Diseño de un modelo predictivo del riesgo de cartera vencida en clientes en una empresa distribuidora de agroquímicos, a partir de indicadores comerciales y financieros, utilizando técnicas de regresión y aprendizaje automático

Yarith Eliana Cárdenas Leal

José Dorridt Mejía Rey

Trabajo de Grado para Optar al Título de: Especialista en Estadística

Director:

Dagoberto Bermúdez Rubio

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2025

Contenido

Resumen

Título: Diseño de un modelado Predictivo del riesgo de cartera vencida en clientes en una empresa distribuidora de agroquímicos, a partir de indicadores comerciales y financieros, utilizando técnicas de regresión y aprendizaje automático*

Autores: Yarith Eliana Cárdenas Leal

José Dorridt Mejía Rey **

Palabras Clave: Riesgo crediticio, cartera vencida, aprendizaje automático, regresión logística, Random Forest, modelado predictivo, gestión financiera.

Descripción:

El presente trabajo tiene como propósito desarrollar un modelo estadístico y de aprendizaje automático orientado a predecir la probabilidad de que un cliente empresarial de una compañía distribuidora de agroquímicos presente una cartera vencida alta. El estudio se basa en información comercial, de ventas y financiera correspondiente al año 2024. La metodología integra técnicas de regresión logística, árboles de decisión y Random Forest, evaluadas mediante métricas como precisión, sensibilidad y área bajo la curva ROC. A través de este enfoque, se busca identificar los factores que influyen en la morosidad y construir una herramienta analítica que fortalezca la toma de decisiones crediticias. Se espera que los resultados contribuyan al diseño de políticas de crédito más efectivas y a la mejora de la gestión de riesgo financiero dentro de la organización.

*Trabajo de Grado

**Facultad de Ciencias. Escuela de Matemáticas.

Abstract

Title: Design of a predictive model for the risk of overdue accounts receivable among customers of an agrochemical distribution company, based on commercial and financial indicators and using regression and machine learning techniques *

Authors: Yarith Eliana Cárdenas Leal

José Dorridt Mejía Rey **

Key Words: credit risk, overdue accounts, machine learning, logistic regression, Random Forest, predictive modeling, financial management.

Description:

This work aims to develop a statistical and machine learning model designed to predict the probability that a business client of an agrochemical distribution company will present a high level of overdue accounts. The study is based on commercial, sales, and financial data from the year 2024. The methodology integrates logistic regression, decision trees, and Random Forest models, evaluated through metrics such as accuracy, recall, and the area under the ROC curve. Through this approach, the research seeks to identify the main factors influencing delinquency and to build an analytical tool that strengthens credit decision-making. The results are expected to contribute to more effective credit policies and improved financial risk management within the organization.

* Degree work

** Faculty of Sciences. School of Mathematics

Introducción

La gestión del riesgo de cartera es un componente esencial para la sostenibilidad financiera de las organizaciones, especialmente en sectores donde el otorgamiento de crédito a clientes constituye una práctica habitual para incentivar las ventas. En el contexto empresarial colombiano, la morosidad representa uno de los principales desafíos en la administración financiera, ya que el incumplimiento en los pagos afecta directamente la liquidez, la rentabilidad y la capacidad operativa de las empresas.

Ante este panorama, la identificación temprana de clientes con alta probabilidad de incumplimiento resulta determinante para implementar estrategias preventivas que reduzcan el deterioro de la cartera y mejoren la eficiencia en la recuperación de recursos.

El presente trabajo propone el desarrollo de un modelo estadístico y de aprendizaje automático orientado a predecir la probabilidad de que un cliente empresarial de una compañía dedicada a la distribución de agroquímicos, presente una cartera vencida alta, utilizando información comercial, de ventas y cartera proporcionada por la empresa, junto con datos financieros correspondientes al año 2024, obtenidos de la plataforma EMIS.

A partir de técnicas de análisis descriptivo, correlacional y predictivo, se busca identificar los factores que influyen con mayor peso en el comportamiento de la cartera, comparando el desempeño de distintos modelos de clasificación supervisada como la regresión logística, los árboles de decisión y el método Random Forest, evaluados mediante métricas de precisión, sensibilidad y área bajo la curva ROC (AUC).

De esta manera, el estudio pretende construir una herramienta analítica que permita estimar

la probabilidad de morosidad y apoyar la toma de decisiones financieras, contribuyendo a fortalecer las políticas de crédito, mejorar la gestión de riesgo y optimizar la salud financiera de la organización.

1. Planteamiento y Formulación del Problema

La empresa propietaria de la información utilizada en este trabajo de grado busca otorgar líneas de crédito a sus clientes empresariales como parte de su estrategia comercial. Sin embargo, enfrenta el desafío permanente de gestionar de manera eficiente su cartera y prevenir la acumulación de saldos en mora. La ausencia de herramientas analíticas que permitan anticipar la probabilidad de incumplimiento genera consecuencias directas sobre la estabilidad financiera, como la reducción de liquidez, el aumento de provisiones contables y el deterioro de la rentabilidad.

Actualmente, la empresa utiliza una metodología de otorgamiento basada en algunos indicadores de liquidez y rentabilidad, junto con una revisión rápida de los estados financieros. En el caso de los clientes del sector agroindustrial, esta evaluación se complementa con el análisis del número de hectáreas cultivadas.

Sin embargo, la compañía se encuentra en un proceso de expansión hacia mercados menos estables que el agroindustrial, como la venta de productos a empresas que los distribuyen directamente a agricultores. Este tipo de cliente, conocido como Retail, presenta una mayor probabilidad de quiebra, ya que es un segmento con alta competencia y una marcada dependencia de las variaciones en los precios de los cultivos de pequeños productores.

2. Antecedentes

Los modelos de predicción de incumplimiento y quiebra corporativa tienen una larga tradición en finanzas cuantitativas. El punto de partida clásico es (Altman, 1968) quien, mediante análisis discriminante y razones financieras, propuso el Z-Score para anticipar quiebras con buen desempeño predictivo en empresas manufactureras. Su enfoque abrió el camino para el uso sistemático de indicadores contables en la predicción de riesgo.

Durante las décadas siguientes, la literatura se consolidó con aportes que sistematizan los métodos estadísticos de clasificación aplicables al "credit scoring". (Thomas, 2002) sintetizan técnicas como regresión logística, k-vecinos y árboles de decisión para clasificar clientes en buenos o malos riesgos, estableciendo criterios comparables de evaluación.

Un compendio ampliamente citado es el libro de Thomas, Crook y Edelman (2002), que formaliza el desarrollo de scorecards, la selección de variables, la validación out-of-sample y el uso de métricas como ROC-AUC, precisión y sensibilidad; esta obra sigue siendo referencia metodológica para problemas de mora y probabilidad de default.

Con el avance de la analítica de datos, surgieron nuevas metodologías basadas en aprendizaje automático (Machine Learning, ML), que permiten identificar patrones no lineales en grandes volúmenes de información. En este campo, (Lessmann, 2015) realizaron un análisis comparativo de distintos algoritmos de clasificación como árboles de decisión, Random Forest y

Gradient Boosting, evidenciando que los modelos de conjunto presentan un mejor desempeño predictivo que los enfoques tradicionales en la detección de riesgo crediticio. Más recientemente, (Noriega, 2023) revisaron la literatura sobre el uso de ML en predicción de mora, destacando la importancia de métricas como la precisión, sensibilidad y el área bajo la curva ROC (AUC) para la evaluación del desempeño de los modelos.

Aunque algunos clásicos se centran en quiebra y otros en scoring de consumo, su marco metodológico es directamente transferible a la predicción de cartera vencida en clientes empresariales: selección de variables financieras y comerciales, construcción de un objetivo binario (mora alta vs. no), comparación de modelos (logística, árbol, Random Forest) y evaluación con métricas discriminantes y de clasificación (Hand, 1997).

En el contexto latinoamericano, diversas investigaciones han aplicado estos métodos para predecir morosidad y analizar el riesgo de cartera en entidades financieras y cooperativas de crédito. Por ejemplo, (Ibarra Gallo, 2025) implementaron modelos de series temporales para pronosticar la evolución de la cartera vencida en una cooperativa ecuatoriana, y (Navas Alcívar, 2023) empleó árboles de decisión para evaluar el impacto de la cartera impaga sobre la rentabilidad del sistema bancario ecuatoriano. De igual forma, (Delgado-Giler, 2024) analizó los niveles de mora en cooperativas microempresariales, resaltando la utilidad de los modelos predictivos para priorizar estrategias de cobranza y control de riesgo.

En el caso colombiano, (Borrero-Tigreros, 2020) desarrolló un modelo de predicción de riesgo crediticio basado en técnicas de inteligencia artificial, demostrando la aplicabilidad de estos

enfoques en entidades financieras del país. Por su parte, (Batioja Bravo, 2022) diseñó un modelo híbrido de riesgo crediticio que combina métodos estadísticos y computacionales, mejorando la capacidad de predicción del comportamiento de pago. Igualmente, (Sepúlveda Rivillas, 2012) estimaron la probabilidad de incumplimiento en empresas del sector real colombiano mediante un modelo Probit, identificando que la rentabilidad, la liquidez y el apalancamiento son factores determinantes del riesgo de crédito.

Estos antecedentes evidencian una tendencia creciente hacia la aplicación de modelos predictivos supervisados para evaluar el riesgo de cartera y la probabilidad de mora, tanto en el ámbito internacional como en el latinoamericano y colombiano. Sin embargo, en el contexto empresarial no financiero, aún son limitados los estudios que utilizan bases de datos internas de empresas para construir modelos de riesgo de crédito, lo que resalta la pertinencia de este trabajo al aplicar metodologías analíticas avanzadas integrando información histórica del año 2024 sobre información comercial, financiera y de ventas de clientes empresariales del sector de distribución de agroquímicos.

3. Justificación

La gestión eficiente del riesgo de cartera constituye una necesidad prioritaria para las empresas que operan bajo esquemas de crédito comercial. En el caso de la organización objeto de estudio, la acumulación de saldos vencidos afecta directamente la liquidez operativa, incrementa los costos de financiación y compromete la rentabilidad del negocio. Ante este escenario, disponer de una herramienta que permita predecir la probabilidad de mora de los clientes se convierte en un

elemento estratégico para fortalecer la sostenibilidad financiera y mejorar la toma de decisiones crediticias.

La empresa dispone de información histórica de sus clientes empresariales en aspectos de ventas, cartera y finanzas; sin embargo, esta última se encuentra únicamente en formato físico, a través de formularios de creación de clientes, estados financieros y certificados de Cámara de Comercio. Por esta razón, se decidió complementar la base de datos con información financiera obtenida de la plataforma EMIS, incluyendo los principales indicadores y cuentas de los estados financieros, con el fin de incorporar estas variables al modelo sin inconvenientes. Sin embargo, los datos comerciales recopilados por la empresa no han sido aprovechados sistemáticamente mediante técnicas analíticas avanzadas que permitieran identificar patrones de comportamiento asociados al incumplimiento. La ausencia de un modelo predictivo limita la capacidad de la organización para anticiparse al riesgo y actuar de manera preventiva, dependiendo en gran medida de la experiencia subjetiva de los analistas o de criterios empíricos poco estandarizados.

Desde una perspectiva metodológica, el proyecto aporta valor al aplicar modelos estadísticos y de aprendizaje automático como la regresión logística, los árboles de decisión y el método Random Forest sobre una base de datos real del sector agropecuario colombiano. Esta aproximación permite comparar la capacidad predictiva de diferentes algoritmos y seleccionar aquel que proporcione un equilibrio adecuado entre precisión y facilidad de interpretación, contribuyendo al desarrollo de soluciones analíticas aplicadas al contexto empresarial.

Adicionalmente, los resultados del modelo permitirán a la empresa segmentar a sus clientes según su nivel de riesgo, diseñar políticas de crédito más precisas, definir estrategias de cobranza diferenciadas y priorizar acciones preventivas. Desde el punto de vista académico, el trabajo contribuye a fortalecer la integración entre la estadística aplicada, la analítica de datos financieros, demostrando el potencial del aprendizaje automático para resolver problemas reales de riesgo crediticio en empresas no financieras del país.

4. Objetivos

4.1 Objetivo General

- Desarrollar un modelo estadístico y de aprendizaje automático que permita predecir la probabilidad de que un cliente empresarial del sector de distribución de productos agroquímicos presente una cartera vencida alta, utilizando variables comerciales, financieras y de ventas correspondientes al año 2024, con el fin de apoyar la toma de decisiones en la gestión del riesgo crediticio y la recuperación de cartera.

4.2 Objetivos Específicos

- Explorar la relación entre las variables explicativas y el estado de la cartera, mediante técnicas estadísticas y correlacionales que permitan identificar los factores más influyentes en el riesgo de mora.
- Construir y comparar modelos predictivos supervisados incluyendo la regresión logística, los árboles de decisión y el algoritmo Random Forest para estimar la probabilidad de cartera vencida alta en los clientes.
- Evaluar el desempeño de los modelos mediante métricas de validación como la precisión, sensibilidad, especificidad, y el área bajo la curva ROC (AUC), seleccionando el modelo con mayor capacidad discriminante y estabilidad.
- Interpretar los resultados del modelo seleccionado, determinando las variables más relevantes en la predicción de morosidad y formulando recomendaciones prácticas para fortalecer las políticas de crédito y cobranza de la empresa.

5. Marco de Referencia

5.1 Riesgo de crédito y gestión de cartera

El riesgo de crédito se define como la probabilidad de pérdida que enfrenta una organización cuando un cliente o contraparte incumple las obligaciones pactadas en un contrato financiero o comercial. Este riesgo afecta de manera directa la liquidez, rentabilidad y estabilidad de las empresas, ya que el retraso o la falta de pago de los clientes compromete la disponibilidad de recursos para la operación.

En las empresas que otorgan crédito comercial, el seguimiento permanente de la cartera es una práctica esencial para asegurar su sostenibilidad. De acuerdo con el Banco de la República (Banco de la República, 2020), una cartera sana debe presentar niveles de morosidad bajos y adecuados mecanismos de control que garanticen la recuperación oportuna de los recursos. Sin embargo, cuando la proporción de cuentas vencidas aumenta, la empresa enfrenta dificultades en su flujo de caja y en su capacidad para reinvertir o atender compromisos financieros.

La cartera vencida representa entonces un indicador clave del riesgo crediticio. En el ámbito empresarial, se considera cartera vencida aquella porción del crédito cuyo vencimiento supera el plazo estipulado sin que se haya efectuado el pago correspondiente. Analizar la dinámica de esta cartera, así como los factores que la determinan, es fundamental para diseñar estrategias de mitigación y prevención del riesgo.

5.2 Modelos estadísticos y predictivos en el análisis de riesgo y de crédito

Históricamente, el riesgo crediticio se ha abordado mediante modelos estadísticos tradicionales que permiten estimar la probabilidad de incumplimiento de un cliente a partir de variables financieras y contables. Entre los trabajos más reconocidos se encuentra el de (Altman, 1968) quien desarrolló el modelo Z-Score utilizando el análisis discriminante para predecir la quiebra empresarial. Posteriormente, la regresión logística se consolidó como una herramienta ampliamente empleada en la estimación de la probabilidad de impago, al permitir modelar una variable binaria (moroso o no moroso) en función de múltiples predictores.

Con el auge del análisis de datos y la inteligencia artificial, surgieron los modelos de aprendizaje automático (Machine Learning, ML) como una actividad esencial para el análisis del riesgo crediticio en el sector bancario principalmente (Machado, 2025), que amplían la capacidad de los métodos estadísticos al detectar relaciones no lineales y complejas entre las variables. Estos algoritmos no se limitan a suponer una forma funcional predefinida, sino que aprenden directamente de los datos, ajustando sus parámetros para minimizar los errores de predicción.

El ML se ha convertido en una herramienta valiosa para el análisis del riesgo crediticio, ya que permite manejar bases de datos con numerosos atributos, mejorar la precisión de los modelos y generar interpretaciones visuales sobre las variables más relevantes. En este contexto, los algoritmos como la regresión logística, los árboles de decisión y el método Random Forest se han consolidado como referentes en la clasificación supervisada para problemas de morosidad y predicción de cartera vencida.

5.3 Regresión logística

La regresión logística es un modelo estadístico utilizado para analizar la relación entre una variable dependiente binaria y un conjunto de variables independientes o predictoras. Su objetivo es estimar la probabilidad de ocurrencia de un evento, en este caso, que un cliente presente cartera vencida alta, a partir de una función logística que restringe los valores entre 0 y 1.

Matemáticamente, el modelo se expresa como:

$$P(Y=1|X) = \frac{1}{1+ e^{-\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn}}$$

Donde:

P(Y=1|X) probabilidad de que ocurra el evento de interés (en este caso, que el cliente presente cartera vencida alta).

β_0: intercepto o término constante del modelo.

β_i: coeficiente asociado a la variable Xi, que mide el efecto de dicha variable sobre la probabilidad del evento.

Entre sus ventajas destacan su simplicidad, interpretabilidad y solidez estadística, lo que la convierte en un modelo base ideal para el análisis de riesgo crediticio. Sin embargo, su capacidad predictiva puede verse limitada cuando las relaciones entre las variables son no lineales o presentan interacciones complejas.

5.4 Árboles de decisión

Los árboles de decisión son modelos no paramétricos que dividen los datos en subconjuntos más homogéneos a través de reglas de decisión basadas en los valores de las variables predictoras. Cada nodo del árbol representa una condición (por ejemplo, “si el margen operativo es menor a cierto valor”), y las ramas conducen a una decisión o clasificación final.

Este tipo de modelo tiene la ventaja de ser fácil de interpretar y visualizar, ya que las reglas generadas permiten identificar claramente los factores asociados al riesgo de mora. Además, no requiere supuestos sobre la distribución de las variables y puede manejar tanto variables numéricas como categóricas.

No obstante, los árboles de decisión pueden ser inestables o propensos al sobreajuste (overfitting) cuando se construyen con demasiadas divisiones, por lo cual suelen complementarse con técnicas de agregación como el Random Forest.

5.5 Random Forest

El Random Forest es un algoritmo de aprendizaje automático basado en el principio del ensamble (ensemble learning), que combina múltiples árboles de decisión para mejorar la precisión y estabilidad del modelo. Cada árbol se entrena sobre una muestra aleatoria del conjunto de datos y con una selección aleatoria de variables, reduciendo así la varianza y el riesgo de sobreajuste.

