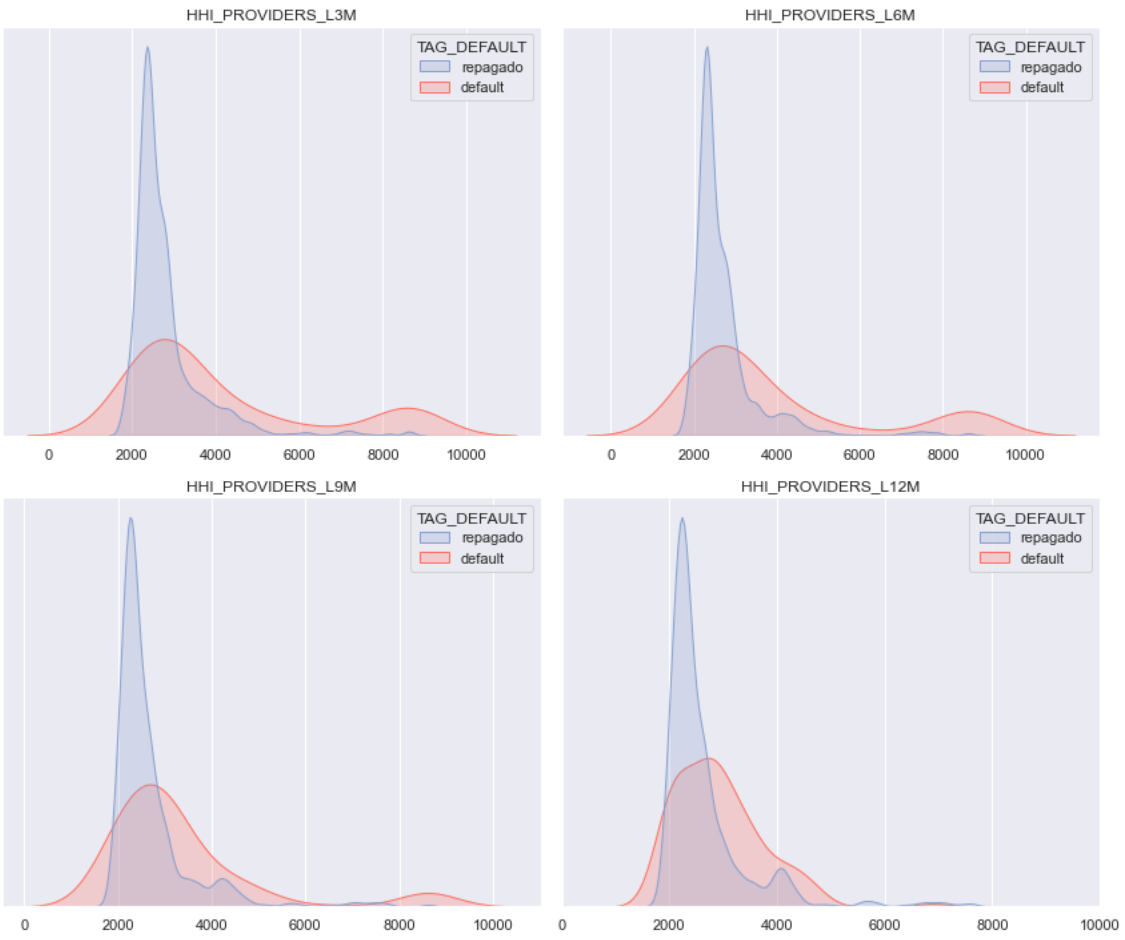
Figura 22) *scatterplot* con regresión lineal y cuadrática de la variación HHI de ventas por segmento de *default* y repagado.



La figura 22) nos ayuda a responder la pregunta de, si la variación de este indicador HHI entre los últimos 12 y 3 meses puede capturar de manera más efectiva el comportamiento de pago en comparación con la variable inicial. Para abordar este interrogante, se realizará un análisis utilizando dos enfoques diferentes: una regresión lineal en el primer gráfico y una regresión cuadrática en el segundo. Estos análisis nos brindarán la oportunidad de examinar cómo se relaciona el comportamiento de pago con la variación del indicador en comparación con la variable inicial. Además, nos permitirá evaluar la eficacia de esta variación como predictor del comportamiento de pago.

Al analizar los gráficos, se observa que ninguno de los dos parece ajustarse completamente al comportamiento de pago; sin embargo, el segundo gráfico, con la regresión cuadrática, parece explicar de manera más efectiva dicho comportamiento. Esto podría atribuirse al hecho de que la función cuadrática tiene la capacidad natural de capturar relaciones no lineales, como las que podrían existir en estos datos. La flexibilidad adicional de la función cuadrática puede permitir una adaptación más precisa a los patrones complejos presentes en el comportamiento de pago.

Figura 23) función de distribución estimada del HHI de las compras de los 3, 6, 9 y 12 meses, por segmento de *default* y re pagado.



La figura 23) muestra la distribución estimada en función de la concentración de las compras los últimos 3, 6, 9 y 12 meses anteriores a la calificación crediticia, tanto para las empresas que incumplieron como para aquellas que no lo hicieron. Este enfoque permite evaluar cómo la concentración de las compras afecta la probabilidad de incumplimiento en diferentes periodos previos a la evaluación crediticia.

Las cuatro variables exhiben un comportamiento similar; sin embargo, las variables que consideran el cálculo de los últimos 3 y 6 meses muestran una mayor concentración de empresas en el segmento de incumplimiento, con un indicador superior a 6000 que en las distribuciones de 9 y 12 meses no están.

Figura 24) *scatterplot* y regresión lineal del HHI de compras de 12 y 3 meses por segmento de *default* y repagado.

La figura 24) muestra con dos gráficos, el primero es un *scatterplot* en donde cada observación representa el HHI de de compras en los últimos 3 meses (eje horizontal) y la diferencia entre el HHI de los últimos 12 y 3 meses en el eje vertical, por segmento de *default* y repago. El segundo gráfico agrega además una regresión lineal para cada grupo basada en estos datos.

