PONTIFICIA UNIVERSIDAD

CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

Gestión de Riesgo Avanzado

"Modelo de Rating"

Aquino Alcántara, Cristian Jhonel (20181407) (50%)

Quispe Robladillo, Almendra Valeria (20193348) (50%)

Lima, 2024

1. Ficha Técnica

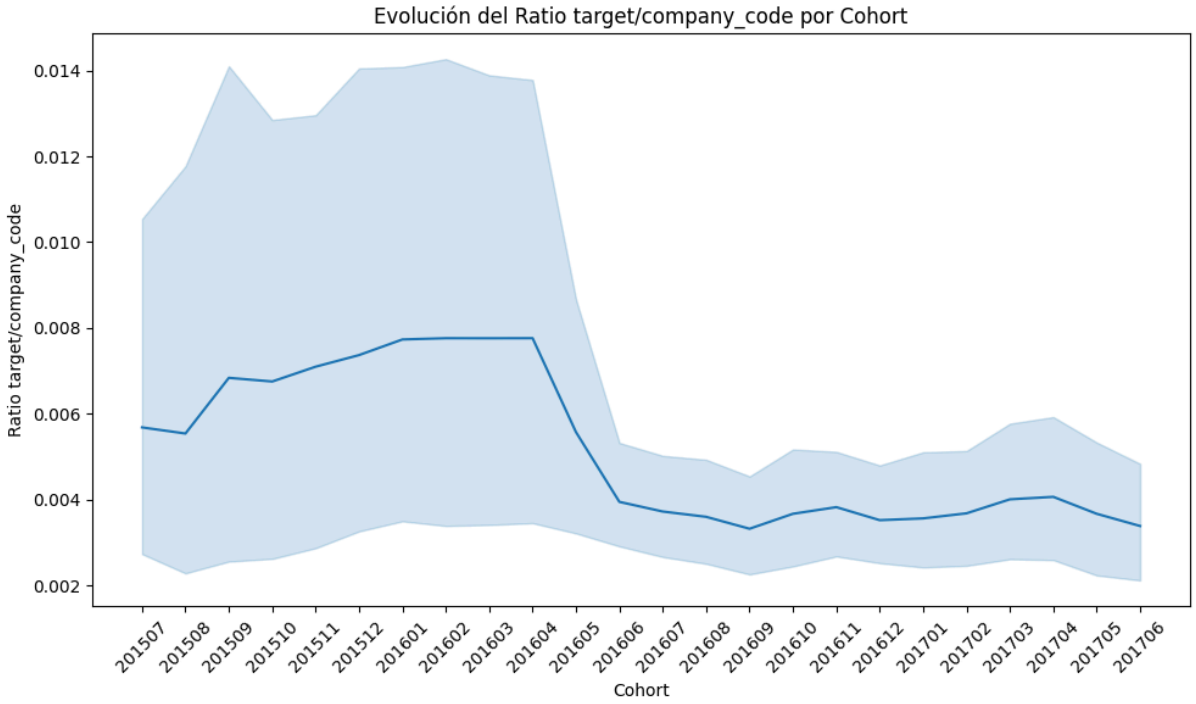
a) Variable Objetivo

La variable objetivo del modelo es una variable binaria llamada “target” que identifica si un cliente cae en mora, es decir, si incurre en un incumplimiento de pago mayor a 90 días. Esta es una categórica en la cual BU representa a los buenos y MA, a los malos.

TARGET		
BU = 0	87, 672	95.73 %
MA = 1	4,260	4.63 %

La tasa de mora general a lo largo de la base de datos es 4.93%; esto quiere decir que la cantidad de clientes que caen en mora es significativamente menor en comparación con los buenos clientes. Esto concuerda con la intuición económica ya que se espera que clientes que son empresas presenten una tasa de mora menor; a diferencia de personas naturales.

En cuanto a la evolución de la tasa de mora en el tiempo, se presenta el siguiente gráfico.



La evolución histórica de la tasa de mora revela fluctuaciones significativas a lo largo del tiempo. Entre julio de 2015 y mediados de 2016, la tasa experimentó un incremento gradual, alcanzando su máximo a principios de 2016. No obstante, en el transcurso del mismo año, se registró una disminución notable. Posteriormente, durante 2017, la tasa se estabilizó. Este comportamiento refleja posibles ajustes en las políticas de crédito o cambios en la calidad crediticia de los clientes.