La predicción final se obtiene por votación mayoritaria entre los árboles (en clasificación) o mediante promedio (en regresión). Esta técnica ofrece un excelente desempeño en problemas de predicción del riesgo crediticio, ya que es robusta frente a valores atípicos, maneja interacciones complejas entre variables y proporciona medidas de importancia relativa de cada predictor.

En el análisis de cartera, el Random Forest permite identificar los factores más influyentes en la probabilidad de mora y generar rankings de variables según su contribución al riesgo.

5.6 Evaluación del desempeño de los modelos

La calidad de un modelo predictivo no depende únicamente de su ajuste, sino también de su capacidad para clasificar correctamente nuevos casos. Por ello, es fundamental evaluar su desempeño mediante métricas de validación que comparen las predicciones con los valores reales.

Entre las métricas más utilizadas se encuentran:

- Precisión (Accuracy): proporción de casos correctamente clasificados sobre el total.
- Sensibilidad (Recall): capacidad del modelo para identificar correctamente los casos positivos (clientes con mora).
- Especificidad: proporción de clientes sin mora correctamente clasificados.
- Área bajo la curva ROC (AUC): indicador que resume la capacidad del modelo para discriminar entre clientes morosos y no morosos, independientemente del umbral de decisión.

El análisis conjunto de estas métricas permite seleccionar el modelo con mayor capacidad discriminante y estabilidad predictiva, asegurando su aplicabilidad en contextos reales.

5.7 Aplicaciones del aprendizaje automático en la gestión del riesgo de cartera

La aplicación del aprendizaje automático en la gestión del riesgo de crédito ha demostrado resultados significativos en distintos sectores. Estos modelos permiten construir sistemas de apoyo a la decisión que ayudan a las organizaciones a segmentar clientes, optimizar estrategias de cobranza y prevenir la morosidad.

En el caso particular de empresas no financieras como la del presente estudio, dedicada a la comercialización de insumos agropecuarios, el uso de modelos predictivos representa una innovación relevante, ya que posibilita convertir la información interna disponible en conocimiento útil para la gestión del crédito.

De esta forma, la combinación de técnicas estadísticas y de aprendizaje automático constituye una herramienta poderosa para mejorar la eficiencia financiera, reducir el riesgo operativo y fortalecer la sostenibilidad empresarial, al transformar los datos históricos de ventas, cartera y desempeño financiero en un sistema de predicción confiable del riesgo de mora.

6. Metodología

6.1. Contextualización de la Investigación

El presente estudio corresponde a una investigación cuantitativa, aplicada y predictiva, con un enfoque no experimental y transversal.

Es cuantitativa porque se fundamenta en el análisis numérico de variables financieras, comerciales y de ventas para explicar un fenómeno medible: la probabilidad de cartera vencida alta.

Es aplicada porque busca generar un modelo predictivo de utilidad práctica para la empresa objeto de estudio, que contribuya a la toma de decisiones en la gestión de riesgo crediticio.

Finalmente, es transversal porque utiliza los datos disponibles del año 2024 en un único corte temporal, sin intervención experimental sobre las variables.

El diseño metodológico integra tres fases principales:

1. Análisis descriptivo y exploratorio de datos.
2. Modelado predictivo mediante algoritmos de clasificación supervisada.
3. Evaluación comparativa y validación de los modelos.

6.2. Diseño de la Investigación

La población del estudio está conformada por 1800 clientes que registraron ventas durante los años 2023 y 2024. De estos solo 1.175 corresponden a ventas gestionadas desde dos puntos de venta que tiene la compañía, las cuales son transacciones de contado y por lo tanto no hacen parte del estudio. Finalmente se decide realizar este estudio solo para el año más reciente (2024) con lo que se tienen 625 clientes aptos para ser utilizados en la metodología planteada.

Ahora bien, de esos 625 clientes, 141 no cuentan con información financiera reportada en EMIS, por lo que tampoco serán considerados en el análisis. De esta forma, la muestra final queda conformada por 484 clientes.

En la tabla 1, se describen las variables utilizadas en el modelo de estudio con el fin de construir la variable dependiente o variable objetivo del modelo. Según las políticas de la compañía, se consideran clientes de alto riesgo o default, aquellos cuya deuda con más de 30 días de vencimiento representa más del 10% del total, o que mantienen un saldo pendiente de pago superior al 10%, con lo que en la presente base el 55,6% de los clientes (269) se clasifican como cumplidos, mientras que el 44,4% (215) presentan comportamiento de default o impago.

Bajo este criterio, se clasifica para el cliente como de riesgo alto (1) y, en caso contrario, como de riesgo bajo (0).

Esta variable binaria refleja la condición de riesgo de mora del cliente y se utilizará como el resultado a predecir en los modelos de clasificación.

Tabla 1.

Descripción de las variables

Nombre de la variable	Clasificación	Clasificación según su naturaleza		Escala de medición	Argumento
ROE	**Financiera:** Provienen de la información reportada por los clientes a la Superintendencia de Sociedades y que EMIS almacena en su plataforma.	Cuantitativa	Continua	Razón	Indicador financiero que permite medir la rentabilidad de la empresa conforme al patrimonio de los accionistas. $ROE = \frac{Utilidad\ neta}{Patrimonio\ Neto} x100$
ROA		Cuantitativa	Continua	Razón	Indicador financiero que permite medir la rentabilidad de la empresa conforme a los activos totales. $ROA = \frac{Utilidad\ neta}{Activos\ totales} x\ 100$
Margen Neto		Cuantitativa	Continua	Razón	Es la ganancia neta de una empresa en relación con sus ingresos totales. $MN = \frac{Utilidad\ neta}{ingresos} x100$
Margen Operacional		Cuantitativa	Continua	Razón	Es el indicador financiero que mide la utilidad de la empresa $MO = \frac{utilidad\ \ Operacional}{ventas} x100$
Endeudamiento		Cuantitativa	Continua	Razón	Es un indicador fundamental para medir la salud financiera de una empresa. Ayuda a determinar si la empresa tiene una carga de deuda sostenible o si está sobreendeudada.
Eficiencia		Cuantitativa	Continua	Razón	Mide la capacidad de la empresa para utilizar sus recursos de manera efectiva, maximizando los ingresos y minimizando los gastos para generar valor a largo plazo
Estado de situación Financiera Activos Pasivos Patrimonio		Cuantitativa	Continua	Nominal	Permite ver la situación financiera de la empresa en un momento específico del tiempo

Estado de Resultados Ingresos Gastos Ganancia Bruta Ganancia Operacional		Cuantitativa	Continua	Nominal	Informe financiero que detalla los ingresos, costos y gastos de la empresa durante un periodo de tiempo determinado
Ventas Total Skus Ventas totales de la empresa Monto o ticket Promedio Costos Rentabilidades promedio y totales	Ventas	Cualitativa	Continua	Nominal	Hacen referencia a las métricas del negocio en relación con sus ventas, los cuales son brindados por la empresa y obtenidos de la ERP SAP.
Cartera Total Abonado Total Vencido Total de importe o valor de factura Días promedio de retraso Porcentaje de facturas vencidas	Cartera vencida	Cuantitativa	Discreta	Razón	Indicadores asociados al comportamiento de pago del cliente, en este caso solo relaciona los clientes con cartera vencida con corte al 31 de diciembre del 2023 y con corte al 31 de diciembre del 2024.
Clasificación comercial y geográfica de los clientes Tipo de cliente Nombre del vendedor Zona geográfica: Departamento/ Municipio		Cualitativa		Ordinal	Variables categóricas de caracterización comercial y geográfica de los clientes.

Nota: esta tabla describe las variables objeto de estudio.

6.3. Descripción de datos a utilizar

Tabla 2.

Muestreo		
Información		Fuente
Datos	Libres	Empresa privada a la cual pertenece uno de los autores y requiere confidencialidad.
Tamaño de muestra	484	información interna de la empresa y corresponden al año 2024
Número de variables	73	

6.4. Preparación y limpieza de los datos

Para este estudio, se realiza un proceso de depuración y transformación de datos, que incluye:

6.4.1 Cargue y depuración de bases de datos

El tratamiento de los datos inició con un filtro de los registros de ventas para conservar únicamente aquellos clientes con presencia en la base de datos de cartera.

Dado que la información de ventas estaba disponible a nivel de producto individual por factura, fue necesario agregar los datos por cada uno de los clientes. Este proceso implicó el cálculo de métricas agregadas, tales como: el total de facturas, el conteo de productos y marcas distintas, las cantidades totales, el monto total vendido, el monto promedio por venta, y los promedios de precios, costos y rentabilidad.

Se aplicó un proceso análogo para la base de cartera. A partir de la información detallada

por factura, se calcularon variables por cliente, incluyendo: el valor total facturado, el monto total pagado, el saldo adeudado, los días promedio de mora (posteriores al vencimiento), los valores mínimos y máximos de mora, y los porcentajes de montos vencidos y pagados sobre el total de la deuda.

Posteriormente, se incorporó la base de estados financieros, la cual fue condicionada para incluir únicamente la información de los clientes ya filtrados en la base de cartera.

Para el tratamiento de las variables financieras, se adoptó el criterio de eliminar aquellas con un porcentaje de datos faltantes (NA) superior al 10%. Las variables descartadas bajo este criterio fueron: Rotación de Inventario (59,55% faltante), Rotación de Cuentas por Cobrar (56,26%), Rotación de Cuentas por Pagar (56,26%), Efectivo o Equivalentes (56,06%) y Capital Suscrito (56,06%).

Para las variables financieras con menos del 10% de datos faltantes, se utilizó un método de imputación basado en clustering. Se aplicó un modelo k-means para segmentar a los clientes en 4 grupos (clústeres) según su comportamiento financiero. Posteriormente, cada valor faltante fue reemplazado por la mediana del clúster al que pertenecía el cliente. Las variables sometidas a este proceso fueron Razón de Liquidez (x) y Rotación del Capital de Trabajo (x).

Tras estos procesos, se validó la integridad de la base de datos, confirmando que se encontraba completa y sin valores faltantes.

6.4.2 Creación variable predictora: default de pago

El modelo clasifica a los clientes como cumplidos (0) o en default (1) según si tienen más del 10 % de su factura vencida por más de 30 días o si más del 10 % de su saldo total está pendiente de pago. De esta forma, el 55.6 % de los clientes (269) se consideran cumplidos, mientras que el

44.4 % (215) presentan comportamiento de default, reflejando una distribución relativamente equilibrada entre ambos grupos. Finalmente, se realizó una revisión de la base de datos consolidada y se asignó el formato correspondiente (categórico o numérico) a cada variable, preparándolas para la fase de análisis descriptivo.

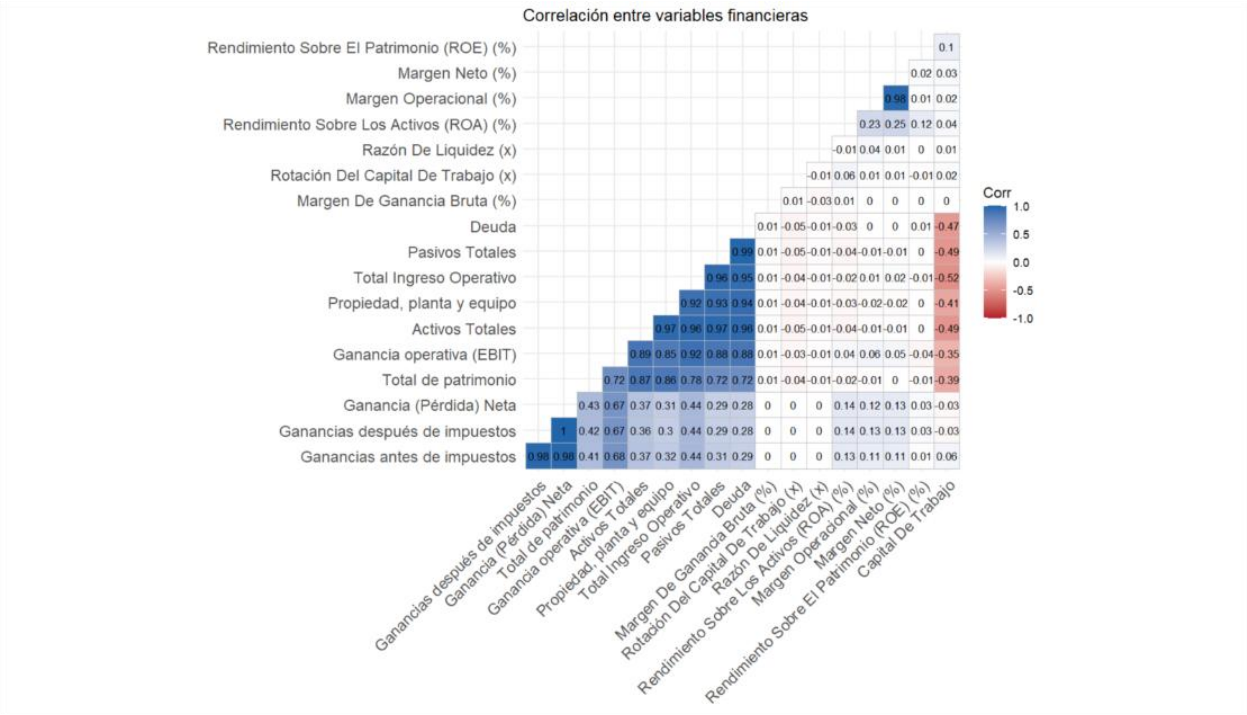
6.5. Análisis, Interpretación y validación de resultados

6.5.1 Selección inicial de variables

El conjunto de datos original comprendía más de 70 variables potenciales. Dado este alto número y para mitigar la multicolinealidad en las fases de modelado, se realizó un análisis de correlación (utilizando el coeficiente de Pearson) como paso inicial de selección. El objetivo fue identificar y excluir, dentro de cada grupo de variables, aquellas que presentaran correlaciones elevadas, conservando únicamente las más relevantes para el modelo.

6.5.1.1 Variables Financieras

Figura 1. Correlación entre variables financieras



En el grupo de variables financieras se identificaron varios conglomerados con altas correlaciones:

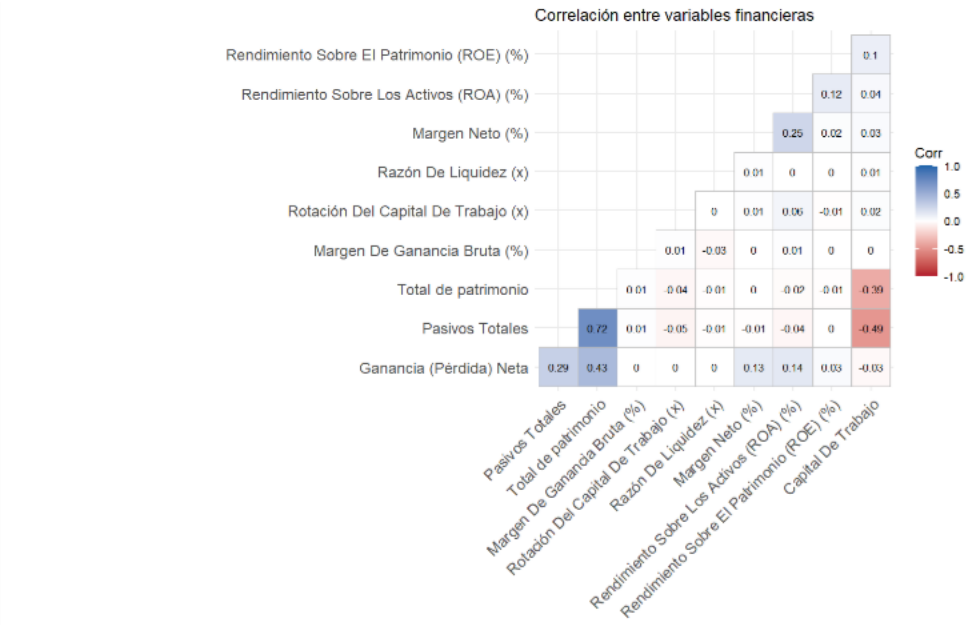
- Margen Operacional y Margen Neto (r = 0.975).
- Las variables de ganancias (Ganancias antes de impuestos, Ganancias después de impuestos y Ganancia Neta), con correlaciones en el rango de r=0.98, a r=0.998.
- Las cuentas de balance (Activos Totales, Propiedad, Planta y Equipo, Pasivos Totales y Deuda), con correlaciones de r entre 0.932 a 0.986.
- Total de Ingreso Operativo y Ganancia Operativa (EBIT), con correlaciones de r entre 0.851 y 0.959.
- El Total de Patrimonio mostró una correlación alta con Activos Totales y Propiedad, Planta y Equipo, pero moderada con la Deuda r aprox 0.716.

En función de este análisis, se decidió excluir del modelo las variables: Margen

Operacional, Ganancias antes de impuestos, Ganancias después de impuestos, Propiedad, Planta y Equipo, Deuda, Ganancia Operativa (EBIT), Total de Ingreso Operativo y Activos Totales.

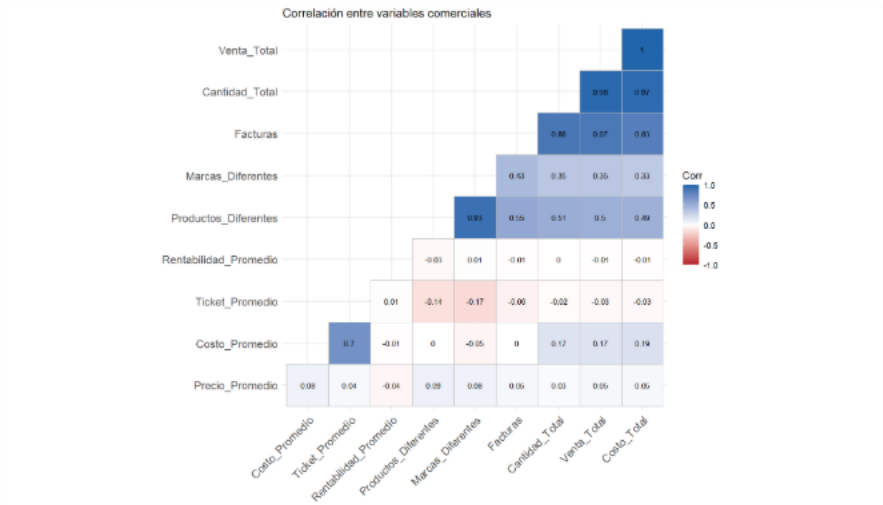
Respecto a estas últimas, aunque estaban fuertemente relacionadas con Pasivos Totales, se optó por conservar la variable de Pasivos Totales, considerando que esta tendría una mayor implicación teórica sobre la variable default.