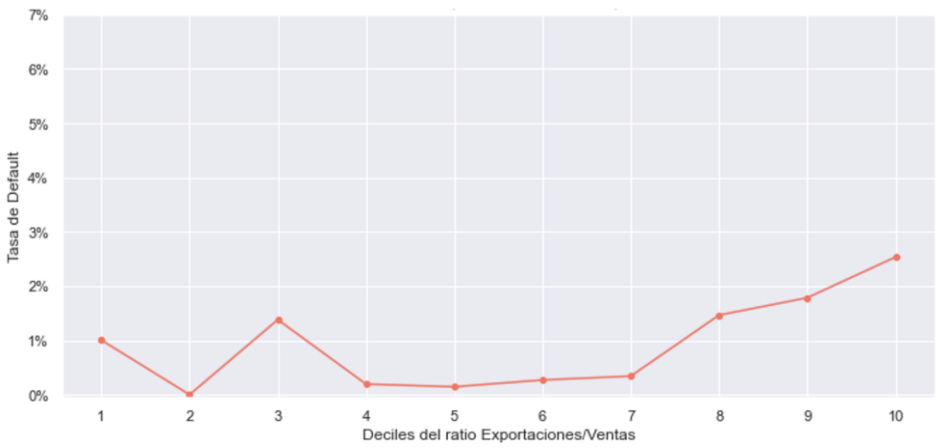
Se puede apreciar en el gráfico una relación más marcada para las observaciones con deterioro superior a los 6000 puntos de HHI se concentra en casos de *default* y una pendiente más vertical en la regresión de en el grupo de *default*.

Figura 25) *Scatterplots* con *splitlines* y regresiones lineales de la variación de HHI de compras y ventas para segmento de *default* y repago.

La figura 25) cuenta con dos gráficos, ambos contienen un *scatterplot* de la diferencia entre el HHI de ventas de 12 y 3 meses en eje horizontal y diferencia entre el HHI de compras de 12 y 3 meses en eje vertical por segmento de repago y *default*. El gráfico ubicado a la izquierda agrega regresiones lineales, mientras que el gráfico de la derecha agrega *splitlines*.

El objetivo de este análisis es entender si el deterioro de ambas variables puede ser un indicador relevante para considerarlo como variable *input.* Sin embargo no pareciera haber un patron que aporte al estudio de este fenomeno.

Figura 26) Evolución de tasa de default por deciles del ratio Exportaciones-Ventas Domésticas



La figura 26) muestra como se distribuye la tasa de *default* con un a través de los deciles basados en la variable RATIO_EXPORTACIONES_SOBRE_VENTAS.

Si bien el objetivo de este gráfico no es evaluar la necesidad de usar la variable de deciles. Se buscaes entender si existe alguna relación entre la tasa de default y el el ratio de exportaciones sobre ventas domésticas.

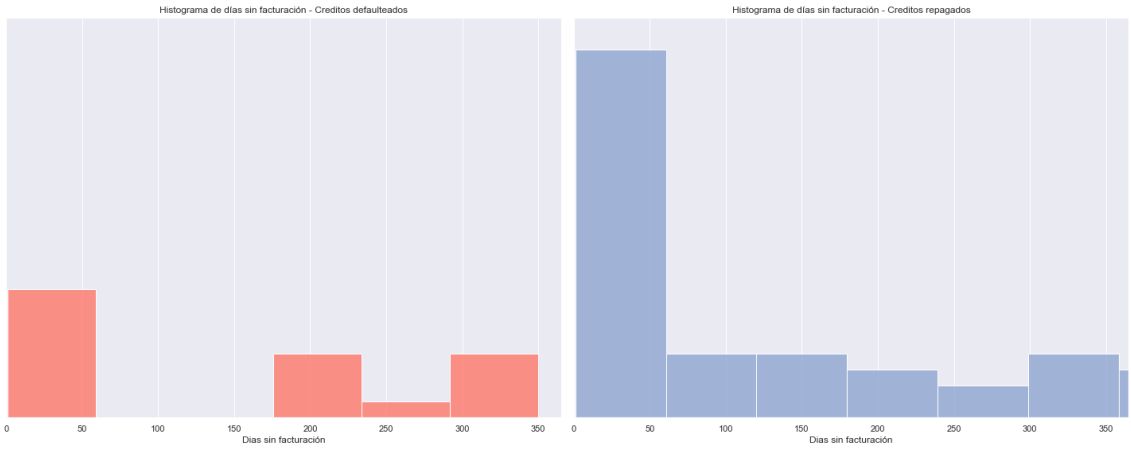
Las empresas con una mayor proporción de exportaciones sobre ventas domésticas, muestran un comportamiento de pago menos favorable a partir del octavo decil.

La figura 27) Evolución de la tasa de *default* por meses acumulados de exportación

La figura 27) muestra la evolución de la tasa de *default*, a través de los meses acumulados de exportaciones en los últimos 12 meses.

Aquellas empresas que registraron una mayor cantidad de meses consecutivos con exportaciones exhiben un comportamiento de pago más saludable, especialmente en el segmento de 11 y 12 meses. En contraste, las empresas con menos de 10 meses de exportaciones presentan un comportamiento más riesgoso, situación que se agrava al disminuir a menos de 6 meses de exportaciones.

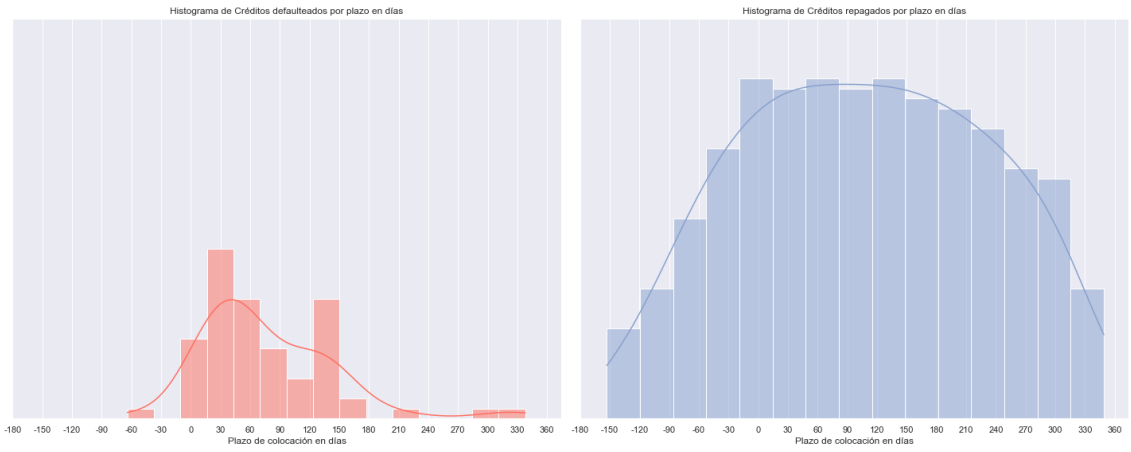
Figura 28) Histogramas de días sin facturación para créditos en *default* y repagados.