Esta reducción en la tasa de mora puede ser un indicador positivo de la salud económica y la efectividad de las políticas de gestión. Factores como el crecimiento económico y la educación financiera pueden haber contribuido a esta tendencia. Una disminución sostenida impacta positivamente el rendimiento del modelo de rating, mejorando el perfil de riesgo de los clientes. Sin embargo, si esta reducción es abrupta, puede ser necesario ajustar el modelo para reflejar adecuadamente el entorno crediticio. Utilizar la

tendencia a nuestro favor es beneficioso, pero es esencial realizar un análisis cuidadoso para mantener la robustez del modelo.

Finalmente, los datos muestran consistentemente una mayor cantidad de clientes que cumplen con sus pagos en comparación con aquellos que caen en mora. Esto refuerza la tendencia positiva mencionada, ya que en todas las cohortes analizadas, la proporción de clientes morosos se mantiene por debajo del 6%, mientras que más del 94% de los clientes cumplen con sus pagos en cada periodo. Por ejemplo, en la cohorte de enero de 2016, de un total de 3,665 clientes, solo 194 (5.59%) incurrieron en mora, lo que refleja que la mayoría de los clientes mantienen un buen comportamiento de pago.

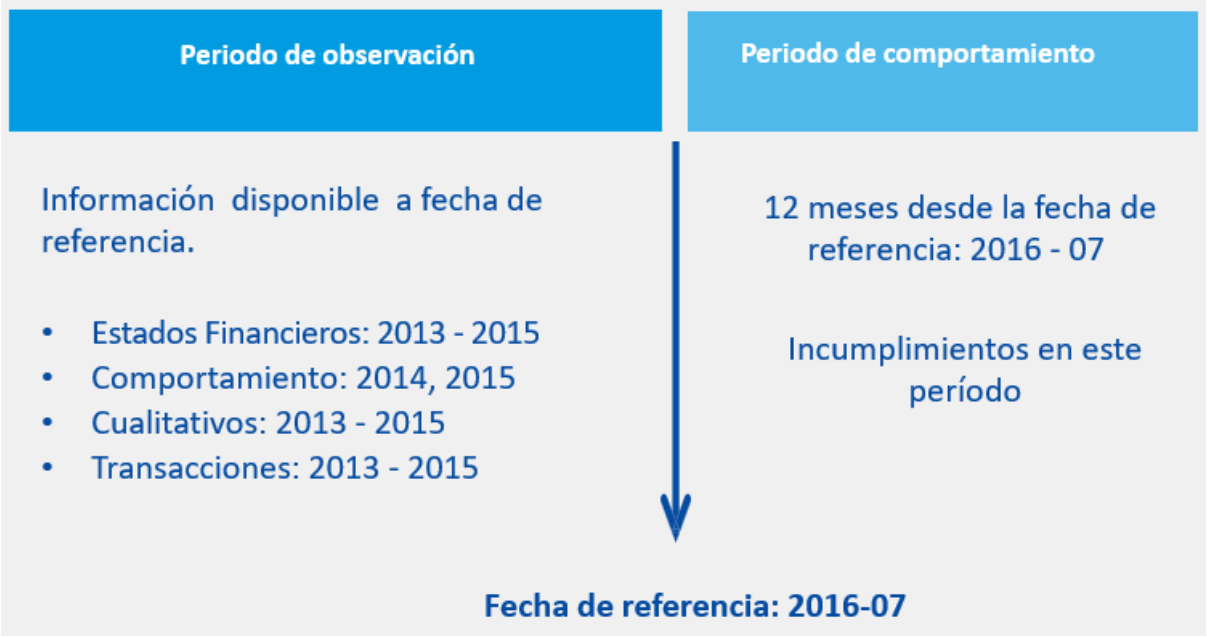
b) Población objetivo

La población objetivo del modelo está compuesta por empresas que han accedido a productos financieros y cuentan con un historial crediticio. Estas empresas pertenecen a los sectores de comercio, industria y servicios.

c) Ventana Temporal

La ventana temporal del modelo, que se extiende desde julio de 2015 hasta junio de 2016, permite observar el comportamiento de la variable objetivo durante un periodo lo suficientemente amplio para capturar tendencias significativas en el riesgo de incumplimiento de los clientes. Este marco temporal no solo proporciona una visión general del comportamiento crediticio, sino que también garantiza que el análisis de la variable objetivo sea óptimo. Al considerar un plazo de un año para evaluar el comportamiento de las empresas, desde julio de 2016 hasta junio de 2017, se asegura que la proporción de operaciones “malas” (morosas) sea representativa y relevante para la modelación, lo que contribuye a una mejor clasificación y pronóstico del riesgo crediticio.