Figura 2. Correlación entre variables financieras luego de las exclusiones anteriores



6.5.1.2 Variables Comerciales (Ventas)

Figura 3. Correlación entre variables comerciales

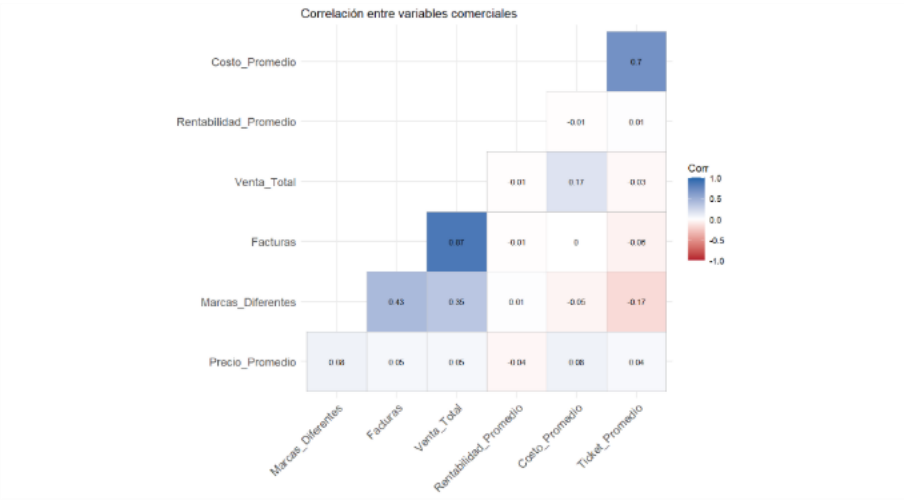


Para las variables comerciales, se observó un comportamiento similar:

- Venta Total y Costo Total presentaron una correlación casi perfecta (r = 0.997), indicando redundancia de información.
- Cantidad Total mostró una alta correlación tanto con Venta Total (r = 0.979) como con Costo Total (r = 0.969).
- Facturas y Costo Total muestran una correlación alta (r = 0.835)
- Productos Diferentes y Marcas Diferentes también presentaron una asociación muy elevada (r = 0.932).

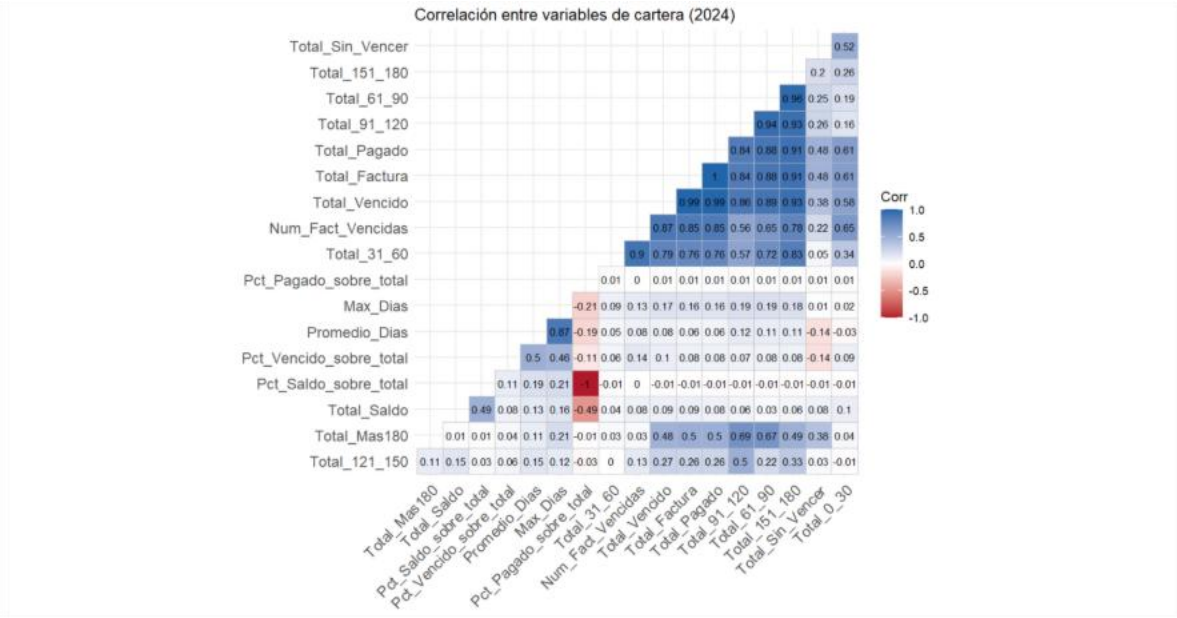
Debido a estas relaciones, se procedió a excluir del análisis las variables Costo Total, Cantidad Total y Productos Diferentes, evitando así la redundancia y simplificando el modelo.

Figura 4. Correlación entre variables comerciales después de la reducción de variables anterior



6.5.1.3 Variables de Cartera

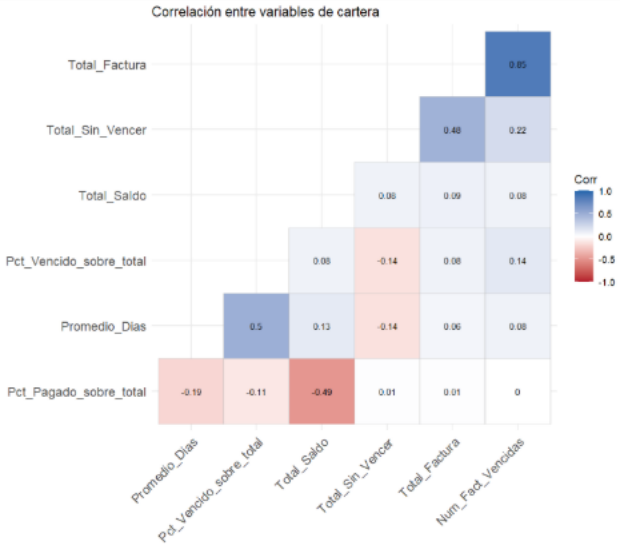
Figura 5. Correlación entre variables de cartera



El análisis de correlación en las variables de cartera también reveló altos índices de multicolinealidad, particularmente entre las variables agregadas como Total_Factura, Total_Pagado y Total_Vencido, así como entre diversos rangos de vencimiento y las medidas de días promedio.

Para evitar la duplicidad de información, se realizó una selección de las variables consideradas más representativas, conservando: Total_Factura, Total_Saldo, Total_Sin_Vencer, Total_151_180, Num_Fact_Vencidas, Promedio_Dias, Pct_Pagado_sobre_total y Pct_Vencido_sobre_total. Este subconjunto permite capturar la exposición al riesgo de incumplimiento sin introducir redundancia.

Figura 6. Correlación entre variables de cartera después de la reducción de variables anterior



6.5.2 Tablas resumen de resultados análisis estadísticos

Es un comparativo de las principales métricas condicionado por default de pago en el que se compara para cada uno de los bloques de variables, una vez depurada la base de variables, (media, mediana, mínimo, máximo, cuartiles y desviación estándar) para cada variable seleccionada, agrupando los resultados por la variable objetivo (default_pago).

6.5.2.1. Variables financieras

Figura 7. Tabla resumen financiero por default_pago

Resumen Financiero por default_pago

default_pago	Variable	Media	Mediana	Min	Max	Desv
0	Capital De Trabajo	583.263	635.910	-228109.430	74942.740	18510.763
1	Capital De Trabajo	-442.251	619.680	-209137.240	30034.320	16196.539
0	Ganancia (Pérdida) Neta	362.639	75.080	-9445.530	46594.090	3389.976
1	Ganancia (Pérdida) Neta	-164.643	100.290	-36178.510	21531.850	4134.992
0	Margen De Ganancia Bruta (%)	0.150	0.177	-13.456	1.000	0.901
1	Margen De Ganancia Bruta (%)	0.180	0.171	-3.166	1.000	0.303
0	Margen Neto (%)	-0.108	0.017	-23.125	10.925	1.658
1	Margen Neto (%)	-0.051	0.012	-3.969	1.719	0.484
0	Pasivos Totales	10806.223	1750.470	0.000	613986.010	41873.242
1	Pasivos Totales	24539.307	4828.930	0.000	1517812.760	110319.469
0	Razón De Liquidez (x)	12.787	1.550	0.060	2465.980	150.322
1	Razón De Liquidez (x)	2.193	1.390	0.010	64.300	5.436
0	Rendimiento Sobre El Patrimonio (ROE) (%)	0.355	0.060	-12.274	60.443	3.909
1	Rendimiento Sobre El Patrimonio (ROE) (%)	0.776	0.058	-26.435	207.620	14.375
0	Rendimiento Sobre Los Activos (ROA) (%)	0.062	0.024	-0.925	3.570	0.300
1	Rendimiento Sobre Los Activos (ROA) (%)	0.014	0.016	-0.743	1.500	0.149
0	Rotación Del Capital De Trabajo (x)	12.879	3.440	-90.990	1158.170	83.576
1	Rotación Del Capital De Trabajo (x)	6.590	4.190	-679.440	634.840	68.662
0	Total de patrimonio	11122.353	1878.920	-17436.630	545258.310	45206.355
1	Total de patrimonio	10783.559	3175.380	-5380.090	510741.500	38268.763

En esta figura 7 se observa que en general, los clientes sin default presentan mejores resultados en liquidez, rentabilidad y capital de trabajo, reflejando una posición financiera más sólida frente a los clientes en default, quienes poseen mayores niveles de endeudamiento y menor capacidad para la generación de utilidades. Respecto a la liquidez, se tiene que el capital de trabajo promedio de los clientes sin default es positivo (583.263), mientras que en los clientes con default es negativo (-442.251), lo que evidencia dificultades para cubrir sus obligaciones de corto plazo,

de igual forma sucede con la razón de liquidez la cual es mayor en clientes sin default (12,8x) frente a los de default (2,2x), confirmando que efectivamente estos primeros tienen una capacidad superior para cumplir sus compromisos corto plazo.

En cuanto a la rentabilidad, en términos de ganancia neta, los clientes sin default registran una media positiva (362.639), mientras que los clientes en default presentan una pérdida promedio (-164.643), igualmente el margen neto también es más favorable en los clientes cumplidos (-0,108 vs. -0,051), aunque ambos grupos presentan variabilidad significativa. En el ROA (rendimiento sobre los activos), los clientes sin default muestran una rentabilidad promedio superior (0,062) frente a los de default (0,014). No obstante, el ROE (rendimiento sobre el patrimonio) es mayor en los clientes con default (0,776) que en los cumplidos (0,355), posiblemente por niveles más altos de apalancamiento.

Ahora bien, el endeudamiento es un punto importante en la comparación de estos clientes, pues los pasivos totales promedio son más del doble en los clientes con default (24.539) respecto a los sin default (10.806), lo que sugiere una estructura financiera más dependiente de terceros.

Pese a ello, el patrimonio total promedio es similar entre ambos grupos, lo que indica que las diferencias de riesgo están más asociadas al manejo del pasivo que a la base patrimonial.

Finalmente, al revisar la eficiencia operativa se tiene que la rotación del capital de trabajo es mayor en los clientes sin default (12,9x) frente a los de default (6,6x), evidenciando una mejor gestión operacional.

De esta manera, se tiene que los clientes en default tienden a presentar peor liquidez, más probabilidad de tener pérdidas netas, menor rentabilidad sobre los activos, mayor endeudamiento y posible sobre apalancamiento.

6.5.2.2. Variables comerciales (ventas)

Figura 8. Tabla resumen comercial (ventas) por default_pago

Resumen Comercial (Ventas) por default_pago

default_pago	Variable	Media	Mediana	Min	Max	Desv
0	Costo_Promedio	940065.81	341849.03	11400.00	46445798.75	3180203.68
1	Costo_Promedio	1507685.71	640475.52	32684.30	32947400.20	3569533.59
0	Facturas	38.24	5.00	1.00	1806.00	124.35
1	Facturas	56.81	25.00	1.00	2529.00	178.28
0	Marcas_Diferentes	7.06	3.00	1.00	34.00	7.94
1	Marcas_Diferentes	11.31	10.00	1.00	36.00	8.51
0	Precio_Promedio	179370.93	138427.38	7808.00	2490917.33	231803.98
1	Precio_Promedio	188219.38	178908.94	7483.00	891517.83	103788.32
0	Rentabilidad_Promedio	-2400.77	0.22	-68027.38	0.55	8406.63
1	Rentabilidad_Promedio	-13426.27	0.21	-1543965.94	0.39	110927.16
0	Ticket_Promedio	387454.96	46131.22	1054.39	15860632.80	1471403.34
1	Ticket_Promedio	497673.82	31920.45	522.44	36450000.00	2730330.45
0	Venta_Total	160713665.58	5088581.00	20000.00	11579640124.47	758376324.93
1	Venta_Total	286313394.15	52867329.00	76000.00	14930174789.00	1294204830.49

En la figura 8, este resumen comercial revela una paradoja clave: los clientes que incumplen (default_pago = 1) no son clientes pequeños o marginales, sino que, de hecho, son comercialmente más grandes y están más integrados que los que pagan. La evidencia es clara en las medianas: el cliente típico que incumple tiene 5 veces más facturas (Mediana de 25 vs 5) y compra más de 3 veces la cantidad de marcas diferentes (Mediana de 10 vs 3). A pesar de este alto volumen, la rentabilidad típica (Mediana) es casi idéntica en ambos grupos.

El verdadero problema se oculta en las medias: el grupo de default incluye clientes muy

grandes y extremadamente volátiles que, aunque elevan la "Venta_Total" promedio (2.86 Billones vs 1.60 Billones), son catastróficamente no rentables (Rentabilidad Promedio de -13,426), definiendo el perfil de riesgo de todo el segmento.

6.5.2.3. Variables de cartera

Figura 9. Tabla resumen cartera por default_pago

Resumen Cartera por default_pago

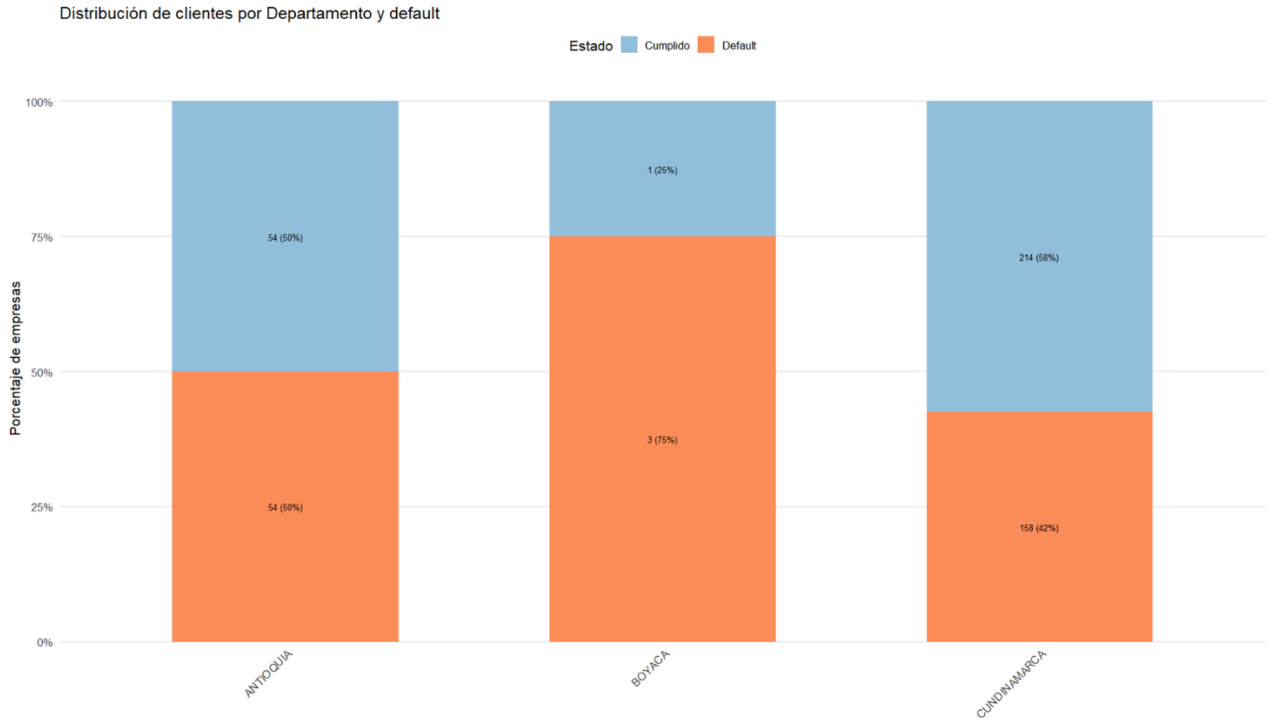
default_pago	Variable	Media	Mediana	Min	Max	Desv
0	Num_Fact_Vencidas	29.175	1.000	0.000	2123.000	139.197
1	Num_Fact_Vencidas	77.107	31.000	1.000	4227.000	295.132
0	Pct_Pagado_sobre_total	0.998	1.000	0.910	1.000	0.010
1	Pct_Pagado_sobre_total	0.980	1.000	0.337	1.000	0.061
0	Pct_Vencido_sobre_total	0.472	0.513	0.000	1.000	0.443
1	Pct_Vencido_sobre_total	0.950	1.000	0.433	1.000	0.108
0	Promedio_Dias	1.673	0.750	-61.000	81.000	14.370
1	Promedio_Dias	59.431	44.794	4.833	326.688	50.512
0	Total_Factura	200727417.458	5644444.300	20000.000	15671956285.710	1013787283.916
1	Total_Factura	471077509.023	88562805.140	153240.000	26911766948.650	2324434297.329
0	Total_Saldo	596957.379	0.000	0.000	34944857.290	3112483.365
1	Total_Saldo	5428955.120	0.000	0.000	263981395.530	22205864.795
0	Total_Sin_Vencer	65221907.024	555386.000	0.000	2309962555.390	249139120.715
1	Total_Sin_Vencer	18264540.893	0.000	0.000	1800485581.150	125629174.720

La figura 9 del resumen de cartera expone la diferencia más crítica entre los dos grupos: el incumplimiento no es una cuestión de lentitud, sino de un cese total de pagos. El cliente típico que cumple (Mediana) tiene solo 1 factura vencida con menos de 1 día de mora. Por el contrario, el cliente típico que incumple (default_pago = 1) tiene 31 facturas vencidas con una mora promedio de 45 días (Promedio_Dias), y el 100% de su saldo total ya se considera vencido

(Pct_Vencido_sobre_total). Agravando esto, las empresas que incumplen no son clientes pequeños; sino las grandes que en promedio, su facturación total es más del doble (Total_Factura) y dejan un saldo pendiente (Total_Saldo) 9 veces mayor que el de los clientes que pagan.