La figura 27) muestra 2 histogramas, que acumula en rangos de a 50 días la cantidad de créditos que fueron originados. El gráfico de la izquierda contiene a los créditos que terminaron en *default* y el de la derecha los repagados.

El histograma de créditos repagados muestra mayor consistencia a lo largo de el eje horizontal, ningún rango quedó sin observaciones. Y podemos encontrar una gran concentración en el rango de 0 a 50. Mientras que el histograma de créditos que entraron el *default* tiene valores sin observaciones en el rango 50 a 200 y no muestra una concentración muy marcada en ninguno de los rangos.

Figura 28) Histogramas de plazo de colocalción para segmentos de *default* y repago.

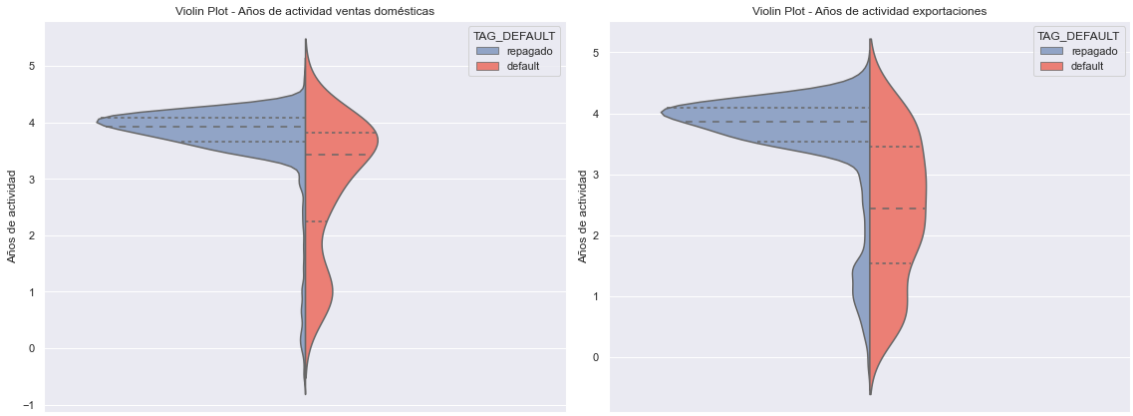


Las figura 28) exhibe dos gráficos de barras enfocados en la variable "Plazo de Colocación", abordando cada grupo por separado: conformado por aquellas que

incurrieron en incumplimiento y otro por las empresas que cumplieron con sus obligaciones.

La distribución de ambos gráficos se solapan. Por lo cual no se ve una relación entre el plazo de colocación y el comportamiento de pago.

Figura 29) *Violinplots* de años de actividad en ventas domésticas y exportaciones por segmento de *default* y repago.



La figura 29) muestra dos gráficos con la distribución estimada en *violinplots* de la de años de actividad en ventas domésticas y exportaciones para los segmentos de repago y *default*.

En ambas representaciones gráficas, se observa un patrón similar. Las empresas con una trayectoria más extensa acumulan un mayor número de buenos pagadores en comparación con aquellas que son más recientes, es decir, con una presencia en el mercado local o internacional inferior a 3 años. Sin embargo, es importante destacar que las empresas más recientes exhiben una proporción significativamente mayor de incumplimientos.

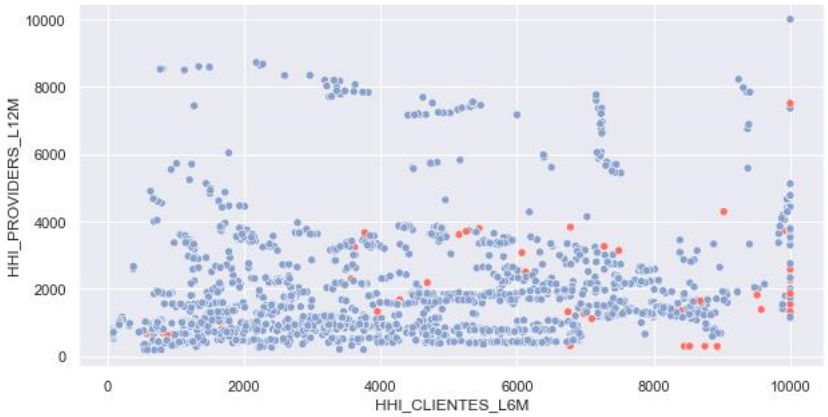
5. Entrenamiento de modelo

En la sección de analisis descriptivo se puso a prueba la aplicación de técnicas de modelado de variables. En esta sección, exploraremos la implementación de las técnicas de *data mining* mencionadas en el apartado de técnicas.