d) Diseño del Experimento



e) Filtros

A partir del código CIU, se ha depurado una variedad de sectores económicos, quedando en la base de datos únicamente los sectores de comercio, industria y servicios. Las empresas institucionales han sido excluidas porque son organizaciones sin fines de lucro y, por ende, no generan ingresos de la misma

manera que las empresas comerciales. Las empresas promotoras e inmobiliarias también se han eliminado, ya que sus estados financieros a menudo reflejan ventas nulas. Por otro lado, las entidades públicas han sido descartadas debido a su respaldo financiero significativo, lo que les permite cumplir con sus obligaciones de pago de manera más consistente. Asimismo, sectores como el agrícola han sido excluidos, ya que su desempeño financiero suele estar sujeto a factores externos, como las condiciones climáticas, que afectan su estabilidad económica. Cabe mencionar que para la realización del presente trabajo se ha usado la base de datos provista en clase; la cual cuenta con 91932 registros y 57 variables de 5825 clientes diferentes.

2. Fuentes de Información

En nuestra base de datos tenemos información de 57 variables; sin embargo no son variables explicativas. Estas se dividen entre 4 categóricas (discretas) y 46 numéricas (continuas). Por otro lado, las fuentes de información se dividen en 4 bloques dependiendo del tipo de información que brindan.

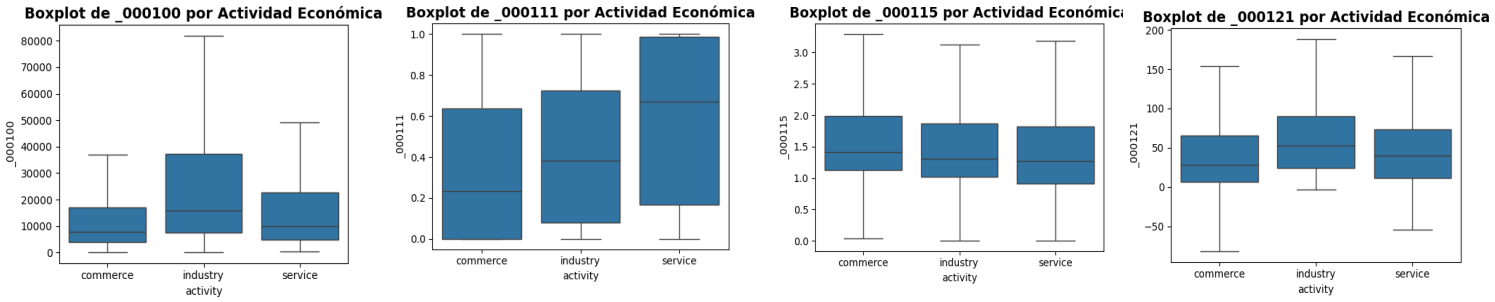
- BLOQUE 1: ESTADOS FINANCIEROS
  Este es el bloque con mayor cantidad de variables, por lo que sólo se muestran las principales.
  - _000115: Ratio de Liquidez General (Activo Corriente / Pasivo Corriente)
  - _000116: Ratio de Prueba Ácida ((Activo Corriente - Existencias) / Pasivo Corriente)
  - _000130: Ratio de Endeudamiento (Patrimonial Pasivo / Patrimonio)
  - _000131: Ratio de Endeudamiento (Pasivo / Activo)
  - _000153: Rentabilidad sobre el Patrimonio (ROE) = Utilidad Neta / Patrimonio
  - _000154: Rentabilidad sobre los Activos (ROA) = Utilidad Neta / Activo
- BLOQUE 2: COMPORTAMIENTO
  - _000409: Incumplimiento de pagos de 8 días en los últimos meses
  - _000410: Incumplimiento de pagos de 15 días en los últimos meses
  - _000416: Promedio de sobregiros pagados a tiempo
  - _000417: Promedio de sobregiros no pagados a tiempo
  - _000423: Incumplimiento de pagos de 8 días en los últimos años
  - _000424: Incumplimiento de pagos de 15 días en los últimos años
- BLOQUE 3: CUALITATIVO
  - _000200: Cuota de Mercado
  - _000204: Relación con Proveedores
  - _000205: Relación con Clientes
  - _000210: Control / Accionariado de Empresa
- BLOQUE 4: TRANSACCIONAL
  - _000304: Variación de cobros por parte de la empresa en el último año
  - _000305: Variación de pagos por parte de la empresa en el último año
  - _000306: Variación de cobros particulares por parte de la empresa en el último año
  - _000307: Variación de pagos particulares por parte de la empresa en el último año

Además, se podrían considerar otros bloque de información como por ejemplo uno sobre relaciones comerciales en el que incluyan variables que brinden información sobre la relaciones de la empresa.