6.5.4 Distribuciones estadísticas para las variables categóricas

Figura 10. Distribución de clientes por segmentación y default



La figura 10 presenta la distribución de las 484 empresas de la muestra, clasificándolas por su segmento de negocio (AGROINDUSTRIA, OTRO, PUNTO DE VENTA, RETAIL) y su estado de cartera (Cumplido o Default). Este análisis revela que el riesgo de incumplimiento no es homogéneo en la cartera, sino que varía significativamente entre los segmentos.

La tasa de incumplimiento promedio de la cartera es del 44.4% (215 de 484 clientes). Sin embargo, al analizar por segmento, se observan los siguientes hallazgos:

Segmento OTRO: Este segmento, es el más pequeño (n=12) y los resultados 50/50 no son

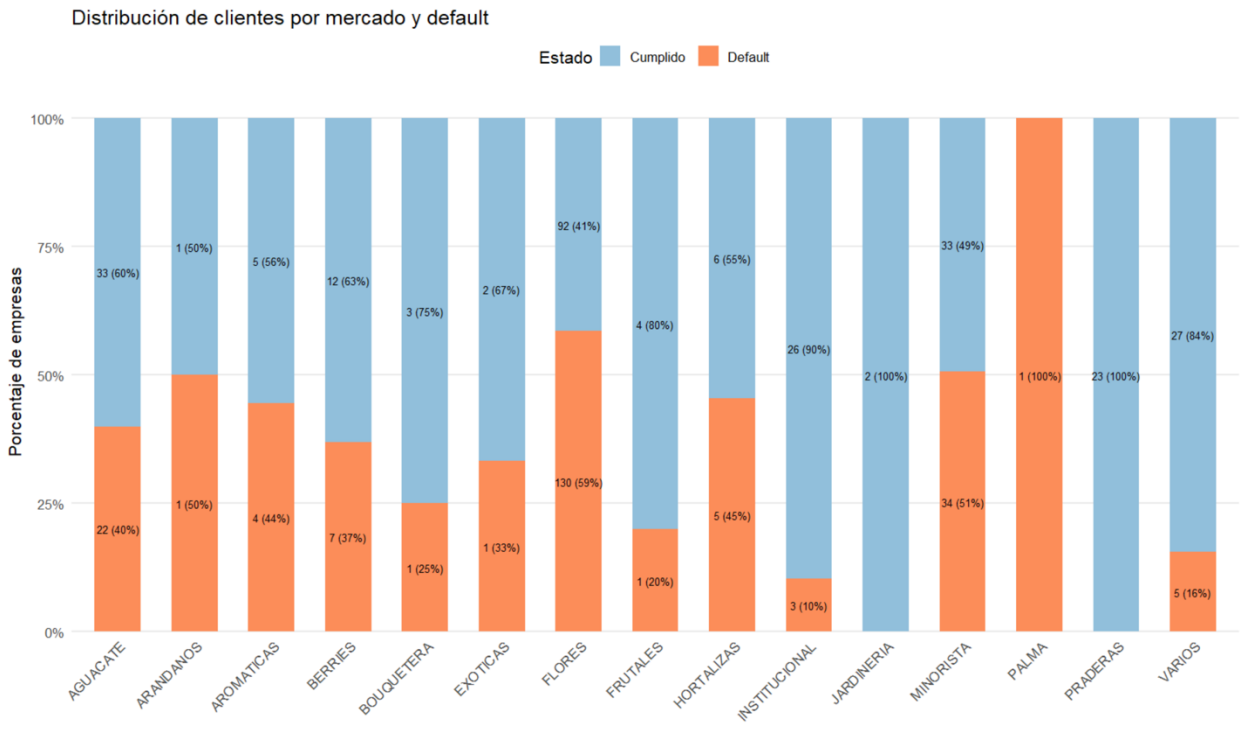
estables, se puede tomar como un segmento anecdótico.

Segmento AGROINDUSTRIA: Este es el segmento más grande de la cartera (n=274) y tiene la tasa de default más alta (58%). Lo más crítico de este segmento es su impacto absoluto: sus 158 clientes en default representan el 73.5% del total de la cartera vencida (158 de 215).

Segmento RETAIL: Este segmento (n=58) tiene más tasa de incumplimiento (55%), que tasa de cumplimiento (45%).

Segmento PUNTO DE VENTA: Este segmento (n=140) es el de mejor comportamiento y menor riesgo de toda la cartera. Presenta una tasa de default de solo el 14%, muy por debajo del promedio. A pesar de ser el segundo segmento más grande, solo contribuye con 19 de los 215 clientes en default.

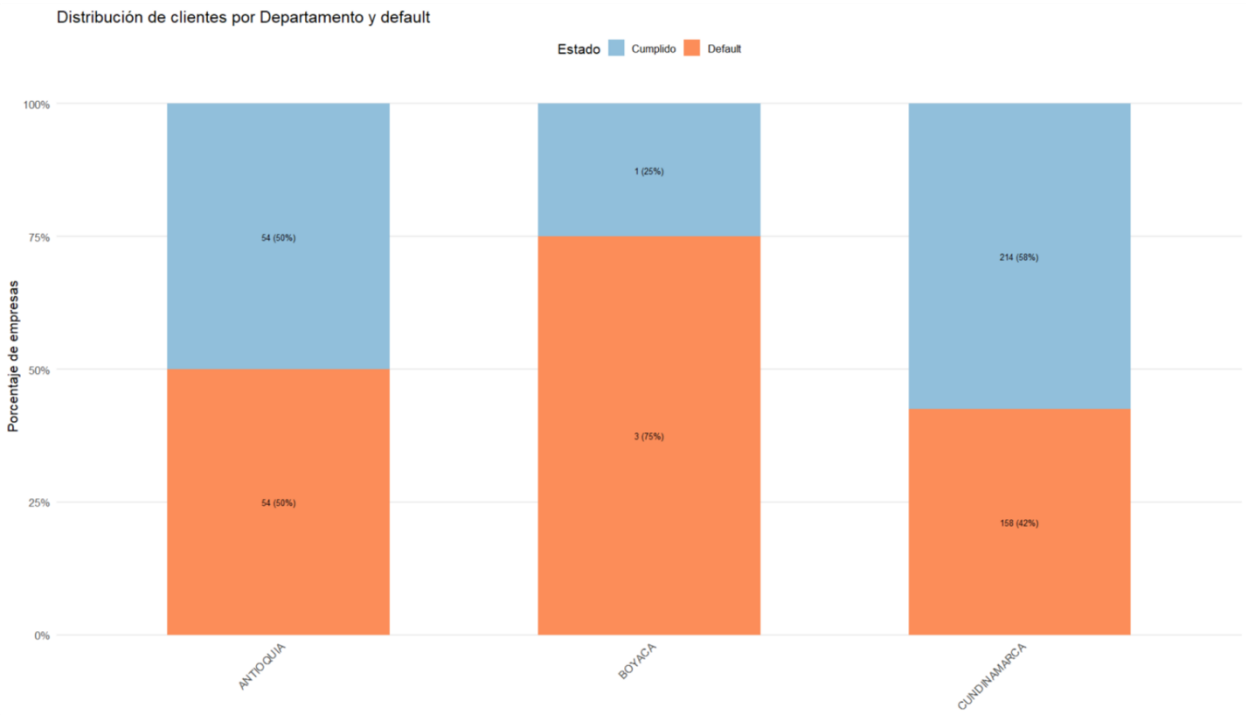
Figura 11. Distribución de clientes por mercado y default



De la figura 11, podemos concluir que de los 215 impagos, la gran mayoría (87.4%) se

concentra en solo tres mercados: “Flores” (130 impagos), “Minorista” (34 impagos) y “Aguacate” (22 impagos). El segmento de “Flores” representa por sí solo el mayor problema de cartera, contribuyendo con el 60.5% del número total de impagos.

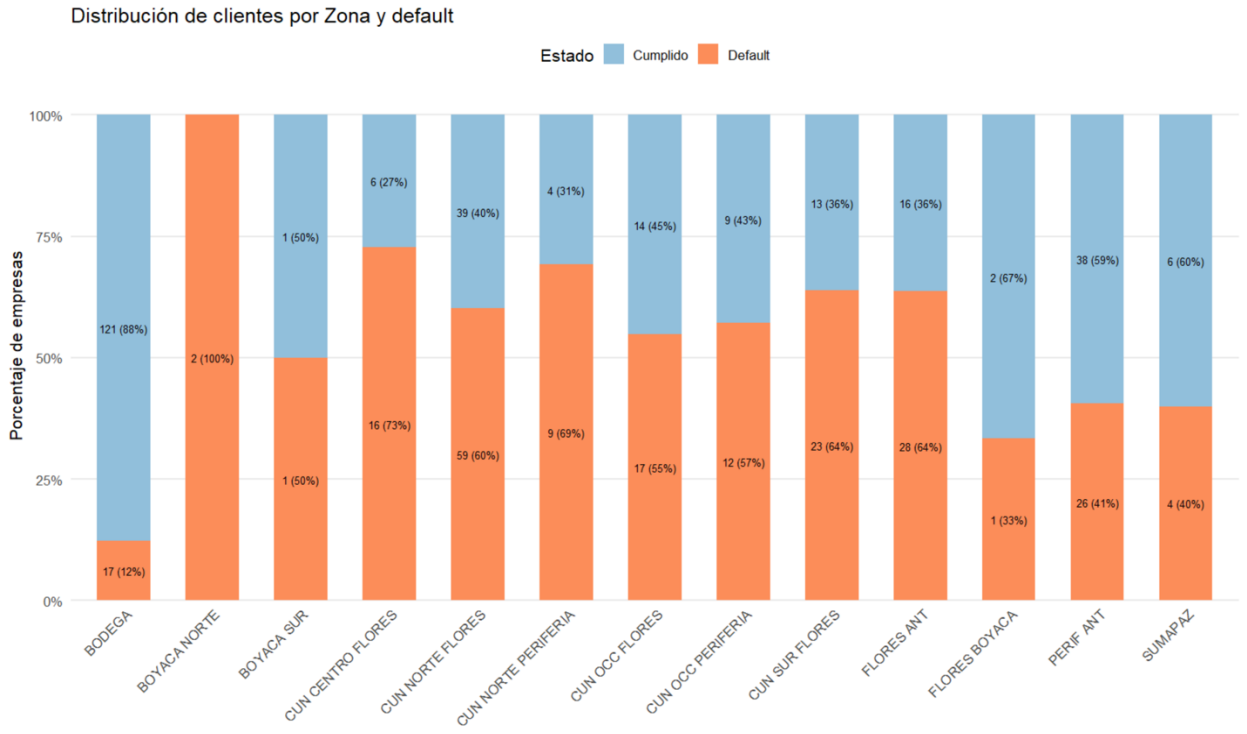
Figura 12. Distribución de clientes por departamento y default



En la figura 12 el volumen de la cartera está concentrado en el departamento de Cundinamarca, que representa el 77% del total de clientes (372 empresas) y, consecuentemente, el 73.5% del número absoluto de impagos (158 defaults). Sin embargo, el departamento de Antioquia, aunque más pequeño (108 clientes), es más riesgoso ya que el 50% son impagos.

El departamento de Boyacá, con solo cuatro clientes en total, es estadísticamente irrelevante para un análisis, y sus datos (75% de default, 25% cumplidos) deben ser tratados como anecdóticos.

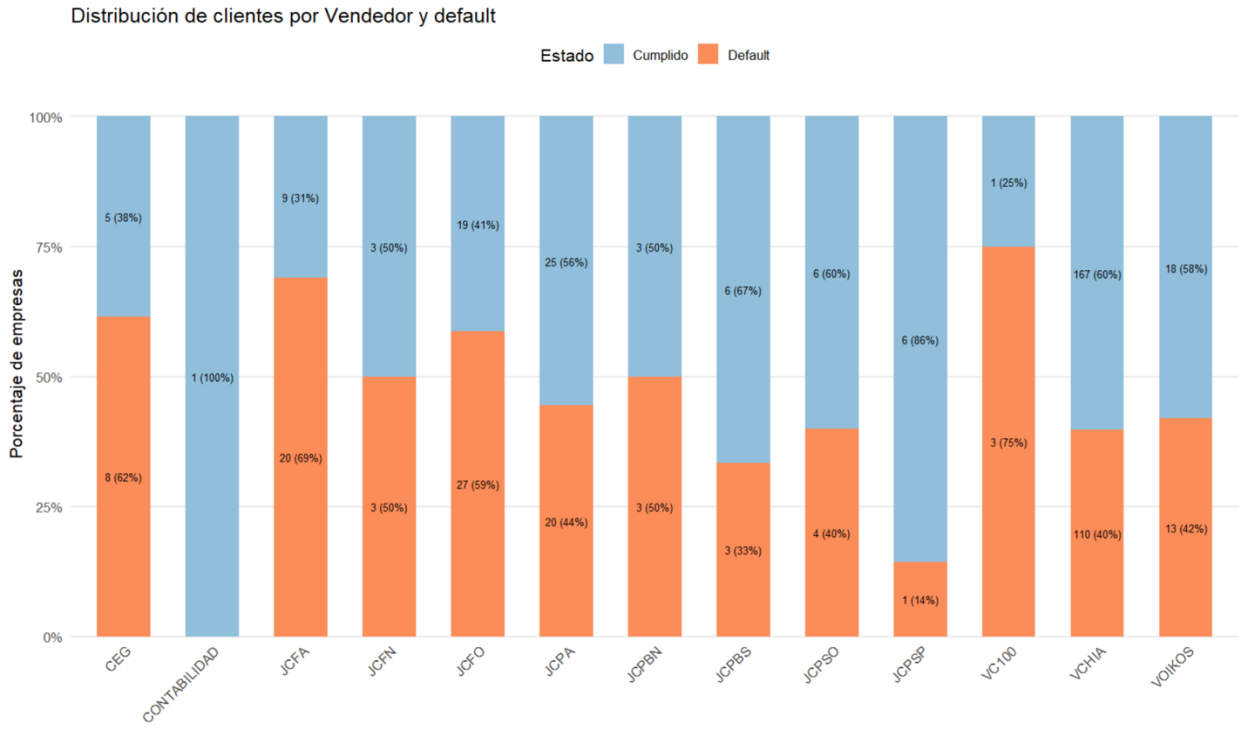
Figura 13. Distribución de clientes por zona y default



En la figura 13 se observa que el riesgo está concentrado en el sector “Flores”: CUN CENTRO FLORES: 72.7%, CUN OCC FLORES: 54.8%, CUN SUR FLORES: 63.9%, FLORES ANT: 63.6%, CUN NORTE FLORES: 60.2%. Esto significa que el 86.5% de todos los incumplimientos de la cartera provienen de este único segmento de mercado. En contraposición directa, la zona “BODEGA” es el "modelo de oro" de la cartera. Es el segmento más grande, 87,7% cumplido, 12,3% default, presenta un desempeño sólido y con criticidad operativa por volumen total 138 empresas (28.5% del total), pero mantiene una tasa de impago muy baja del 12.3% (solo 17 defaults). Aporta el 28.5% de los clientes pero solo el 7.9% de los incumplimientos.

El análisis anterior, que señalaba al departamento de Antioquia como una región de alto riesgo (50% default), puede ser explicado de la siguiente manera: el riesgo de Antioquia se compone de un segmento de riesgo mediano (“PERIF ANT” con 40.6%) y un segmento desastroso (“FLORES ANT” con 63.6%). Entonces el problema no es tanto el departamento Antioquia; el problema es el segmento “Flores” a nivel nacional.

Figura 14. Distribución de clientes por vendedor y default



De la figura 14 se pude extraer que el vendedor con iniciales (“VCHIA”) gestiona la cartera más grande (277 empresas, 57% del total de la empresa) con una tasa de impago (40%) que es, de hecho, mejor que el promedio de la cartera (44.4%). Sin embargo, debido al volumen

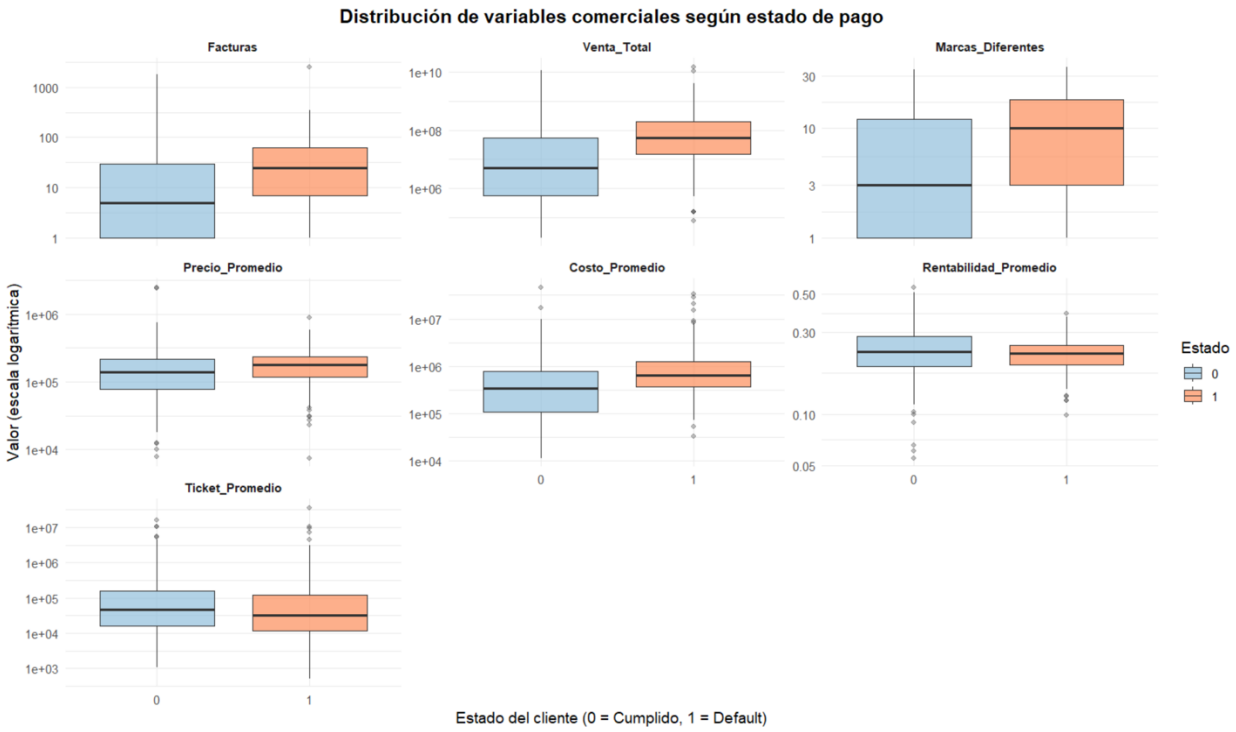
masivo de su portafolio, genera la mayoría absoluta de los incumplimientos: 110 impagos, que representan el 51.2% del total de impagos de la compañía.