5.1 Aplicación de *k-means*

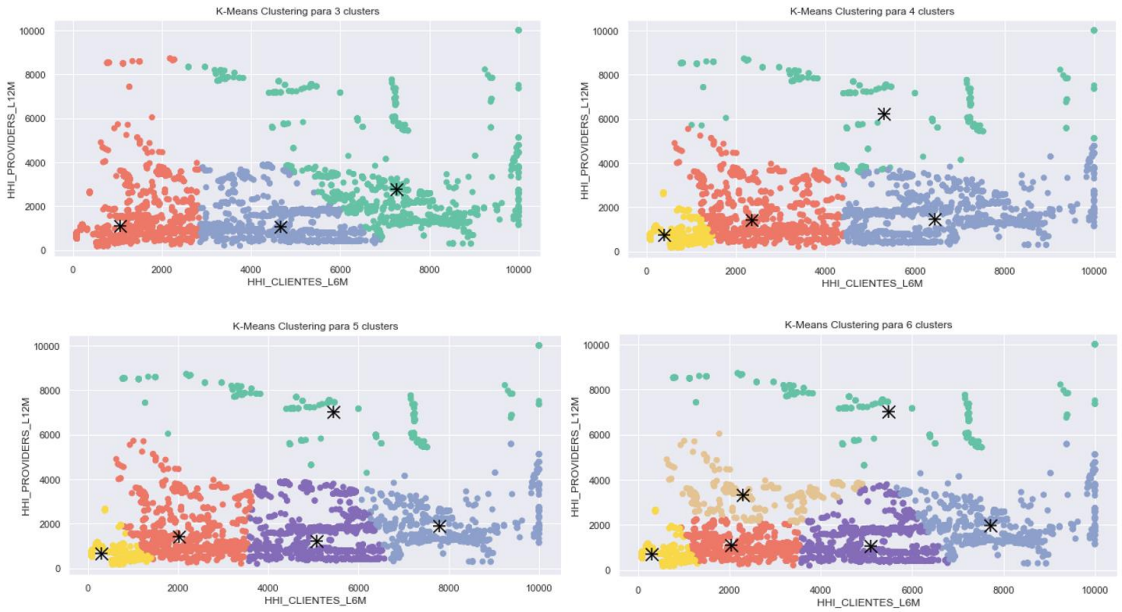
Como fue mencionado en el apartado de técncias, *k-means* es una técnicas de aprendizaje no supervisado. El objetivo en esta aplicación es generar una variable discriminante sin utilizar el *default* en la generación de la misma pero que colabore a segmentar las empresas que han incurrido en incumplimiento crediticio y aquellas que no lo han hecho. Este análisis se basará en el Índice de Herfindahl-Hirschman (HHI) de clientes durante los últimos 6 meses y el HHI de proveedores durante los últimos 12 meses.

Figura 30) scatterplot de HHI ventas 12 meses y compras de 6 meses por segmento de repago y *default*.



La figura 30) muestra que si bien existe una tendencia de mayor incidencia de incumplimiento en empresas con una concentración elevada en ventas, no se aprecia una relación positiva entre el incumplimiento y la concentración en compras.

Figura 31) aplicación de *K-means* para 3, 4, 5 y 6 *clusters*.



La figura 31) muestra como el algoritmo de k-means trabaja en base a los datos presentados en la figura 30.

Se presentan cuatro gráficos distintos, manteniendo los ejes de la figura 30. El primer cuadro, ubicado en la parte superior izquierda, exhibe el algoritmo con k = 3, es decir, con tres categorías diferentes. En el siguiente gráfico, situado en la parte superior derecha, muestra el algoritmo con cuatro categorías. El gráfico de la parte inferior izquierda representará el algoritmo con cinco categorías, y finalmente, en la parte inferior derecha, se presentará el algoritmo con seis categorías.

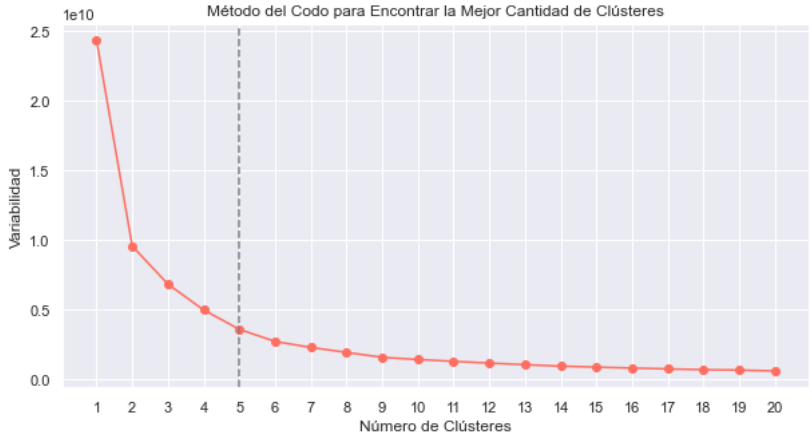
La elección de las cantidades de categorías para se basará en el método del codo como fue anticipado, una técnica que comúnmente es utilizada para determinar el número óptimo de grupos en un análisis de *k-means*. Este método implica realizar una serie de ejecuciones del algoritmo de *k-means* con un rango de valores de k y luego graficar la suma de las distancias al cuadrado de cada punto al centroide más cercano en función de k.

5.1.a Optimizando k a través del método del codo

Como fue revisado en el apartado de técnicas, el punto de inflexión en la curva, tomo el nombro de "codo", indica el número óptimo de *clusters*. Este punto representa el punto en el que el incremento en la variabilidad explicada por agregar otro *cluster* es significativamente menor que en pasos anteriores, lo que sugiere que añadir más clusters no mejora sustancialmente la estructura del agrupamiento.

Por lo tanto, al elegir visualizar estos cuatro valores específicos de k (3, 4, 5 y 6), buscamos identificar el punto en la curva donde se produce el codo, indicando así el número óptimo de categorías para este análisis específico. Esto proporcionará una comprensión más clara y significativa de la distribución de las empresas en función de su comportamiento de incumplimiento y su nivel de concentración en ventas y compras.

Figura 32) método del codo para optimización de k.



La figura 32) muestra como disminuye la variabilidad de los grupos a través de dos gráficos combinados de *scatterplot* y *lineplot* que se conoce como método del codo por el efecto visual que genera el gráfico.