3. Segmentación o Sectorización

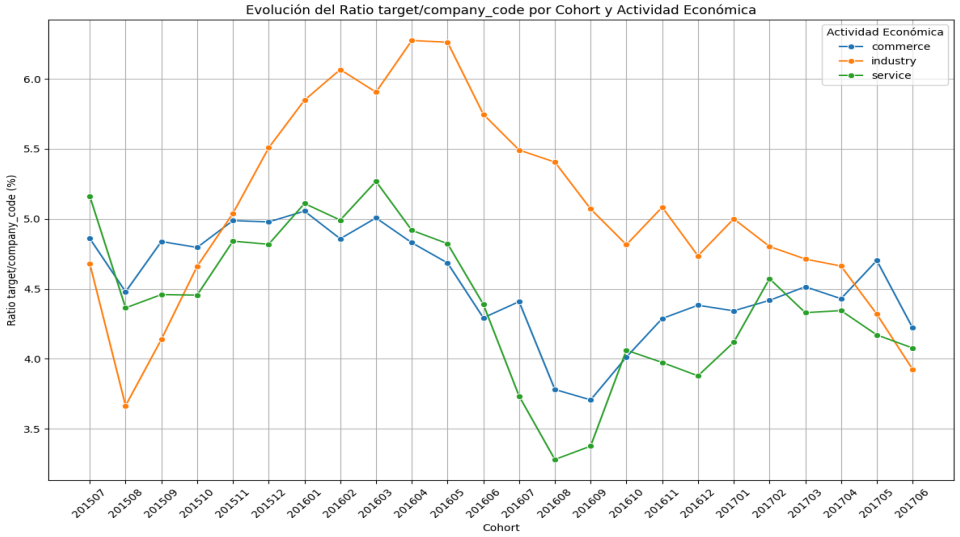
La segmentación es dividir la base de datos en segmentos de acuerdo a un criterio; esta partición se debe a que cada segmento muestra un comportamiento diferenciado, por lo que evaluar todo en su conjunto en un solo modelo no sería lo óptimo. En el presente apartado se evalúa y plantea la posibilidad de desarrollar modelos diferenciados para segmentos.
Esta segmentación podría darse según el tipo de actividad que las empresas realizan, en nuestro caso tenemos 3 sectores económicos: comercio, industria y servicios. De acuerdo a lo visto en el código se han tomado las siguientes gráficas, las cuales son representativas de comportamientos diferenciados entre estos sectores.

- Comportamiento de variables



Las diferencias observadas en los boxplots reflejan las características específicas de cada sector económico. En la variable activo (100), la mediana más alta en industria sugiere que estas empresas requieren mayores inversiones en activos fijos, mientras que servicios y comercio dependen más de activos intangibles. Para el ratio de obligaciones financieras estructurales (111) el sector servicios presenta la mediana más alta, lo que indica un mayor uso de financiamiento a largo plazo, posiblemente debido a la necesidad de estabilidad financiera, mientras que la industria se sitúa en un nivel intermedio y comercio en el más bajo, reflejando un menor apalancamiento. En el ratio de liquidez general (115) aunque las cajas son de tamaño similar, la mayor mediana de comercio podría indicar una gestión más eficiente del capital de trabajo en comparación con industria y servicios. Finalmente, el periodo de cobro de clientes (121) muestra que la mediana de industria es la más alta, lo que puede reflejar ciclos de cobro más largos en comparación con el servicio y el comercio, que tienden a tener procesos más rápidos.

- Comportamiento de la mora



El comportamiento observado en la mora puede deberse a factores económicos específicos que afectaron a cada sector en esos periodos. El crecimiento en la mora del sector industrial entre agosto de 2015 y abril de 2016 podría estar relacionado con dificultades en la producción o la caída de la demanda, lo cual incrementó el riesgo crediticio. Por otro lado, la tendencia similar en la mora de los sectores de industria y servicios hasta mayo de 2016 sugiere que ambos sectores podrían haber enfrentado condiciones macroeconómicas adversas. La posterior reducción de la mora en el sector servicios a niveles bajos (menores a 3.5) puede indicar una recuperación más rápida en la capacidad de pago de las empresas de servicios, posiblemente por una mejora en la demanda de sus servicios.