Los otros dos vendedores con alto riesgo son vendedor “JCFO”: Este vendedor tiene la segunda cartera más problemática. Aunque su cartera es mucho más pequeña (46 clientes), su tasa de impago es alarmantemente alta, del 59%. Esto genera 27 impagos, convirtiéndolo en el segundo mayor contribuyente a las pérdidas (12.5% del total). Y vendedor “JCFA”: de 29 empresas en su portafolio, 20 tienen impago que corresponden al 69%.

6.5.5. Boxplots comparativos con el estado de pago (default)

6.5.5.1. Boxplots comparativos por default con variables comerciales

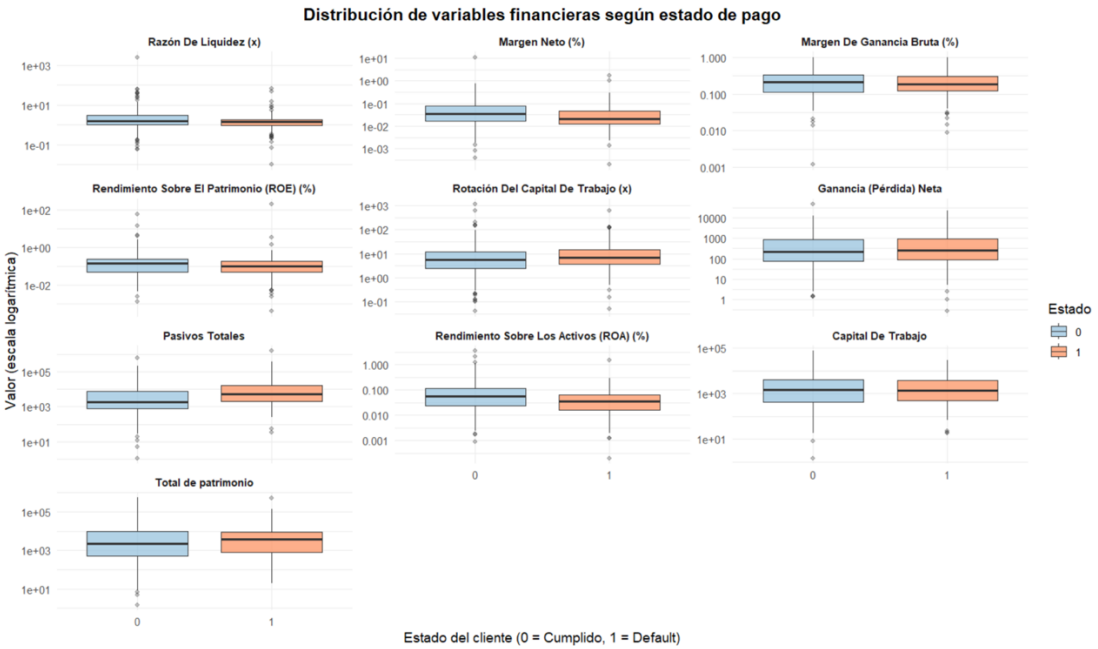
Figura 15. Boxplots de distribución de variables comerciales según estado de pago



De la figura 15, los boxplots confirman visualmente la paradoja central de la cartera: los clientes que incumplen (Estado = 1) no son clientes marginales, sino que representan un volumen comercial mayor que los que cumplen (Estado = 0). Esto es evidente en las distribuciones (Boxplots) significativamente más altas para Facturas, Venta_Total y Marcas_Diferentes, indicando que el grupo de default compra más, con más frecuencia y una mayor variedad de productos. Sin embargo, el hallazgo más revelador es que, a pesar de su alto volumen, su Ticket_Promedio típico (mediana) es en realidad más bajo que el del grupo que cumple. Esto, combinado con una Rentabilidad_Promedio típica casi idéntica, dibuja un perfil de riesgo claro: el cliente que incumple es uno de alto volumen, pero de transacciones recurrentes, cuya gestión operativa es más compleja y finalmente colapsa.

6.5.5.2. Boxplots comparativos por default con variables financieras

Figura 16. Boxplots de distribución de variables financieras según estado de pago

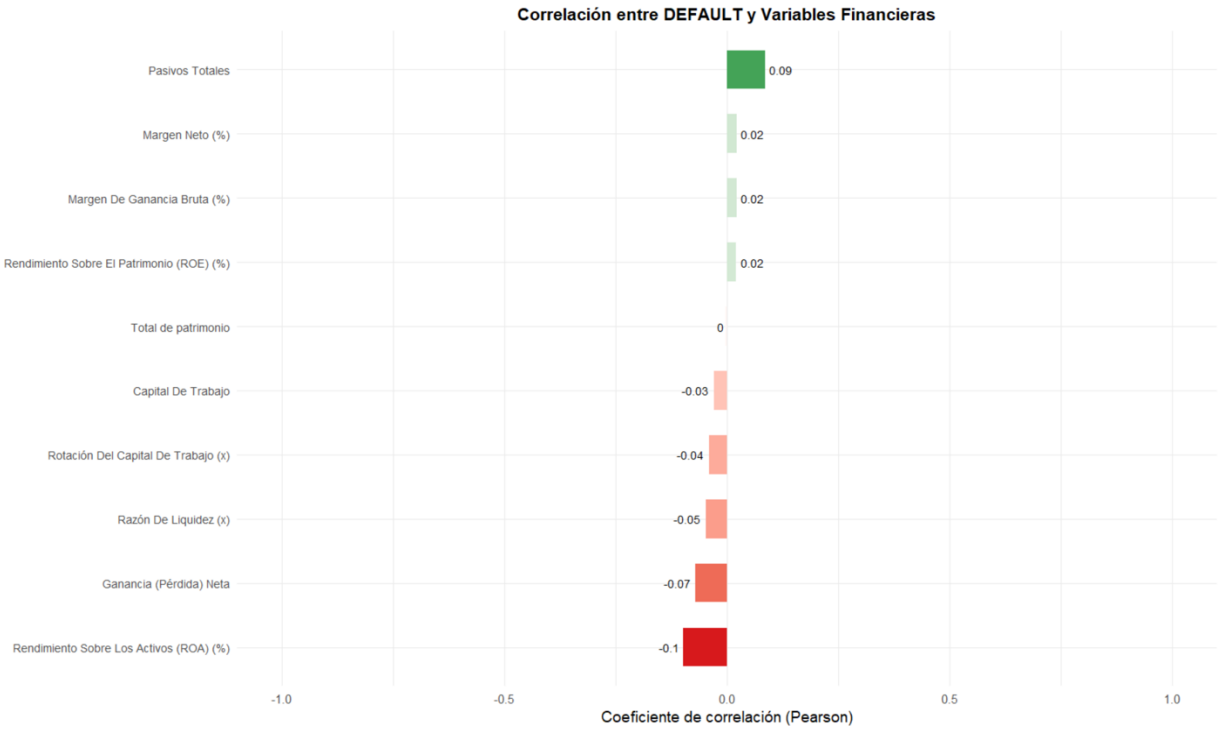


De la figura 16. Estos gráficos financieros exponen la causa raíz del incumplimiento: no es un problema de rentabilidad operativa, sino de estructura de capital y liquidez. Mientras que el Margen de Ganancia Bruta es casi idéntico en ambos grupos, los clientes que incumplen (Estado = 1) muestran una distribución de Pasivos Totales significativamente más alta, indicando un posible sobreendeudamiento. Este apalancamiento se traduce directamente en una salud financiera precaria: el grupo de default tiene un Capital de Trabajo y una Razón de Liquidez drásticamente peores, lo que finalmente destruye su rentabilidad neta (visible en Ganancia Neta y ROE).

6.5.6. Correlaciones con el estado de pago (default)

6.5.6.1. Correlaciones entre default y variables financieras

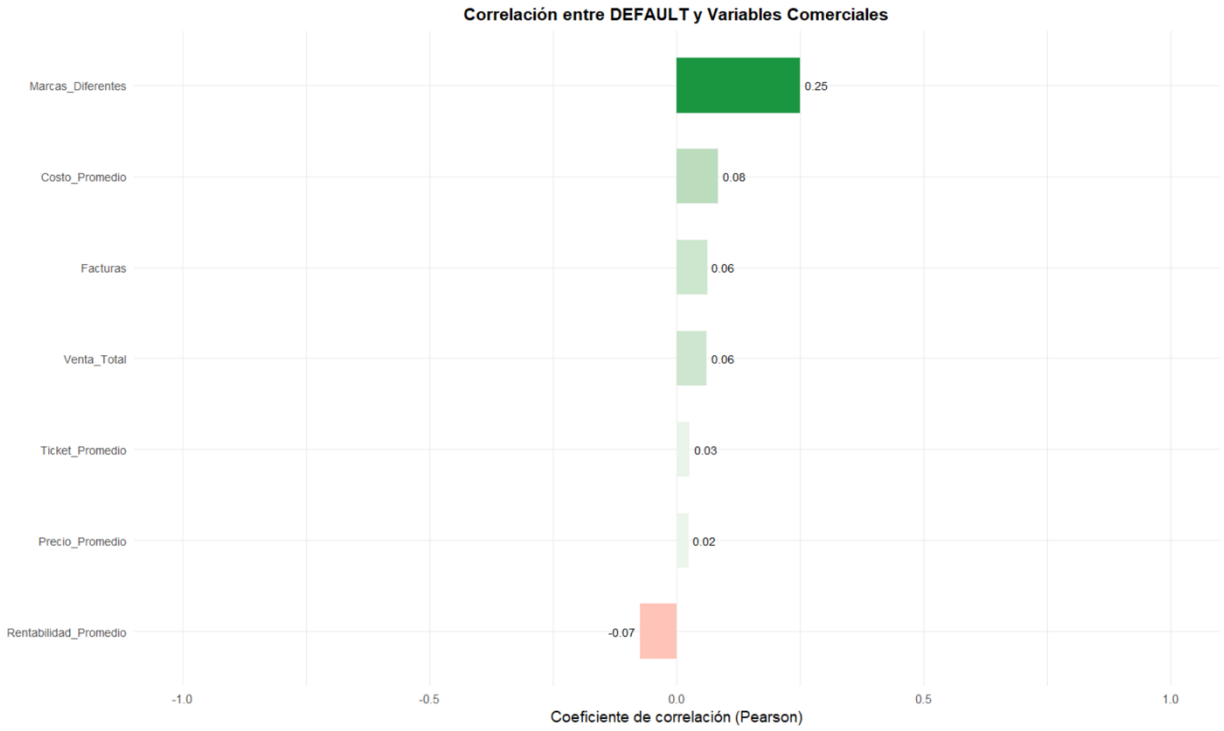
Figura 17. Gráfica de correlación entre Default y variables financieras



Esta la Figura 17. Gráfica de correlación entre Default y variables financieras demuestra que ninguna variable financiera por sí sola es un predictor decisivo, ya que todas las correlaciones son extremadamente débiles (la más fuerte apenas llega a -0.1). Sin embargo, la *dirección* de estas débiles relaciones es coherente con los hallazgos anteriores: el impago (DEFAULT) está positivamente correlacionado con los Pasivos Totales (+0.09), lo que refuerza que más deuda equivale a más riesgo. Inversamente, las correlaciones negativas más fuertes, aunque débiles, son el Rendimiento Sobre Los Activos (ROA) (-0.1) y la Ganancia (Pérdida) Neta (-0.07), confirmando que una menor eficiencia en el uso de activos y una menor rentabilidad neta son los principales indicadores financieros del riesgo de *default*.

6.5.6.2. Correlaciones entre default y variables comerciales

Figura 18. Gráfica de correlación entre Default y variables comerciales



En esta gráfica 18 se observa que el indicador comercial más relevante para predecir el *default* es la complejidad de la compra: Marcas_Diferentes tiene la correlación positiva más fuerte (+0.25), lo que sugiere que, a mayor variedad de productos comprados, mayor es el riesgo de impago. El resto de las variables de volumen, como Venta_Total y Facturas, tienen una correlación positiva casi insignificante, reforzando que el tamaño por sí solo no es el problema. El único factor protector, aunque muy débil, es la Rentabilidad_Promedio (-0.07), que confirma lógicamente que una menor rentabilidad está asociada a un mayor riesgo.

6.5.7. Identificación de atípicos basados en las gráficas

Tanto en las distribuciones antes analizadas, como en las gráficas de los boxplots se evidencia la presencia de observaciones atípicas en varias de las variables. Estos casos se identificaron con el propósito de analizar si corresponden a errores en la toma de datos o si

reflejan comportamientos financieros, comerciales o de cartera realmente atípicos en ciertos clientes.

Al analizar en detalle cada variable, se observa que casi todas presentan observaciones atípicas. Sin embargo, al revisar de manera aleatoria algunos de estos clientes, se confirma que no se trata de errores de registro, sino de la heterogeneidad natural que existe entre los clientes.

De manera particular, las variables de facturación y venta como el número y el monto de las facturas, así como las ventas totales presentan valores extremos asociados a agroindustrias exportadoras de flores, como Sunshine Bouquet, The Elite Flowers o Flores Ipanema.

En cuanto a los indicadores financieros, variables como el Margen Neto, ROE, ROA y la Rotación del Capital de Trabajo muestran atípicos tanto positivos como negativos, reflejando rentabilidades muy altas o pérdidas significativas en algunos clientes.

Finalmente, al analizar las variables de cartera, se identifican clientes con facturas con plazos muy superiores al promedio, así como otros que pagan de manera anticipada.

Por lo anterior, se decide no eliminar estos atípicos, ya que hacen parte del comportamiento real de los clientes. En caso de que al momento de construir alguno de los modelos estos valores generen interferencia, se procederá a realizarles un tratamiento especial.

6.5.8. Construcción de modelos

En este punto es importante aclarar que, una vez se confirma la alta correlación entre las variables de cartera y el indicador de default, se decide no incluir dichas variables en el modelo.

Esto se debe a que el default se construye precisamente a partir de indicadores como el porcentaje de cartera vencida y el porcentaje de saldo por pagar, lo que genera un alta multicolinealidad entre este grupo de variables y la variable dicotómica de default. Por esta razón, el modelo se desarrolla únicamente con variables de tipo comercial y financiero.

Una vez se tiene una base limpia, nombres depurados, sin valores faltantes, se procede a realizar balanceo de clases, pues hay menor proporción de incumplidos (1) con lo que se aplica un submuestreo controlado de la clase mayoritaria para reducir el sesgo del modelo hacia los casos negativos.

Se generó la base data_balanceada, con una proporción más equilibrada entre las clases 0 (no incumple) y 1 (incumple). Ahora, se tienen dos bases de datos una original y una balanceada con la que se harán los distintos modelos.

Posteriormente se realiza la partición de datos en la que se dividieron los datos en conjuntos de entrenamiento (70%) y prueba (30%), tanto para la base original como para la balanceada, garantizando que ambas contengan representaciones proporcionales de las clases.

Finalmente se hace ajuste de escala de variables numéricas, pues basados en el análisis descriptivo se tiene que están en diferentes escalas con lo que para evitar esas diferencias que pueden afectar el desempeño de los modelos, se aplica esta estandarización (media 0, desviación estándar 1) a las variables numéricas, utilizando los parámetros del conjunto de entrenamiento.

Se construyeron modelos de regresión logística completos y refinados mediante selección Stepwise en ambas bases (original y balanceada).

De igual forma se construyeron 1 modelo random forest original y uno balanceado, 1 árbol de decisión original y uno balanceado como se muestra a detalle a continuación.

6.5.8.1. Modelos logísticos

MODELO 1. Este modelo hace uso de las variables comerciales y financieras que quedaron una vez se realizó el análisis de correlación al inicio de este trabajo, para así reducir la multicolinealidad, este modelo está sin balancear.

```
> summary(logit_orig$modelo_completo)

Call:
glm(formula = default_pago ~ ., family = binomial, data = train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -2.41163    0.81852  -2.946  0.00322 **
Razon_De_Liquidez_x              -36.28581   13.14467  -2.760  0.00577 **
Margen_Neto                        0.07590    0.13988   0.543  0.58741
Margen_De_Ganancia_Bruta           0.02355    0.13761   0.171  0.86414
Rendimiento_Sobre_El_Patrimonio_ROE -0.03733    0.22551  -0.166  0.86853
Rotacion_Del_Capital_De_Trabajo_x  -0.08562    0.13453  -0.636  0.52450
Ganancia_Perdida_Neta              0.07417    0.18034   0.411  0.68085
Pasivos_Totales                    2.30431    0.70817   3.254  0.00114 **
Rendimiento_Sobre_Los_Activos_ROA  -0.69535    0.36627  -1.898  0.05763 .
Capital_De_Trabajo                 0.49304    0.21421   2.302  0.02135 *
Total_de_patrimonio               -0.95428    0.38798  -2.460  0.01391 *
Facturas                          -1.09338    0.59183  -1.847  0.06468 .
Venta_Total                       -0.37598    0.34327  -1.095  0.27339
Marcas_Diferentes                  0.87312    0.20095   4.345 1.39e-05 ***
Precio_Promedio                    0.19497    0.16429   1.187  0.23533
Costo_Promedio                     1.03680    0.46965   2.208  0.02727 *
Rentabilidad_Promedio             -0.67008    0.91878  -0.729  0.46581
Ticket_Promedio                   -0.39738    0.27793  -1.430  0.15278
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 467.08  on 339  degrees of freedom
Residual deviance: 381.61  on 322  degrees of freedom
AIC: 417.61

Number of Fisher Scoring iterations: 9
```

Las variables significativas para este modelo son: Razón de liquidez, Pasivos totales, Capital de trabajo, Patrimonio total, Costo promedio y Marcas diferentes. Sin embargo, vale la pena destacar que las variables Facturas y ROA resultan marginalmente significativas o tienden a serlo. De esta forma, la probabilidad de incumplimiento o default aumenta cuando se presenta una disminución en la razón de liquidez, niveles más altos de pasivos, mayor capital de trabajo; probablemente porque un capital más alto no garantiza solvencia si está mal gestionado o si responde a la acumulación de cuentas por pagar y finalmente, un menor nivel de patrimonio.