La elección específica de mostrar el algoritmo con k = 5 se basa en el punto de inflexión identificado en la curva del método del codo. Después de realizar varias ejecuciones del

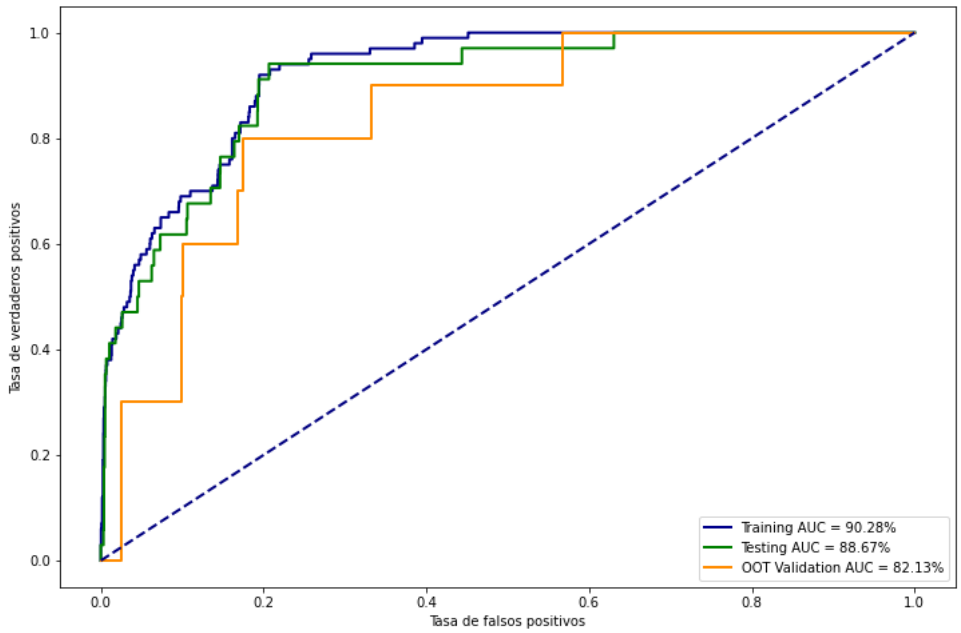
algoritmo de k-medias con diferentes valores de k y graficar la suma de las distancias al cuadrado en función de k, se observa que el punto de inflexión se encuentra en k = 5.

5.2 Aplicación modelo de regresión logística

5.2.a Variante 1 del modelo de regresión logística

A continuación, se presenta la primer variante del modelo de regresión logística. El presente modelo toma como variables de *input* todas las variables mencionadas en la definición de variables predictoras, excepto las variables de VENTAS_DOMESTICAS_M1 en consecutivo hasta VENTAS_DOMESTICAS_M18 y SHARE_VENTAS_DOM_M1 en consecutivo hasta SHARE_VENTAS_DOM_M18 para evitar multicolinealidad.

Figura 33) curva *AUC* en los conjuntos de datos de *train*, *testing* y *out of time* (OOT) del la variante 1 del modelo de regresión logística.



La figura 33 muestra la el area bajo de curva del modelo variante 1 de regresión logísitca para los conjuntos de datos *train*, *testing* y *out of time.* Se puede observer la el AUC en

el conjunto de *train* supera a la del conjunto de *testing* , y esta última a su vez supera a la del conjunto OOT. Este fenómeno es habitual, ya que el aprendizaje en un conjunto de datos puede experimentar cierto grado de sobreajuste. A pesar de ello, al comparar el rendimiento en train y testing, donde la diferencia es menor al 5%, esta relación tiene sentido.

5.2.b Variante 2 del modelo de regresión logística

En la segunda iteración del modelo, se incorporaron técnicas de re-muestreo mencionadas previamente, tomando como base las variables de *input* que fueron utilizadas en el modelo de la versión 1. La variante 2 incorpora estrategias de *undersampling*, *oversampling* y combinaciones de ambas como fue mencionado en el apartado de técnicas.

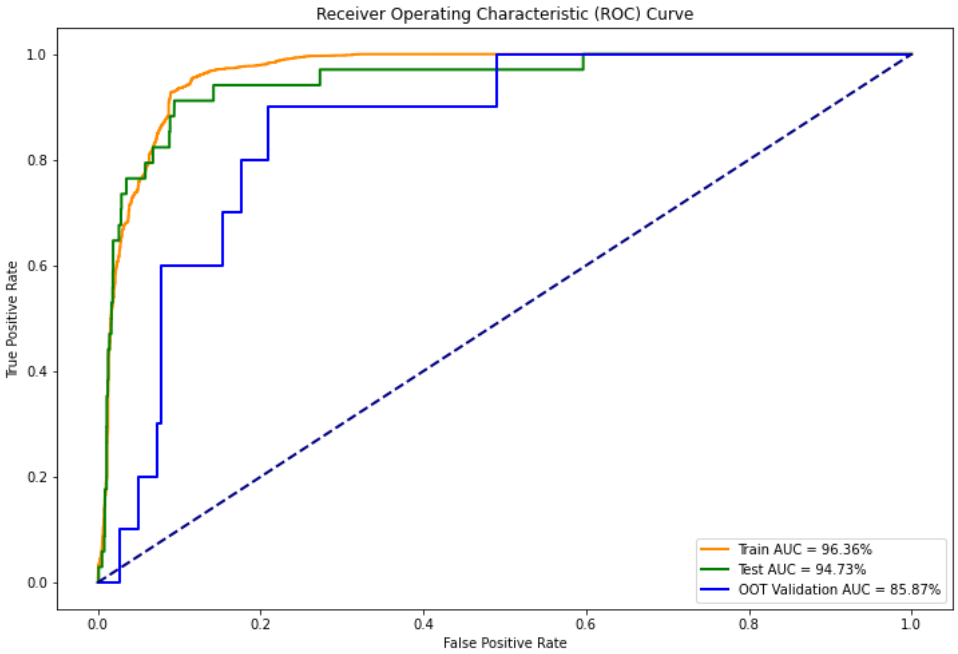
Tabla 4) Resumen de variantes con *undersampling*/*oversampling* del modelo de regresión logística

Variante	Undersampler	Oversampler	%clase minoritaria	AUC en OOT
2.a	1	-	50%	82.46%
2.b	0.6	-	37%	83.22%
2.c	0.3	-	23%	81.43%
2.d	-	1	50%	85.69%
2.e	-	0.6	37%	85.60%
2.f	-	0.3	23%	85.43%
2.g	0.6	0.6	37%	85.87%
2.h	0.5	1	50%	83.52%

La tabla 4) se comparan los indicadores AUC en el conjunto de datos *out of time* (OOT) para evaluar el rendimiento de las distintas variantes. A continuación, se presenta una tabla que resume las variantes empleadas y su respectivo desempeño.