Entonces, en base a los análisis realizados se encontró diferencias significativas entre los sectores comercio, industria y servicios; por lo que plantear modelos diferentes para cada sector es adecuado. Cabe mencionar que otra posible solución es realizar una sectorización; a través de la cual se

estandarizan las variables para hacerlas comparables. De esta manera no se partiría la base de datos según la actividad que realicen las empresas,

4. Análisis Univariante

A continuación se realizará el análisis univariante de las variables explicativas. En el caso de las numéricas se evalúa las siguientes estadísticas: máximo, mínimo, media, desviación estándar y los percentiles (25%, 50% y 75%); en el caso de las categóricas se evalúan las frecuencias.

a. Variables Numéricas

El análisis descriptivo de los datos muestra algunos indicios de variables con grandes desviaciones y anomalías en los datos.

Se identificó variables como 100 (activo), 102 (Obligaciones Financieras), y 124 (Periodo de pago a proveedores) que presentan desviaciones estándar muy elevadas en comparación con sus medias. Por ejemplo, 100 tiene una media de 24,680.85 pero una desviación estándar de 48,895.03, lo que indica una gran dispersión en los datos.
En segundo lugar, se encontraron anomalías o valores sospechosos. En el caso de la variable “age” se presenta un valor mínimo de -3, lo cual es improbable debido a que esta variable indica los años de operación que tiene la empresa. Si bien es posible que esta anomalía se deba a un error en la base de datos, también puede hacer referencia a los años de inversión previos al comienzo de operaciones.

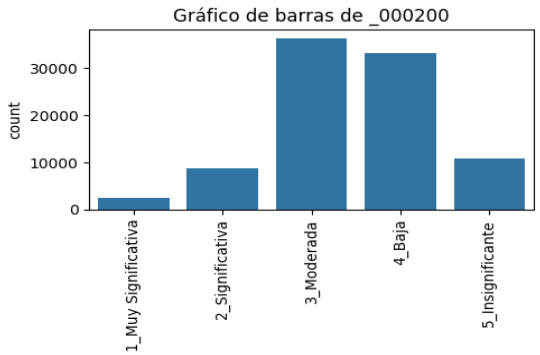
b. Variables Categóricas

i. Análisis Estadístico

Con respecto al análisis estadístico de las categóricas se muestra la siguiente tabla que detalla la categoría más común de cada una.

VARIABLE	CANTIDAD DE CATEGORÍAS	CATEGORÍA MÁS COMÚN	FRECUENCIA
Target	2	BU = bueno	87 672
Activity	3	commerce = comercio	38 671
200: Cuota de mercado	5	3_ Moderada	36 408
204: Relación con proveedores	3	1_ Amplia Variedad	79 555
205: Relación con clientes	3	4_ Amplia Variedad	85 873
210: Control / Accionariado de la empresa	5	4_Empresas restantes	42 021

En el caso de la variable 200 - cuota de mercado la distribución de la frecuencia se presenta en el siguiente gráfico. Se puede observar que las categorías de muy significativa, significativa e insignificante se pueden agrupar en una sola, con la finalidad de mantener una volumetría regular entre las 3 categorías resultantes.



c. Detección de valores atípicos

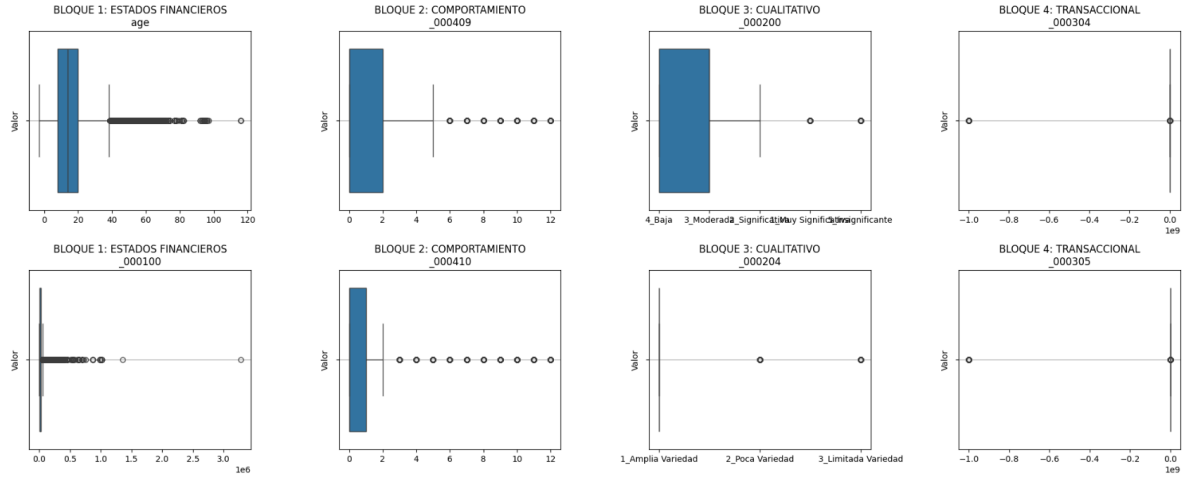
i. Duplicados

En cuanto a valores duplicados, se resalta que no se encontraron filas duplicadas en la base de datos.