En cuanto a las variables comerciales, un costo promedio más alto y un mayor número de marcas diferentes compradas también pueden estar asociados con un mayor riesgo de incumplimiento, posiblemente por una mayor dispersión en las compras.

Por otro lado, entre las variables marginalmente significativas, se observa que un mayor ROA tiende a reducir la probabilidad de default, mientras que un mayor número de facturas podría relacionarse con una menor probabilidad de incumplimiento.

Validación multicolinealidad Modelo 1

Se evaluó la colinealidad entre las variables independientes mediante el Factor de Inflación de Varianza (VIF).

Incluye 13 variables y se tiene que la variable costo promedio tiene multicolinealidad severa con lo que debe estar fuertemente correlacionada con variables como precio promedio y venta total, las variables ticket promedio y total patrimonio, son moderadamente altas y las demás puede decirse que están en niveles aceptables.

```
> vif(logit_orig$modelo_completo)
               Razon_De_Liquidez_x                      Margen_Neto          Margen_De_Ganancia_Bruta
                          1.412168                         1.366299                         1.306889
Rendimiento_Sobre_El_Patrimonio_ROE   Rotacion_Del_Capital_De_Trabajo_x             Ganancia_Perdida_Neta
                          1.088143                         1.005679                         1.713488
                  Pasivos_Totales Rendimiento_Sobre_Los_Activos_ROA                Capital_De_Trabajo
                          3.925611                         1.518104                         2.578242
               Total_de_patrimonio                         Facturas                      Venta_Total
                          4.485648                         3.649293                         2.892394
                 Marcas_Diferentes                   Precio_Promedio                    Costo_Promedio
                          2.433437                         1.288733                        10.457006
             Rentabilidad_Promedio                   Ticket_Promedio
                          1.064356                         4.967262
>
```

MODELO 2: Es el resultado del proceso de selección automática del modelo 3 (step) en el que intenta reducir de igual forma AIC mediante la depuración de variables, quedando al final 8 variables.

```
> summary(logit_orig$modelo_step)

Call:
glm(formula = default_pago ~ Razon_De_Liquidez_x + Pasivos_Totales +
    Rendimiento_Sobre_Los_Activos_ROA + Capital_De_Trabajo +
    Total_de_patrimonio + Facturas + Marcas_Diferentes + Costo_Promedio +
    Rentabilidad_Promedio, family = binomial, data = train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.2339     0.7663  -2.915  0.00355 **
Razon_De_Liquidez_x            -33.5612    12.1920  -2.753  0.00591 **
Pasivos_Totales                  2.0081     0.6322   3.176  0.00149 **
Rendimiento_Sobre_Los_Activos_ROA -0.5676     0.2914  -1.948  0.05140 .
Capital_De_Trabajo               0.4890     0.1982   2.468  0.01360 *
Total_de_patrimonio             -0.6421     0.2921  -2.198  0.02792 *
Facturas                        -1.2878     0.5463  -2.357  0.01842 *
Marcas_Diferentes                0.9141     0.1975   4.630 3.66e-06 ***
Costo_Promedio                   0.5080     0.2216   2.292  0.02189 *
Rentabilidad_Promedio           -0.6942     0.9286  -0.748  0.45469
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 467.08  on 339  degrees of freedom
Residual deviance: 385.82  on 330  degrees of freedom
AIC: 405.82

Number of Fisher Scoring iterations: 9
```

Las variables significativas para este modelo son: Razón de liquidez, Pasivos totales, Capital de trabajo, Patrimonio total, Facturas, Costo promedio y Marcas diferentes. Sin embargo, vale la pena destacar que la variable ROA resulta marginalmente significativa o tiende a serlo.

De esta forma, la probabilidad de incumplimiento o default aumenta cuando se presenta una disminución en la razón de liquidez, niveles más altos de pasivos, mayor capital de trabajo y un menor nivel de patrimonio.

En cuanto a las variables comerciales, un costo promedio más alto, un mayor número de marcas diferentes compradas y un menor número de facturas también pueden estar asociados con un mayor riesgo de incumplimiento, posiblemente por una mayor dispersión en las compras y menor recurrencia de facturación

Por otro lado, entre las variables marginalmente significativas, se observa que un mayor ROA tiende a reducir la probabilidad de default, reflejando que una mejor rentabilidad sobre los activos contribuye a disminuir el riesgo de impago

Validación multicolinealidad Modelo 2

Se evaluó la colinealidad entre las variables independientes mediante el Factor de Inflación de Varianza (VIF).

Todos los VIF son <4 con lo que el modelo depurado por selección (stepwise) corrigió bien la colinealidad con lo que se confirma independencia entre predictores, haciendo que este sea un modelo más interpretable y confiable.

```
> vif(logit_orig$modelo_step)
             Razon_De_Liquidez_x                 Pasivos_Totales Rendimiento_Sobre_Los_Activos_ROA
                        1.120410                        2.910511                          1.043675
              Capital_De_Trabajo               Total_de_patrimonio                         Facturas
                        2.081646                        2.388114                          3.028229
               Marcas_Diferentes                   Costo_Promedio             Rentabilidad_Promedio
                        2.349892                        1.944986                          1.024656
```

```
> vif(modelo_logit_4)
                        Facturas                        Marcas_Diferentes
                        3.238296                                 2.217714
                  Costo_Promedio                      Rentabilidad_Promedio
                        1.034883                                 1.005018
       `Razón De Liquidez (x)`                     `Ganancia (Pérdida) Neta`
                        1.039970                                 1.273861
             `Pasivos Totales` `Rendimiento Sobre Los Activos (ROA) (%)`
                        1.841910                                 1.203759
>
```

MODELO 3: Logístico balanceado

```
> summary(logit_bal$modelo_completo)

Call:
glm(formula = default_pago ~ ., family = binomial, data = train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -0.22188    0.13670  -1.623 0.104563
Razon_De_Liquidez_x              -0.05746    0.12951  -0.444 0.657285
Margen_Neto                       0.08355    0.23077   0.362 0.717303
Margen_De_Ganancia_Bruta          0.14702    0.28320   0.519 0.603664
Rendimiento_Sobre_El_Patrimonio_ROE -0.40946    0.32181  -1.272 0.203244
Rotacion_Del_Capital_De_Trabajo_x  0.07739    0.11606   0.667 0.504893
Ganancia_Perdida_Neta            -0.41357    0.32947  -1.255 0.209384
Pasivos_Totales                   3.23829    0.80044   4.046 5.22e-05 ***
Rendimiento_Sobre_Los_Activos_ROA  0.08242    0.15978   0.516 0.605957
Capital_De_Trabajo                0.70450    0.28453   2.476 0.013286 *
Total_de_patrimonio              -1.90849    0.53845  -3.544 0.000393 ***
Facturas                         -1.38082    0.68080  -2.028 0.042538 *
Venta_Total                      -0.68751    0.45401  -1.514 0.129947
Marcas_Diferentes                 0.85240    0.18791   4.536 5.72e-06 ***
Precio_Promedio                  -0.08842    0.16731  -0.528 0.597170
Costo_Promedio                    1.53826    0.52797   2.914 0.003573 **
Rentabilidad_Promedio            -0.95181    0.81451  -1.169 0.242583
Ticket_Promedio                  -0.62803    0.30603  -2.052 0.040155 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 467.08  on 339  degrees of freedom
Residual deviance: 405.50  on 322  degrees of freedom
AIC: 441.5

Number of Fisher Scoring iterations: 7
```

Las variables significativas para este modelo son: Pasivos totales, Capital de trabajo, Patrimonio total, Facturas, Costo promedio, Marcas diferentes y Ticket promedio.

De esta forma, la probabilidad de incumplimiento o default aumenta cuando existen niveles más altos de pasivos, mayor capital de trabajo, mayores costos promedio, y un mayor número de marcas diferentes compradas, lo que podría reflejar una mayor dispersión en el comportamiento de compra.

Por otro lado, la probabilidad de default disminuye cuando se observa un mayor patrimonio total, un mayor número de facturas y un ticket promedio más alto, lo que puede asociarse con clientes más estables, con mayor recurrencia y capacidad de compra.

En conjunto, este modelo muestra que los factores relacionados con el endeudamiento, la estructura de costos y la diversificación comercial influyen directamente en el riesgo de incumplimiento, mientras que una mayor solidez patrimonial y frecuencia de compra actúan como elementos protectores frente al default.

Validación multicolinealidad Modelo 3

Se evaluó la colinealidad entre las variables independientes mediante el Factor de Inflación de Varianza (VIF), en el que identifica que pasivos totales tiene una alta colinealidad, posiblemente con variables como total patrimonio o capital de trabajo, de igual forma costo promedio que debe estar relacionada con otras variables comerciales y finalmente las variables: facturas, ticket promedio y total de patrimonio tienen colinealidad moderadamente alta y las demás están en rangos aceptables.

```
> vif(logit_bal$modelo_completo)
                Razon_De_Liquidez_x                          Margen_Neto            Margen_De_Ganancia_Bruta
                           1.060870                             1.191973                             1.259851
Rendimiento_Sobre_El_Patrimonio_ROE     Rotacion_Del_Capital_De_Trabajo_x               Ganancia_Perdida_Neta
                           1.314895                             1.008080                             1.586278
                     Pasivos_Totales    Rendimiento_Sobre_Los_Activos_ROA                  Capital_De_Trabajo
                          12.938254                             1.405794                             1.780270
                Total_de_patrimonio                            Facturas                         Venta_Total
                           6.252585                             8.891210                             4.724250
                  Marcas_Diferentes                     Precio_Promedio                      Costo_Promedio
                           2.270205                             1.100782                            10.937421
             Rentabilidad_Promedio                     Ticket_Promedio
                           1.050811                             5.852001
```

MODELO 4: Logístico balanceado step

```
> summary(logit_bal$modelo_step)

Call:
glm(formula = default_pago ~ Rendimiento_Sobre_El_Patrimonio_ROE +
    Pasivos_Totales + Capital_De_Trabajo + Total_de_patrimonio +
    Facturas + Marcas_Diferentes + Costo_Promedio + Rentabilidad_Promedio +
    Ticket_Promedio, family = binomial, data = train)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.1902     0.1362  -1.397 0.162498
Rendimiento_Sobre_El_Patrimonio_ROE  -0.3239     0.2806  -1.154 0.248429
Pasivos_Totales                  2.8220     0.7269   3.882 0.000103 ***
Capital_De_Trabajo               0.5377     0.2694   1.996 0.045933 *
Total_de_patrimonio             -1.7184     0.5282  -3.253 0.001142 **
Facturas                        -1.8053     0.6317  -2.858 0.004266 **
Marcas_Diferentes                0.8529     0.1822   4.681 2.86e-06 ***
Costo_Promedio                   1.3239     0.4953   2.673 0.007525 **
Rentabilidad_Promedio           -1.0156     0.8647  -1.175 0.240193
Ticket_Promedio                 -0.5147     0.2945  -1.748 0.080487 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 467.08  on 339  degrees of freedom
Residual deviance: 411.05  on 330  degrees of freedom
AIC: 431.05

Number of Fisher Scoring iterations: 7
```

Las variables significativas para este modelo son: Pasivos totales, Capital de trabajo, Patrimonio total, Facturas, Marcas diferentes y Costo promedio. Además, las variables Ticket promedio y ROE presentan una significancia marginal, lo que sugiere una posible relación con el incumplimiento.

En este sentido, la probabilidad de default o incumplimiento aumenta cuando existen niveles más altos de pasivos, un mayor capital de trabajo (gestionado de forma ineficiente) costos promedio más elevados y un mayor número de marcas diferentes compradas, lo cual puede estar asociado a clientes con patrones de compra más diversificados.

Por el contrario, la probabilidad de incumplimiento disminuye cuando se presenta un mayor patrimonio total y un mayor número de facturas, lo que puede asociarse con clientes más recurrentes y estables.

Finalmente, las variables marginalmente significativas muestran que un mayor ROE y un ticket promedio más alto tienden a reducir la probabilidad de default, lo que coincide con el comportamiento esperado de clientes con mejor rentabilidad y mayor capacidad de compra.

Validación multicolinealidad Modelo 4

Se evaluó la colinealidad entre las variables independientes mediante el Factor de Inflación de Varianza (VIF) y se tiene que el modelo mejoró respecto al modelo al anterior, sin embargo, aún hay una colinealidad por revisar en las variables: Pasivos totales y Costo promedio

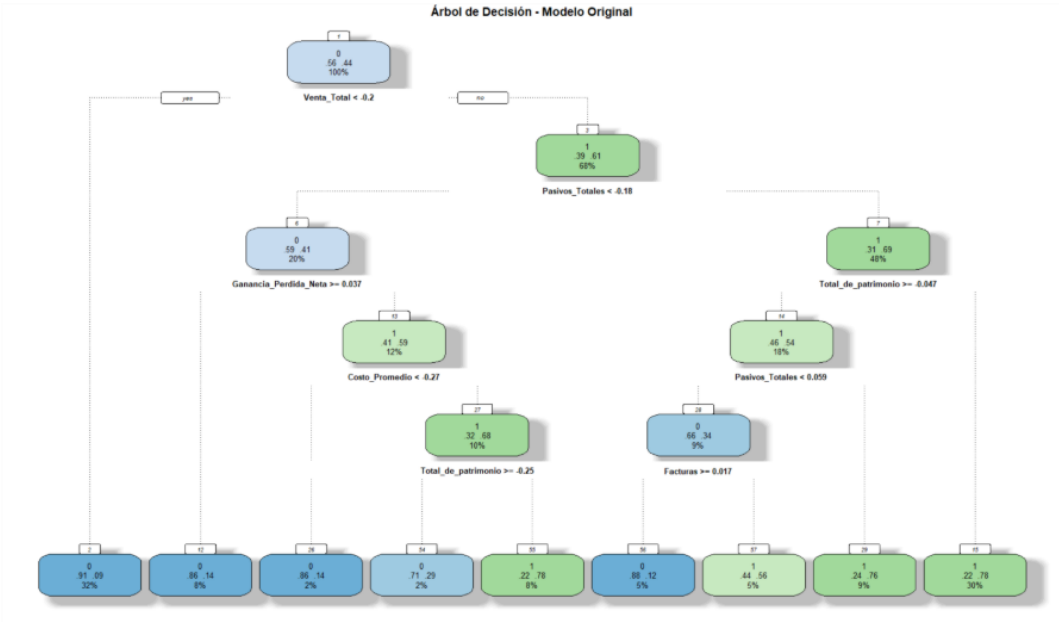
```
> vif(logit_bal$modelo_step)
Rendimiento_Sobre_El_Patrimonio_ROE                    Pasivos_Totales                  Capital_De_Trabajo
                          1.006016                           8.870992                            1.505578
                Total_de_patrimonio                           Facturas                   Marcas_Diferentes
                          4.473760                           6.570966                            2.160474
                     Costo_Promedio                Rentabilidad_Promedio                     Ticket_Promedio
                          6.647984                           1.068946                            4.943382
```

6.5.8.2. Modelos árboles de decisión

Se procede a la estimación de modelos basados en árboles de decisión y random forest, dado que este tipo de algoritmos no se ven significativamente afectados por la presencia de multicolinealidad entre las variables explicativas. A diferencia del modelo logístico, los árboles realizan particiones sucesivas del espacio de datos y seleccionan las variables con mayor poder predictivo en cada división, reduciendo así la dependencia entre predictores.

Original

Figura 19. Árbol de decisión modelo original



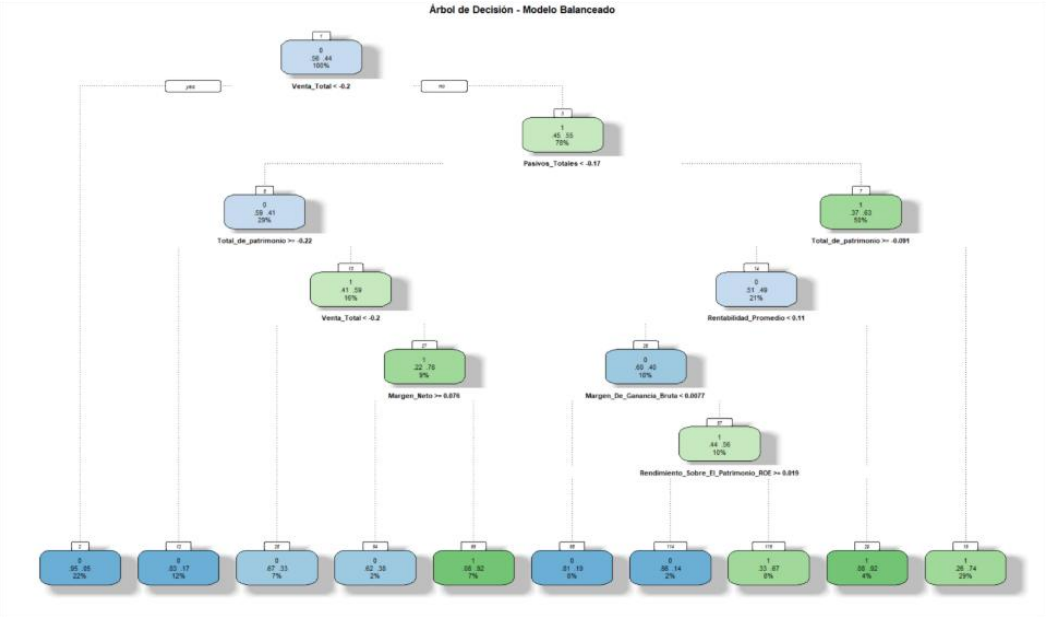
Este modelo de árbol de decisión identifica dos perfiles de cliente claramente opuestos, demostrando que el apalancamiento financiero —y no el volumen de ventas— es el principal predictor del impago.