Cabe destacar que, en el caso de la estrategia mixta, se llevó a cabo inicialmente el *undersampling* en el conjunto de entrenamiento, seguido por el *oversampling*.

Figura 34) curva *AUC* en los conjuntos de datos de *train*, *testing* y *out of time* (OOT) del la variante 2.g del modelo de regresión logística.



La figura 34 muestra la el area bajo de curva del modelo variante 2.e de regresión logísitca para los conjuntos de datos *train*, *testing* y *out of time.*

Se puede observar el AUC en el conjunto de *out of time* del 85.87% lo cual muestra ser una variante superior al modelo variante 1.

5.3 Variante de modelo basado en árboles de *boosting*

A continuación vamos a aplicar los distintos modelos de árbol que fue revisado en el apartado de técnicas. Como fue mencionado en dicho apartado, el mismo hace ejercicio de técnicas propias del modelo de *bagging* y *random forest* por lo cual se procederá a entrenar solamente el modelo de boosting dado que es el más complete desde el punto de vista de técnicas.

Este utilizará las mismas variables de *input* que fueron mencionadas para la variante 1 y 2 de la regresión logística.

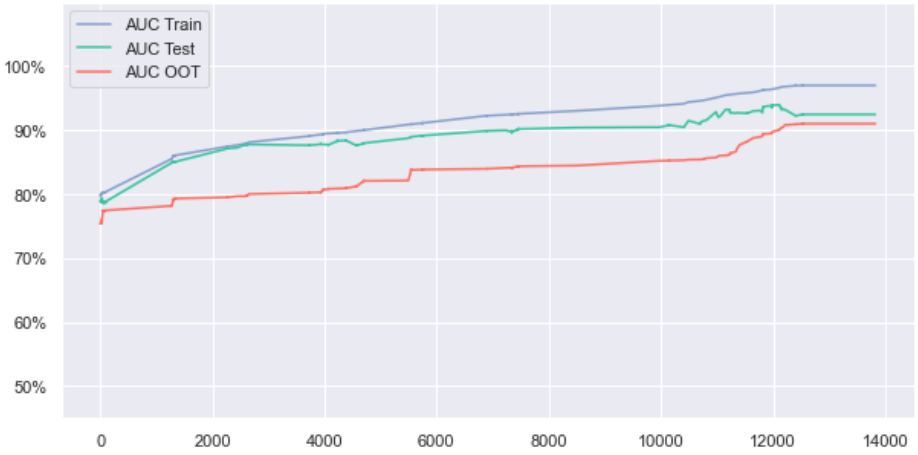
El entrenamiento de un modelo de árboles de *boosting*. implica el ajuste de nueve hiperparámetros diferentes, los cuales fueron detallados en la sección correspondiente a las técnicas empleadas. Una parte crucial del proceso de entrenamiento es la identificación del conjunto óptimo de hiperparámetros que permiten construir modelos de manera efectiva y precisa. En este contexto, se utilizó una grilla de hiperparámetros aleatoria, la cual se detalla a continuación:

Tabla 5) Grilla de hiperparámetros.

Nombre de variable	Valores
Máxima profundidad	[3, 5, 7, 11]
Tasa de aprendizaje	[0.1, 0.01]
Cantidad de estimadores	[100, 200, 300]
Peso mínimo de nodo	[10, 30, 50, 70]
Gama	[0, 0.2]
Submuestra	[0.2, 0.4, 0.8]
Muestreo de columnas por arbol	[0.2, 0.4, 0.6, 0.8]
Regularizacion alfa	[0, 0.1, 0.5]
Regularizacion lambda	[1.5, 2]

La tabla 5) muestra la grilla de hiper parámetros del modelo de boosting. Disponemos de cuatro valores para los hiperparámetros de máxima profundidad, mínimo de nodo y muestreo de columnas por arbol; tres valores para cantidad de estimadores, submuestra y regularizacion alfa; y 2 valores para tasa de aprendizaje, gama y regularizacion lambda. En conjunto, esto nos brinda la posibilidad de generar hasta 13,824 combinaciones únicas de hiperparámetros. El conjunto de hiperparámetros óptimo será mostrado en el apéndice.

Figura 35) evolución del AUC para los 14,000 modelos estimados en train, testing y out of time



La figura 35 presenta un gráfico que ilustra la evolución del rendimiento del modelo, ordenado de menor a mayor. En este contexto, el modelo con peor rendimiento en entrenamiento es representado por el número 1, mientras que el número 13,824 corresponde al modelo con el mejor rendimiento en entrenamiento.

Una observación de gran importancia es que el rendimiento durante el conjunto de *train* (representado por la línea azul) es consistentemente superior al rendimiento del conjunto de *testing* (línea verde), mientras que este último, a su vez, supera el rendimiento del conjunto *out of time* (línea roja).

El valor del Área bajo la Curva (AUC) para este modelo es del 90.99%, lo que lo sitúa por encima del rendimiento observado en la versión 4 del modelo de regresión lineal. Por consiguiente, esta versión del modelo de *boosting* ha sido seleccionada como la que posee mayor capacidad predictiva.

6. Reconto de tesis para la conclusión

En esta tesis, se ha investigado la aplicación de técnicas de modelado de variables financieras para predecir el incumplimiento financiero en empresas. Se comenzó con un análisis descriptivo y se procedió a la implementación de técnicas de data mining, destacando especialmente el uso de la técnica de *k-means*, modelos de regresión logística y árboles.

Resultados y Hallazgos

Aplicación de k-means

Se exploró la relación entre el Índice de Herfindahl-Hirschman (HHI) de clientes y proveedores y el incumplimiento financiero. Se observó una tendencia de mayor incidencia de incumplimiento en empresas con alta concentración en ventas, pero no se encontró una relación positiva con la concentración en compras.

Modelo de Regresión Logística

Se desarrollaron varias variantes de modelos de regresión logística, variando las variables utilizadas. Se observó un fenómeno de sobreajuste, con un mejor rendimiento en el conjunto de entrenamiento que en el de validación cruzada y Out Of Time (OOT).

Modelo Basado en Árboles de Boosting

Se entrenó un modelo utilizando el método de Árboles de Boosting, explorando una amplia gama de hiperparámetros. Se identificó un conjunto de parámetros que demostró un rendimiento superior, superando incluso al modelo de regresión logística.