ii. Outliers

Con respecto a los outliers, estos se identificaron usando la regla del rango intercuartílico IQR, la cual define como outliers a los valores que se ubican por debajo de Q1 - 1.5 * IQR y, también, a los que están por encima Q3 + 1.5 * IQR. En este caso se define el umbral (IQR) de 1%. Por ejemplo para el caso de la variable age se encontró 4589 outliers. En general el promedio de outliers presentes en cada variable numérica es el 8.64%

Con respecto al tratamiento de outliers se plantea como opción eliminarlos debido a que alteran la distribución de las variables.



Como parte de un análisis adicional se muestra este gráfico, en el cual se muestran los boxplots de acuerdo a los bloque de información, estos corresponden a la variable “age” y 100 (activo). Para ambos casos se puede observar una mayor cantidad de outliers en los bloques de estados financieros y variables comportamentales.

iii. Valores nulos

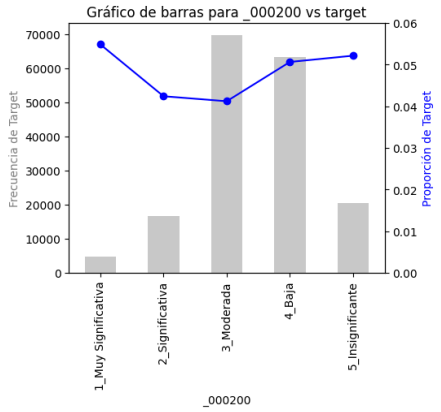
Además de lo que se identificó anteriormente, se encontraron valores como nulos (NaN) los cuales pueden ser identificados por presentarse como -99999999 o como NaN. En el primer caso se encontró que en general el promedio de valores -999999999.00 por variable es 3962.74, es decir el 4.31%.En el segundo caso, variables como 100-activo presenta un porcentaje de 14% de missings; el promedio general de valores NaN por variable: 1580.74, lo cual es el 1.72%.

Respecto al tratamiento de los valores nulos se plantea como opción eliminarlos debido a que no representan un porcentaje alto del total de las observaciones de la base de datos.

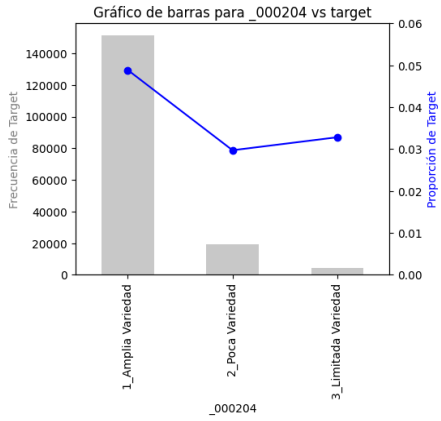
5. Análisis Bivariante

Variables Categóricas

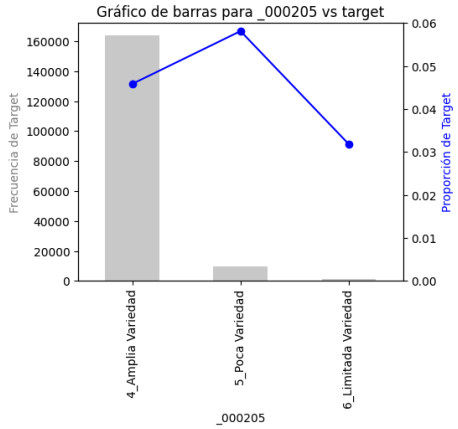
El análisis de las variables relacionadas con la tasa de mora revela patrones importantes.



En el caso de la cuota de mercado (Variable 200), su valor de IV es 0.01, lo que la clasifica como no predictiva. Esto se debe a que no tiene una relación lógica que respalde la idea de que las empresas con una cuota de mercado muy significativa presenten mayores tasas de mora en comparación con aquellas con una cuota de mercado moderada o significativa.