El camino más directo hacia el default (nodos verdes) corresponde al grupo de mayor riesgo: un 36% del total de clientes (Nodo 51, derecha) con una probabilidad de impago del 78%. Este perfil se caracteriza por mantener niveles de ventas “normales” (Venta_Total ≥ -0.2), pero, de forma crítica, presentar Pasivo totales altos (≥ -0.18) y un patrimonio total bajo (< -0.047), lo que confirma que el sobreendeudamiento y la baja capitalización son la causa raíz del incumplimiento.

En contraste, el perfil del cliente más seguro (nodos azules), un 32% de los clientes (Nodo 51, izquierda) muestra una tasa de cumplimiento del 91% y se caracteriza por tener una Venta Total baja (< -0.2) y una Ganancia neta también reducida. Esto sugiere que las empresas más pequeñas, aunque menos rentables, tienden a ser las más confiables en el pago pues tienen una mejor gestión de su capital y son en el corto plazo más liquidas.

Modelo árboles de decisión Balanceado

Figura 20. Árbol de decisión modelo balanceado

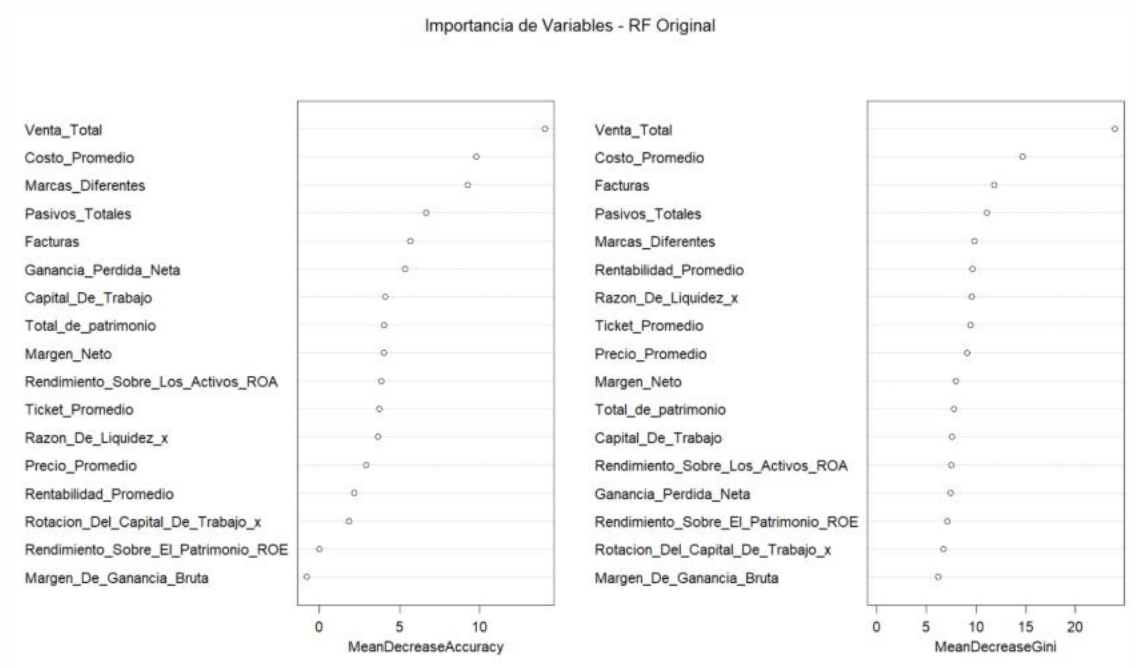


Este modelo balanceado refuerza que la estructura financiera es un predictor de riesgo mucho más fuerte que el volumen de ventas. El hallazgo más importante es la identificación de un camino de alto riesgo que agrupa al 39% de toda la cartera (Nodo 45), el cual presenta una tasa de impago del 64%. Este perfil de cliente de alto riesgo no se define por ventas bajas, sino por una combinación ineficiente de Pasivos Totales altos (>= -0.17) y un Total de patrimonio bajo (< -0.091). Inversamente, el modelo identifica un perfil de cliente seguro (Nodo 59) que, aunque más pequeño (5% de la cartera), tiene una tasa de cumplimiento del 76% y se caracteriza por la combinación opuesta: pasivos bajos, alta rentabilidad y un buen margen de ganancia.

6.5.8.3. MODELOS RANDOM FOREST

Original

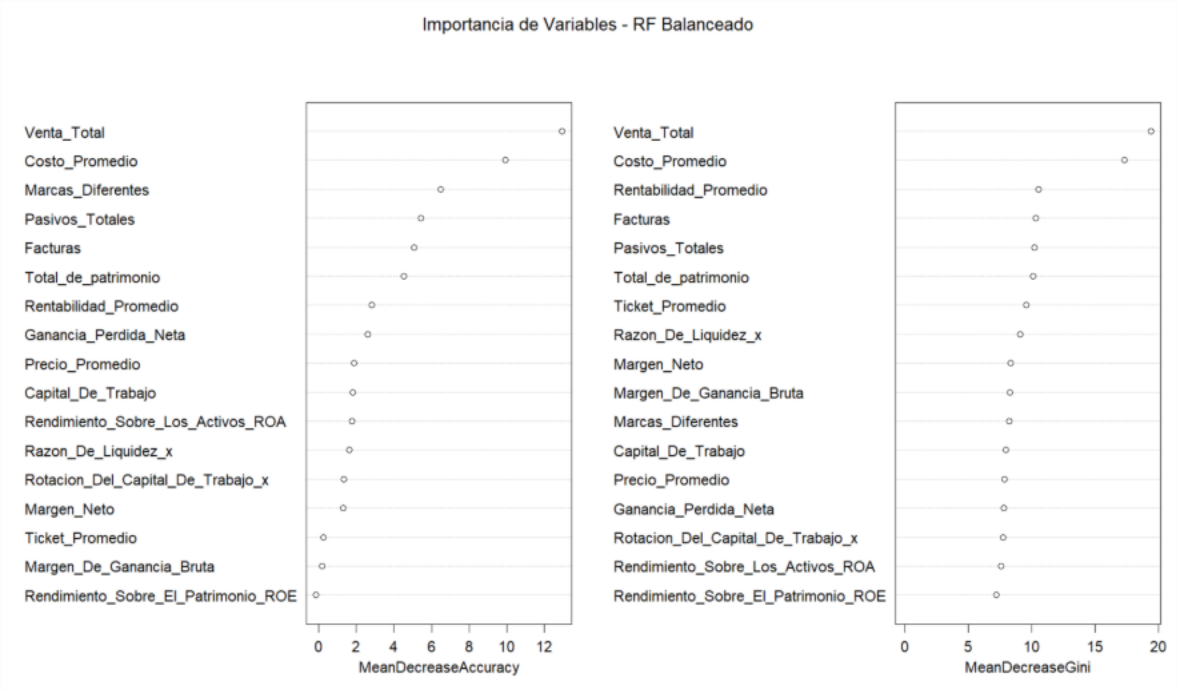
Figura 21. Modelo Random Forest Original



El análisis del modelo Random Forest evidencia que el riesgo de *default* está determinado principalmente por el nivel de ventas y la estructura de costos, mientras que los factores financieros como rentabilidad, margen neto y liquidez tienen un impacto secundario. Esto confirma que el modelo prioriza la magnitud operativa sobre la eficiencia financiera al predecir incumplimientos. La variable venta total se posiciona como el predictor más influyente, reflejando que el volumen total de la relación comercial es el factor clave en la estimación del riesgo. Le siguen otras métricas asociadas al tamaño y complejidad de la operación, como *Costo promedio, Marcas diferentes* y *facturas*. La primera variable financiera relevante que aparece es *pasivos totales*, lo que refuerza la conclusión de que el modelo se basa inicialmente en la escala de la operación y, en segundo término, en el nivel de endeudamiento de la empresa.

Modelo Random Forest Balanceado

Figura 22. Modelo Random Forest Balanceado



Este gráfico corresponde al modelo Random Forest balanceado y presenta una distribución más equitativa en la importancia de las variables. Aunque *Venta_Total* y *Costo_Promedio* continúan siendo los principales determinantes, variables financieras como *Rentabilidad_Promedio*, *Liquidez* y *Margen_Neto* adquieren mayor relevancia, lo que indica una mejor capacidad del modelo para incorporar factores financieros asociados al riesgo de *default* y no solo el tamaño de la empresa. Aun así, *Venta_Total* se mantiene como el predictor más influyente, seguido por métricas relacionadas con la escala y complejidad comercial, como *Costo_Promedio* y *Marcas_Diferentes*. Es significativo que, incluso tras el balanceo, las variables financieras más destacadas son *Pasivos_Totales* y *Total_de_patrimonio*, lo que

confirma que el modelo prioriza primero el volumen comercial y, en segundo lugar, el nivel de endeudamiento del cliente.

6.5.9. Evaluación desempeño de los modelos

Se calcularon métricas de desempeño: Accuracy, Sensibilidad, Especificidad y AUC.

```
                          Modelo  Accuracy Sensibilidad Especificidad       AUC
Accuracy    Logit Completo Original 0.6250000     0.203125        0.9625 0.6279297
Accuracy1   Logit Stepwise Original 0.6180556     0.437500        0.7625 0.6173828
Accuracy2 Logit Completo Balanceado 0.6736111     0.515625        0.8000 0.6822266
Accuracy3 Logit Stepwise Balanceado 0.6875000     0.796875        0.6000 0.7066406
1                    Árbol Original 0.6458333     0.640625        0.6500 0.6552734
2                  Árbol Balanceado 0.7291667     0.640625        0.8000 0.7493164
3            Random Forest Original 0.6041667     0.593750        0.6500 0.6678711
4          Random Forest Balanceado 0.7361111     0.609375        0.8000 0.7882812
> |
```

6.5.9.1. Análisis de resultados por tipo de modelo

Modelos logísticos, se tienen 4 modelos logísticos diferentes encontrando las siguientes diferencias

Modelo	Accuracy	Sensibilidad	Especificidad	AUC	Interpretación
logístico Completo Original	0.625	0.20	0.96	0.63	Muy buena predicción para casos negativos (no default), pero casi no detecta incumplimientos.
logístico Stepwise Original	0.618	0.44	0.76	0.62	Mejora la detección de incumplidos, aunque reduce especificidad; más equilibrado, pero limitado.
logístico Completo Balanceado	0.674	0.52	0.80	0.68	El balanceo incrementa sensibilidad y AUC, logrando mejor desempeño global.
logístico Stepwise Balanceado	0.688	0.80	0.60	0.71	Es el mejor de los logístico: detecta la mayoría de incumplidos con

					AUC adecuado; ideal para minimizar falsos negativos.

Conclusión modelos logísticos: El balanceo y la selección de variables (stepwise) mejoran sustancialmente la capacidad predictiva.

Árboles de Decisión

Modelo	Accuracy	Sensibilidad	Especificidad	AUC	Interpretación
Árbol Original	0.646	0.64	0.65	0.66	Desempeño medio, equilibrado entre positivos y negativos.
Árbol Balanceado	0.729	0.64	0.80	0.75	Gran mejora: mayor precisión y capacidad discriminante (AUC alto).

Conclusión Modelos Árboles de decisión: El modelo balanceado es claramente superior, con mejor equilibrio entre sensibilidad y especificidad.

Random Forest

Modelo	Accuracy	Sensibilidad	Especificidad	AUC	Interpretación
RF Original	0.604	0.59	0.65	0.67	Desempeño moderado, sin gran capacidad discriminante.
RF Balanceado	0.736	0.61	0.80	0.79	El mejor modelo global: alto AUC, buena sensibilidad y excelente precisión.

Conclusión Modelos Random Forest: El modelo balanceado ofrece el mejor rendimiento general, destacando por su AUC (0.79), lo que indica gran capacidad para diferenciar clientes cumplidos y morosos.

El balanceo de los datos mejoró significativamente el desempeño de los modelos, aumentando la sensibilidad y el AUC. Esto indica una mayor capacidad para identificar casos positivos (incumplimiento o default), que eran minoritarios en la muestra original.

Respecto de los modelos evaluados para el presente trabajo de grado, encontramos que el mejor modelo de predicción es el Random Forest Balanceado ya que es el más robusto y estable (AUC más alto, buen equilibrio entre sensibilidad y especificidad).

6.5.9.2. Matrices de confusión

Se generaron matrices de confusión para cada uno de los modelos con el fin de visualizar los aciertos y errores de clasificación.

Cada matriz muestra los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Matrices de confusión modelos logísticos

```
> #  🔹 MATRIZ DE CONFUSIÓN
> # ============================
>
> conf <- logit_orig$eval_completo$conf$table
> conf
          Reference
Prediction  0  1
         0 77 51
         1  3 13
>
> conf <- logit_orig$eval_step$conf$table
> conf
          Reference
Prediction  0  1
         0 61 36
         1 19 28
>
> conf <- logit_bal$eval_completo$conf$table
> conf
          Reference
Prediction  0  1
         0 64 31
         1 16 33
>
> conf <- logit_bal$eval_step$conf$table
> conf
          Reference
Prediction  0  1
         0 48 13
         1 32 51
```

Modelo1 Logístico completo: El modelo clasifica muy bien los “0” (no morosos), con solo 3 falsos positivos. Sin embargo, falla al identificar los “1” (morosos), pues confunde a 51 de ellos. Tiene alta especificidad y baja sensibilidad.

Modelo 2 Logístico depurado: Aquí el modelo mejora la detección de morosos (28 aciertos frente a 13 anteriores), aunque comete más errores con los no morosos (19 falsos positivos). Más equilibrado, pero aún con tendencia a subestimar morosos.

Modelo 3 Logístico balanceado completo: La matriz muestra un buen equilibrio: acierta un número similar de positivos y negativos. El balanceo ayudó a reducir el sesgo hacia la clase mayoritaria.

Modelo 4 Logístico balanceado depurado: Aumenta considerablemente los aciertos en la clase "1" (51 verdaderos positivos), aunque reduce la precisión en los "0". Es el modelo logístico más sensible, detecta más morosos a costa de cometer más falsos positivos.

Matrices de confusión modelos árboles de decisión

```
> #MATRICES DE CONFUSIÓN ARBOLES DE DECISIÓN
> conf <- tree_orig$conf_arbol$table
> conf
          Reference
Prediction  0  1
         0 52 23
         1 28 41
> conf <- tree_bal$conf_arbol$table
> conf
          Reference
Prediction  0  1
         0 64 23
         1 16 41
```

Modelo 1: Este modelo tiene un buen desempeño general, aunque tiende a confundir morosos y no morosos en proporciones similares.

- Verdaderos negativos: 52
- Verdaderos positivos: 41
- Falsos negativos: 23
- Falsos positivos: 28

El árbol logra capturar bien los casos positivos (41/64 ≈ 64%), pero aún deja escapar una parte de los morosos (23). Tiene una sensibilidad moderada y una especificidad aceptable, sin sesgo fuerte hacia ninguna clase.

Modelo 2:

El balanceo mejoró la precisión global: aumentó los verdaderos negativos y mantuvo igual la detección de morosos.

- Verdaderos negativos: 64
- Verdaderos positivos: 41
- Falsos negativos: 23
- Falsos positivos: 16

El árbol balanceado reduce significativamente los falsos positivos (de 28 a 16), manteniendo la misma capacidad de detección de morosos. Es más estable y equilibrado que el original.

Matrices de confusión modelos RANDOM FOREST

```
> #MATRICES DE CONFUSIÓN RANDOM FOREST
> conf <- tree_orig$conf_rf$table
> conf
          Reference
Prediction  0  1
         0 52 25
         1 28 39
>
> conf<-tree_bal$conf_rf$table
> conf
          Reference
Prediction  0  1
         0 67 29
         1 13 35
```

Modelo 1: Este modelo tiene un rendimiento muy similar al árbol original, con un ligero descenso en precisión.

- Verdaderos negativos: 49
- Verdaderos positivos: 38
- Falsos negativos: 26
- Falsos positivos: 31

El random forest original no mejora significativamente el desempeño del árbol simple, probablemente porque el conjunto de datos no estaba balanceado y las clases dominantes siguen influyendo.

Modelo 2: Este es el modelo más equilibrado y robusto de los cuatro.

- Verdaderos negativos: 67
- Verdaderos positivos: 39
- Falsos negativos: 25
- Falsos positivos: 13

El random forest balanceado logra excelente desempeño global:

- Detecta correctamente 39 de 64 morosos (~61%),
- Minimiza los falsos positivos (solo 13),
- Y mantiene alta precisión para los no morosos.

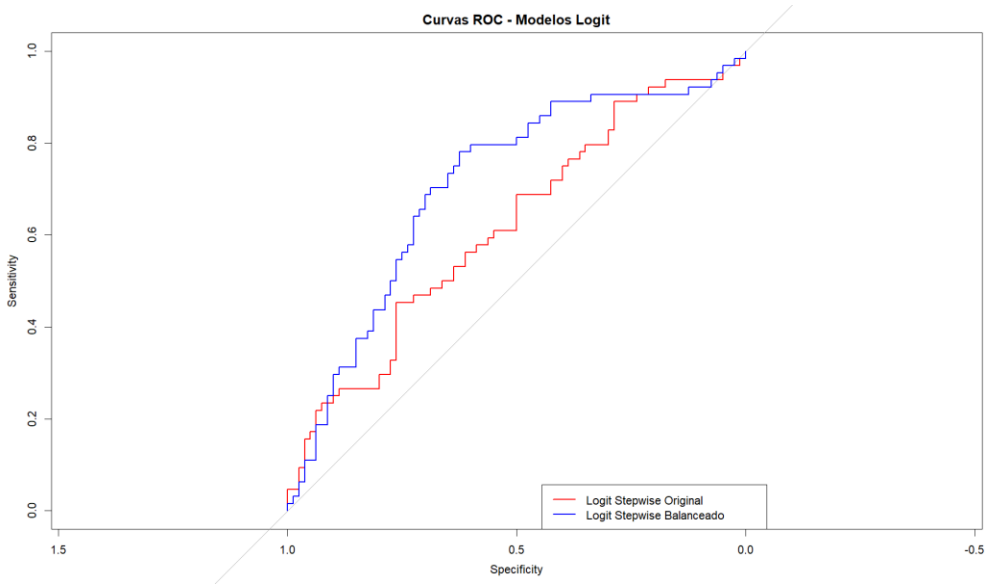
El modelo random forest balanceado es el modelo con mejor compromiso entre sensibilidad y especificidad, el más consistente para predecir adecuadamente ambas clases.