Selección de modelo

Tabla 6) resumen de AUC en conjunto de datos *out of time* para las tres variantes.

Nombre de la variante	*AUC* en *out of time*
Variante 1 Regresión logística	82.13%
Variante 2 Regresión logística	85.87%
Árboles de boosting	90.99%

La tabla 6 muestra un resumen de la performance de las tres variantes de modelos que fueron detalladas en el apartado de entrenamiento.

El modelo con major performance corresponde al modelo de árboles de *boosting* con una performance de 90.99% por lo cual es el modelo seleccionado para implementar en el proceso de negocio de la empresa *fintech*.

Feature Engineering variables descartadas

Se generaron nuevas variables basadas en la distancia euclidiana hacia los centroides de las categorías "default" y "no default". A pesar de su potencial utilidad, se descartaron debido a la falta de relación con el modelo de *K-means* y la complejidad adicional que añadían al análisis.

Sugerencias para Investigaciones Futuras

Se podría explorar el refinamiento de los modelos existentes, así como la incorporación de técnicas de ensemble y redes neuronales para mejorar la precisión predictiva. Como también la creación de nuevas variables o datos al modelo como pueden ser los bureau de crédito externos.

En resumen, esta tesis ha proporcionado una exploración detallada de técnicas de modelado para predecir el incumplimiento financiero en empresas. Los resultados obtenidos sugieren la viabilidad de estos enfoques, pero también señalan áreas para futuras investigaciones y mejoras en la metodología.

7. Glosario

- Entidades crediticias: Instituciones financieras que ofrecen servicios de crédito, como préstamos y líneas de crédito.

- Segmentación de clientes: División de la base de clientes de una entidad en grupos homogéneos con características similares, con el fin de adaptar estrategias de marketing y riesgo específicas a cada grupo.

- *Factoring* financiero: Proceso mediante el cual una empresa vende sus facturas pendientes a un tercero (la empresa de *factoring*) a cambio de un anticipo de efectivo.

- Límites de crédito: La cantidad máxima de dinero que una institución financiera está dispuesta a prestar a un cliente.

- Gestión efectiva del riesgo crediticio: Proceso de identificación, evaluación y mitigación de los riesgos asociados con el otorgamiento de crédito para minimizar las pérdidas y maximizar la rentabilidad.

- Probabilidad de incobrabilidad: La posibilidad de que un prestatario no cumpla con sus obligaciones de reembolso, lo que puede resultar en pérdidas financieras para la entidad crediticia.

- Bancos digitales: Instituciones financieras que operan exclusivamente en línea, ofreciendo servicios financieros a través de plataformas digitales sin sucursales físicas.

- Fintech: Empresas emergentes en el sector financiero que utilizan la tecnología para ofrecer servicios financieros de manera innovadora y eficiente.

- Originaciones: Conjunto de de préstamos o créditos otorgados en un período.

- Evaluación de riesgos: Proceso de análisis de la probabilidad de pérdida y la magnitud de las pérdidas potenciales asociadas con una actividad específica, como el otorgamiento de crédito.

- *Factoring* internacional: Proceso financiero que implica la venta de cuentas por cobrar a una entidad financiera, generalmente en el contexto de transacciones comerciales internacionales.

- Segmentación por Volumen de Exportaciones: Se analiza la relación entre el tamaño de la empresa exportadora y la tasa de incumplimiento. Se observa que, en general, a medida que el tamaño de la empresa aumenta, la tasa de incumplimiento tiende a disminuir, excepto para el segmento Large/Corporate, donde las empresas grandes muestran una tasa de incumplimiento inferior a pesar de tener un volumen de exportaciones menor en comparación con las empresas Corporate.

- Política de crédito: Refiere a las reglas formales de que son definidas para garantizar la eficiente gestion crediticia. La politica de crédito tiene los siguientes componentes: Tasa de interés, CAP, plazo de colocación y RCI.

- CAP: Es el monto máximo al que se puede endeudar un cliente.

- Plazo de colocación: Es la unidad de tiempo desde la generación del crédito hasta el vencimiento.

- RCI: Representa un porcentaje máximo al cual se le puede prestar a un cliente en función de la relación cuota ingreso.

- Tasas de interés: Porcentaje que se aplica sobre el monto principal de un préstamo o crédito y que representa el costo del crédito para el prestatario.

- Modelos estadísticos de caja negra: Un modelo estadístico de caja negra es aquel en el que se observan los datos de entrada y salida pero el funcionamiento interno del modelo no es objetevo de análisis. Esto ocurre principalmente porque dichos modelos funcionan de manera compleja y dificil de interpretar.

- Bengio, Y., & Bergstra, J. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research.

- Chen, J., & Guo, C. (2021). "An Application of XGBoost Algorithm in Enterprise Credit Risk Assessment." Journal of Physics: Conference Series. Este estudio aplica el algoritmo XGBoost en la evaluación del riesgo crediticio empresarial.

- Chen, Y., Ding, S., Li, W., & Yang, S. (2018). Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China. IEEE Access.

- Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery.

- Haibo, H., & Yunqian, M. (2013). Imbalanced Learning Foundations, Algorithms, and Applications.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in Python. New York: Springer.

- Rhoades, S. A. (1993). The Herfindahl-Hirschman Index. Federal Reserve Bulletin, 79(3), 188-189.

- Rikkers, F., & Thibeault, A. (2015). A Structural form Default Prediction Model for SMEs, Evidence from the Dutch Market. Multinational Finance Journal.

- Wang, J., & Shi, Q. (2021). "A credit scoring model based on ensemble learning." Journal of Computational Science. Este artículo propone un modelo de puntuación crediticia basado en aprendizaje en conjunto, incluyendo técnicas de boosting.

- Yang, J., Hu, C., & Li, D. (2022). "Credit Risk Evaluation Based on a Comprehensive Credit Scoring Model Combining XGBoost and Random Forest." Mathematical Problems in Engineering. Esta investigación utiliza una combinación de XGBoost y Random Forest para la evaluación del riesgo crediticio.

- Yang, S., & Zhang, H. (2018). Comparison of Several Data Mining Methods in Credit Card Default Prediction. Intelligent Information Management, 10(5).