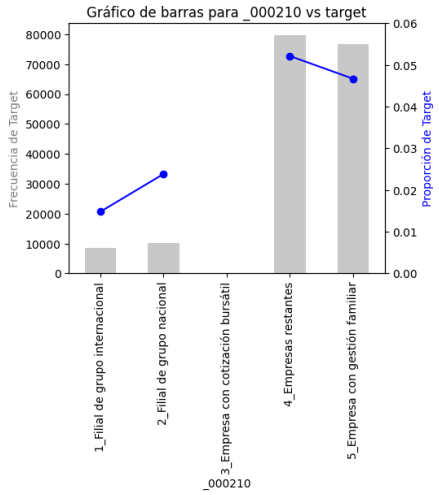


Respecto a la relación con proveedores (Variable 204), con un IV de 0.02, se encuentra en el límite entre no predictiva y débilmente predictiva. El análisis muestra que las empresas con una mayor diversidad en sus relaciones con proveedores experimentan una tasa de mora más elevada, cercana al 5%. En contraste, aquellas con una variedad limitada de relaciones presentan tasas de mora más bajas, alrededor del 3%.



En cuanto a la relación con clientes (Variable 205), su IV es de 0.00, lo que la califica como una variable no predictiva. A pesar de que, en teoría, se esperaría que las empresas con una relación limitada con sus

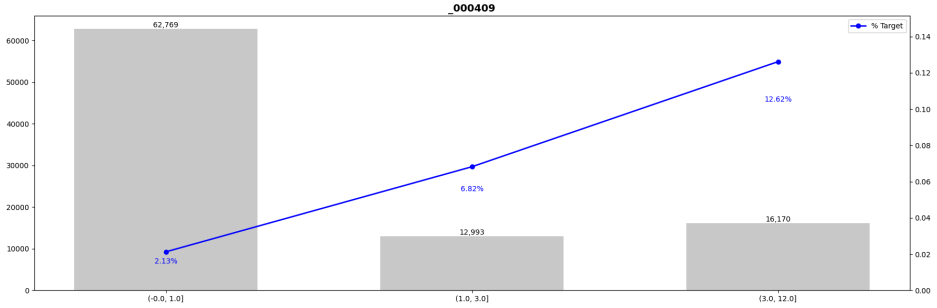
clientes tendieran a mostrar una mayor tasa de mora que aquellas con una relación más amplia, este comportamiento no se refleja en los datos.



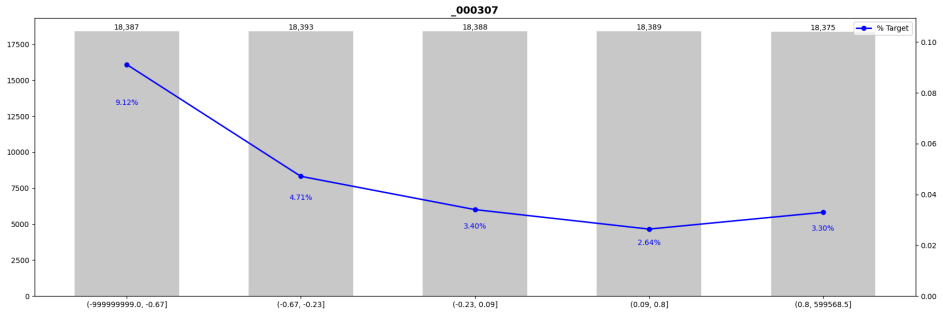
Por último, el control accionario (Variable 210) presenta el IV más alto entre las variables categóricas, con un valor de 0.07, lo que lo sitúa dentro de la categoría de predicción débil. Las empresas gestionadas de manera familiar y otras de naturaleza similar presentan tasas de mora más altas (5% y 4%, respectivamente), mientras que las filiales de grupos internacionales y nacionales tienen tasas significativamente más bajas, cercanas al 1% y 2%.

Variables Numéricas

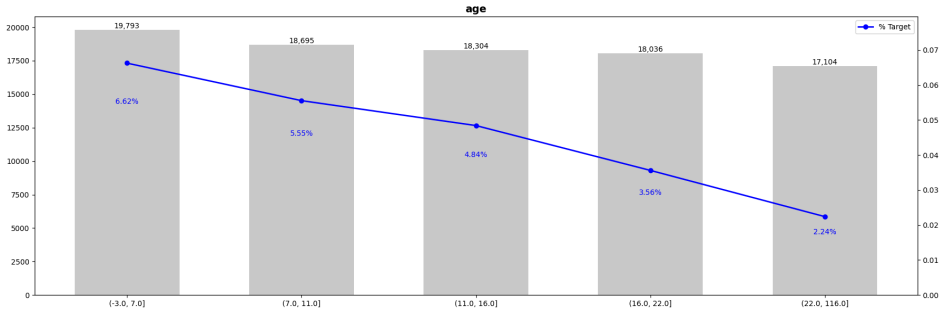
El análisis de las variables numéricas relacionadas con la tasa de mora revela patrones importantes en su relación con el incumplimiento de pago.