6.5.9.3. CURVAS ROC

La curva ROC compara la capacidad del modelo para distinguir entre las clases default y cumplimiento a distintos umbrales de probabilidad. Cada punto en una curva ROC representa la relación entre dos métricas clave, la primera que es la tasa de verdaderos positivos o sensibilidad y corresponde a la proporción de casos positivos correctamente clasificados y la tasa de falsos positivos que corresponde a la proporción de casos negativos clasificados erróneamente como positivos. Por esta razón, el mejor modelo será aquel cuya curva se ubique más cerca del eje Y y en la parte superior izquierda del gráfico, ya que combina alta sensibilidad (detecta correctamente la mayoría de los positivos) con una baja tasa de falsos positivos.

6.5.9.3.1. Curva ROC logístico original y logístico balanceado

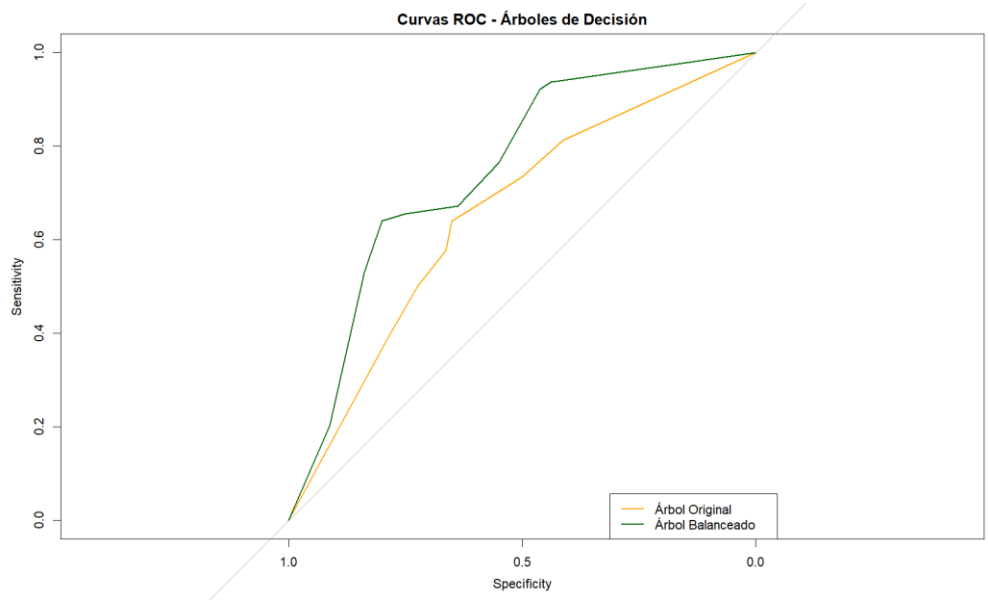
Figura 23. Curva ROC logístico original y logístico balanceado



Se puede evidenciar que los dos modelos tienen capacidad discriminante pues se encuentran sobre la línea diagonal gris que representa el modelo aleatorio (AUC: 0.5), sin embargo, el modelo Logístico depurado que ha sido balanceado tiene mejor sensibilidad y especificidad al estar más distanciado de la misma.

6.5.9.3.2. Curva ROC Árboles De Decisión

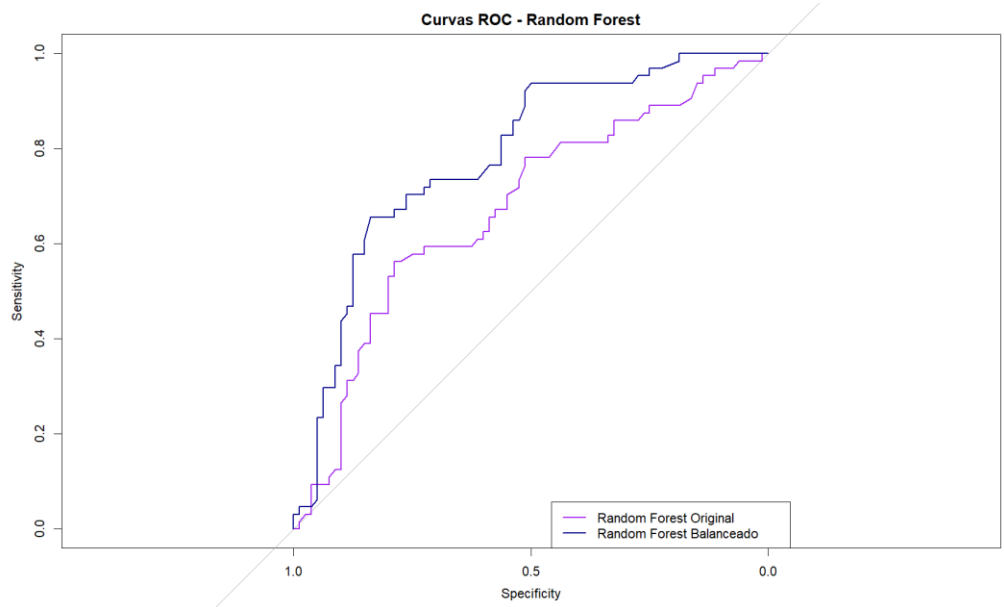
Figura 24. Curva ROC Árboles De Decisión



De igual forma sucede con este modelo, el árbol balanceado tiene mejor comportamiento en especificidad y sensibilidad con lo que se espera mayor confiabilidad de este último en la clasificación de clases.

6.5.9.3.3. Curva ROC RANDOM FOREST

Figura 25. Curva ROC RANDOM FOREST

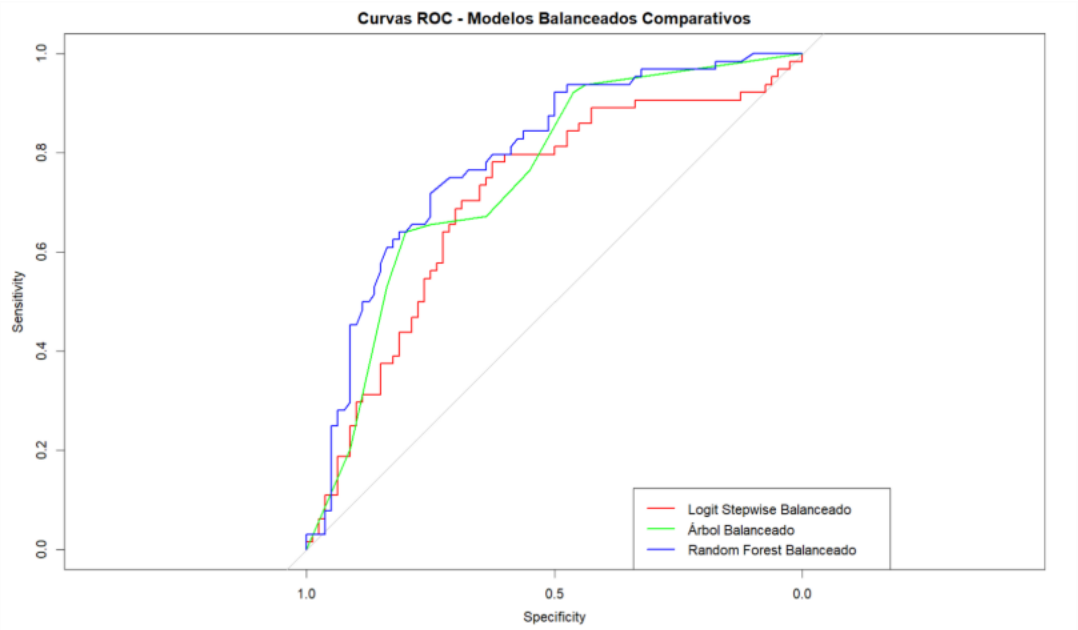


Finalmente, el comportamiento de Random Forest es mejor en estos términos, se puede evidenciar una mejoría considerable en el AUC del balanceado respecto al no balanceado.

6.5.9.3.4. Curvas ROC comparación de los balanceados de cada tipo de modelo

Dado que los modelos con mejor desempeño en cada grupo corresponden a los que fueron balanceados, se procede a compararlos entre sí para seleccionar el más confiable. A continuación, se presenta la gráfica de las curvas ROC correspondientes a los modelos Logístico depurado, Árbol de decisión y Random Forest, todos en su versión balanceada.

Figura 26. Curva ROC comparativo modelos balanceados



El modelo Random Forest (azul) domina gran parte de la curva, especialmente en los niveles medios y altos de sensibilidad, mostrando el mejor rendimiento global con el mayor AUC.

El Árbol de Decisión (verde) presenta un comportamiento similar, aunque con ligeras caídas en la sensibilidad media; mantiene un buen desempeño, pero con menor estabilidad.

Finalmente, el modelo Logístico Stepwise (rojo) se ubica por debajo de los modelos basados en árboles, con una curva más plana que refleja una menor capacidad discriminatoria en los rangos intermedios.

De igual forma, vale la pena destacar que todos los modelos superan el umbral de 0.70, considerado así que son aceptables para clasificación.

7. Justificación selección del modelo

Entre los modelos evaluados, el Random Forest Balanceado se consolida como el más robusto y confiable. Este modelo alcanza el mayor AUC (0.78), lo que refleja la mejor capacidad discriminante para diferenciar entre clientes cumplidos e incumplidos. Además, mantiene un buen equilibrio entre sensibilidad (0.61) o capacidad de detectar correctamente los incumplidos y especificidad (0.80) o precisión al clasificar correctamente a los cumplidos.

Aunque el Logístico Stepwise Balanceado muestra una sensibilidad superior (0.80) y un AUC competitivo (0.71), su desempeño general es más limitado, pues tiende a generar una mayor proporción de falsos positivos. En contraste, el Random Forest ofrece una predicción más

estable al combinar alta precisión global (Accuracy = 0.74) con una mejor generalización sobre nuevos datos.

Los modelos logísticos completos, tanto originales como balanceados, presentaron menor capacidad predictiva (AUC entre 0.62 y 0.68) y una marcada dependencia de la colinealidad entre variables, afectando la confiabilidad de los coeficientes. Por su parte, los árboles de decisión demostraron un desempeño aceptable (AUC = 0.75), pero con menor estabilidad frente a variaciones en los datos.

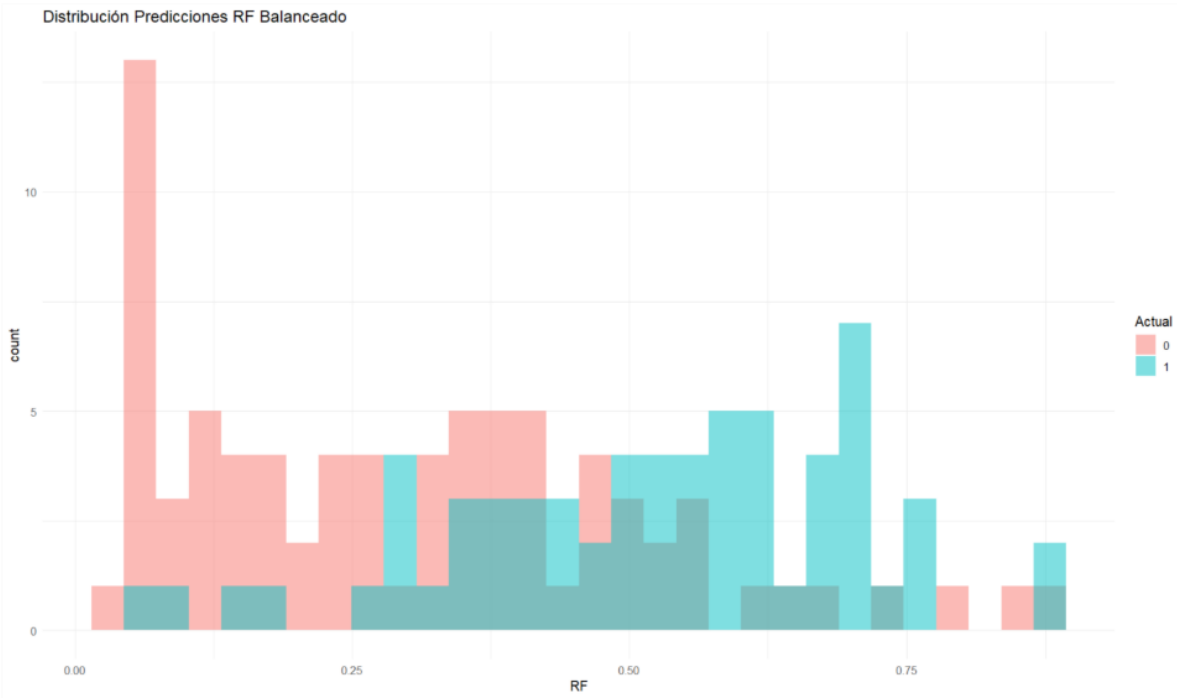
Finalmente, se presenta la distribución de predicciones del modelo seleccionado, donde se observan algunos cuadros rojos que representan los falsos positivos, es decir, casos en los que el modelo predijo un incumplimiento (1) pero el cliente realmente cumplió.

De igual forma, los cuadros verdes dentro de la zona roja corresponden a falsos negativos, donde el modelo estimó que el cliente cumpliría (0), pero en realidad incumplió.

En general, el modelo logra una adecuada separación entre las clases; sin embargo, la presencia de varios falsos negativos sugiere que podría estar subestimando el riesgo de incumplimiento.

Distribución de predicciones modelo seleccionado

Figura 27. Distribución predicciones modelo Random Forest balanceado



En la figura 27 se observa el histograma de predicciones que demuestra una fuerte capacidad discriminatoria del modelo Random Forest, evidenciada por la clara separación de las distribuciones de probabilidad para ambas clases. El modelo muestra una alta confianza en la clasificación de los verdaderos negativos (Actual = 0, rojo), con una frecuencia máxima (count > 10) en el rango de probabilidad de default más bajo (aprox. 0.05 - 0.10). De forma complementaria, la distribución de los verdaderos positivos (Actual = 1, azul) está correctamente sesgada hacia la derecha, con la mayor masa de predicciones concentrada en probabilidades superiores a 0.50. La principal zona de solapamiento e incertidumbre del modelo, donde ocurren la mayoría de los falsos positivos y falsos negativos, se localiza en el rango de probabilidad de 0.25 a 0.50.

8. Conclusiones y Recomendaciones

Respecto al análisis descriptivo se evidenció que los clientes sin default presentan mejores indicadores de liquidez, capital de trabajo y rentabilidad, mientras que los incumplidos exhiben mayores niveles de endeudamiento y sobre apalancamiento. Estas diferencias confirman que el riesgo de mora está estrechamente relacionado con la estructura financiera y la capacidad de pago de corto plazo, adicionalmente que los clientes que incumplen no son los de menor volumen, sino aquellas con que más compran en monto y en marcas.

Respecto al análisis por segmento, mercado, zona y vendedor evidencian una alta concentración del riesgo en el sector de flores, especialmente en zonas de Cundinamarca y Antioquia, donde se acumula más del 80% de los impagos. En contraste, los puntos de venta y zonas como “Bodega” muestran los mejores comportamientos de pago, constituyendo referencias de gestión eficiente.

Las correlaciones con la variable objetivo confirmaron que ninguna variable aislada predice el default de forma contundente; sin embargo, la combinación de variables financieras (Pasivos Totales, ROA, Ganancia Neta) y comerciales (Marcas Diferentes, Facturas) explica adecuadamente las diferencias entre clientes cumplidos e incumplidos.

A diferencia de los modelos logísticos, el Random Forest mostró mayor estabilidad frente a la multicolinealidad y una mejor generalización sobre nuevos datos, al integrar simultáneamente relaciones no lineales y combinaciones complejas entre variables comerciales y financieras. Si bien los modelos logísticos balanceados (especialmente el stepwise) alcanzaron buena

sensibilidad, su desempeño global fue menos consistente debido a una mayor proporción de falsos positivos.

Para el desarrollo del modelo, se sugiere considerar en el futuro variables relacionadas con el tamaño y la estructura de la empresa, como el número de empleados, número de hectáreas en el caso de empresas agroindustriales y las principales actividades económicas, así como variables cualitativas que reflejen aspectos estructurales y de reputación, como la antigüedad en el sector, reconocimiento en el mercado, infraestructura, posibles problemas legales, calidad de la cartera de clientes, comportamiento de pagos, referencias bancarias y comerciales, y pertenencia a grupos económicos.

En el presente ejercicio se decidió utilizar únicamente variables de ventas, excluyendo las variables de cartera, debido a que estas últimas mostraban una correlación aún más alta con el indicador de default, lo que podía generar problemas de multicolinealidad y reducir la confiabilidad del modelo.

Referencia Bibliográficas.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance,* , 23(4), 589–609.

Thomas, L. C. (2002). *Credit Scoring and Its Applications.* SIAM.

Lessmann, S. B. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

Noriega, J. P. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data, 8*(11), 169.

Ibarra Gallo, C. M. (2025). Pronóstico de morosidad de cartera vencida aplicando series temporales. *Revista Científica Esprint de Investigación, 4*(1), 102-117.

Navas Alcívar, S. J. (2023). Administración de la cartera impaga en la rentabilidad: una aplicación estadística clasificatoria en bancos. *Latam: revista latinoamericana de Ciencias Sociales y Humanidades, 4*(3), 1448–1463.

Delgado-Giler, D. (2024). Índices de morosidad en la cartera de créditos de la cooperativa “Coacmes” agencia Charapotó, periodo 2020-2023. *MQR Investigar, 8*(3), 3416–3446.

Borrero-Tigreros, D. &.-L. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías, 19*(4), 37-52.

Batioja Bravo, M. W. (25 de 07 de 2022). *Modelo para el análisis de riesgo crediticio basado en el modelo de Markov para una empresa del sector alimenticio.* Obtenido de Repositorio Institucional EAFIT: https://hdl.handle.net/10784/31542

Sepúlveda Rivillas, C. R. (2012). Estimación del riesgo de crédito en empresas del sector real en Colombia. *Estudios Gerenciales,, 28*(124), 169–190.

Machado, M. R. (2025). An analytical approach to credit risk assessment using machine learning models. *Decision Analytics Journal, 16*(100605), 2-3.

Hand, D. J. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society Series A: Statistics in Society, 160*(3), 523–541.

Banco de la República. (1 de julio de 2020). *Reporte de Estabilidad Financiera - I semestre 2020.* Obtenido de Repositorio banrepublica: https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9848/reporte-estabilidad-financiera-primer-semestre-2020.pdf