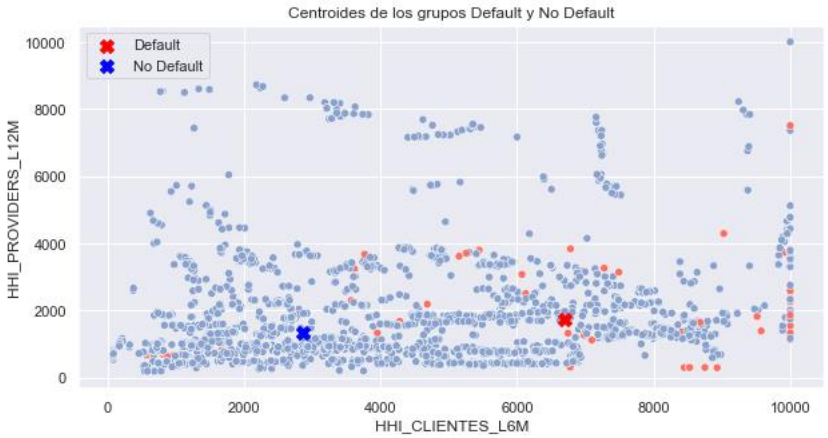
- Zhang, Y., Li, X., & Geng, X. (2021). "An improved XGBoost model for credit risk assessment based on adaptive random search algorithm." International Journal of Financial Engineering and Risk Management. Este estudio explora mejoras en el modelo XGBoost para la evaluación del riesgo crediticio.

- Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning. USA: O'Reilly.

Apéndice A: Técnicas de *Feature Engineer* descartadas.

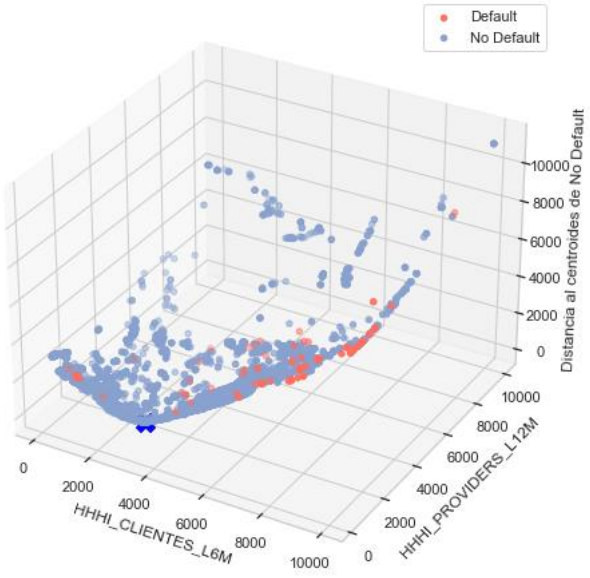
Para este trabajo, se llevó a cabo la generación de dos nuevas variables que capturan la distancia euclidiana hacia los centroides de las categorías "default" y "no default", tomando en consideración las variables de Índice de Herfindahl-Hirschman (HHI) de clientes y proveedores. Estas variables también se emplearon para construir el conjunto de características utilizado en el algoritmo *K-means*. Sin embargo, en esta aplicación en particular, se introduce una variable adicional que no guarda relación con el modelo de *K-means*.

A continuación se presenta una descripción visual, donde cada punto representa el HHI de los proveedores y clientes de una empresa en el momento de solicitar el crédito. Las empresas que lograron repagar el crédito se muestran en azul, mientras que aquellas que entraron en incumplimiento se destacan en rojo. Las cruces indican los centroides de cada grupo, que representan el valor promedio del HHI en cada uno.

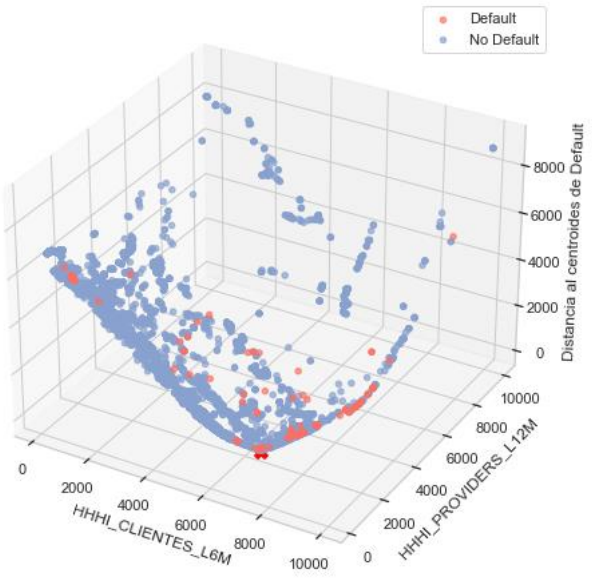


Basándonos en la representación visual anterior, se procedió a calcular dos nuevas variables, que consisten en la distancia euclidiana desde cada punto de datos hasta los respectivos centroides de las categorías "repago" y "incumplimiento".

A continuación, se presenta un gráfico en tres dimensiones que amplía la visualización anterior. Además de las dimensiones que muestran el Índice de Herfindahl-Hirschman (HHI) de los proveedores y clientes en el momento de la solicitud de crédito, se incorpora una tercera dimensión que representa la distancia de cada punto al centroide del grupo de empresas que realizaron el repago de su crédito.



Finalmente, el último gráfico de esta sección mantiene los mismos ejes del gráfico anterior. Sin embargo, en esta representación se añade una dimensión adicional que muestra la distancia euclidiana desde cada punto de datos hasta el centroide del grupo de empresas que entraron en incumplimiento (*default*).



Apéndice B: Parámetros del modelo de boosting que optimizan el AUC

Tabla 7) Conjunto de hiperparámetros que óptimo en OOT del modelo de boosting:

Nombre del parámetro	**Valor**
Máxima profundidad	11
Tasa de aprendizaje	0.1
Cantidad de estimadores	300
Peso mínimo de nodo	10
Gama	0
Submuestra	0.8
Muestreo de columnas por arbol	0.6
Regularizacion alfa	0
Regularizacion lambda	1.5