Respecto al incumplimiento de pago de 8 días en los últimos meses (Variable 409), esta es una variable que presenta una tendencia creciente. Tiene un valor de IV (Information Value) alto de 0.687, lo que la clasifica como una variable altamente predictiva. Esto es lógico desde el punto de vista financiero, ya que a mayor número de impagos de 8 días en los últimos meses, mayor es la probabilidad de caer en mora.



En cuanto a la variación de pagos particulares por parte de la empresa en el último año (Variable 307), esta es una variable con comportamiento fluctuante. Presenta un IV de 0.239, lo que indica que tiene un poder predictivo moderado. La forma de la gráfica se justifica por la naturaleza de la variable, que refleja una variación en los pagos.



Finalmente, respecto a la edad de la empresa (Variable Age), esta es una variable decreciente en términos de riesgo de mora. Cuenta con un IV de 0.134, lo que también la clasifica como una variable con poder predictivo moderado. Esto tiene sentido lógico, ya que a mayor edad de una empresa, es probable que tenga más estabilidad y recursos financieros, lo que reduce la probabilidad de incumplimiento de pagos.

6. Análisis Multivariante

- **Correlación**

En el análisis multivariante analiza la correlación entre las variables explicativas; debido a que se quiere evitar la presencia de multicolinealidad en el modelo, es decir que el modelo no tenga variables que reporten la misma información. La multicolinealidad puede llevar a estimaciones inexactas de los coeficientes de regresión, dificultando la interpretación de los resultados y aumentando la varianza de las estimaciones. Para este análisis se realizó un mapa de calor, el cual facilita la identificación de variables altamente correlacionadas. Cabe mencionar que se estableció un umbral de 0.7 para la correlación. Este valor es comúnmente utilizado en la literatura estadística para identificar relaciones fuertes entre variables, lo que permite excluir aquellas que proporcionan información redundante y garantizar así la estabilidad y precisión del modelo (Cohen, 1988). Un análisis exhaustivo de las correlaciones ayuda a optimizar la selección de variables, lo que puede mejorar el rendimiento del modelo y evitar problemas en la generalización de los resultados. El mapa de calor y las variables correlacionadas, por encima de 0.70, se pueden observar en la sección de Anexos.

Se halló una correlación de 0.9 entre las variables 102 (Obligaciones Financieras Totales) y 110 (Obligaciones Financieras Estructurales - Largo Plazo). Esta alta correlación se debe a que ambas variables están relacionadas con la deuda financiera de la empresa, ya que las obligaciones financieras estructurales son una parte importante de las obligaciones financieras totales. Por lo tanto, cuando aumenta el total de la deuda financiera, es muy probable que también aumenten las obligaciones a largo plazo, lo que genera una alta correlación entre ambas. De manera similar, se encontró una correlación de

0.9 entre las variables 409 (Incumplimiento de pagos de 8 días en los últimos meses) y 423 (Incumplimiento de pagos de 8 días en los últimos años). Esto ocurre porque los incumplimientos recientes tienden a reflejar patrones históricos de incumplimiento, indicando que una empresa con incumplimientos recientes también es probable que haya tenido problemas de pagos en el pasado.

Cabe mencionar que si bien el análisis de correlaciones nos brinda las variables que reportan la misma información se utiliza adicionalmente como criterio de elección de variables al IV (Information Value). Esto quiere decir que la variable que posea un mayor IV es la variable que no se elimina.

7. MODELADO

a. Selección de Variables

Las variables seleccionadas para el modelo se eligieron mediante un análisis exhaustivo que incluyó análisis univariante, bivariante y multivariante. En el análisis univariante, se optó por variables con alta calidad de información y un comportamiento coherente. El análisis bivariante verificó que las variables candidatas presentaran una tendencia clara hacia la variable objetivo. Luego, se estudió la correlación, estableciendo un umbral de 0.70 para identificar pares altamente correlacionados. Se filtraron variables numéricas y categóricas con un Índice de Información (IV) entre 0.1 y 0.6, indicando que son informativas pero no redundantes. A partir de la matriz de correlación, se eliminaron las variables con el IV más bajo en cada par correlacionado, priorizando las más informativas. En los anexos se incluye la tabla de las variables preseleccionadas.

b. Modelado 1:

En este primer modelado, utilizaremos la regresión logística para examinar la relación entre las variables seleccionadas y la probabilidad de default, basándonos en los tramos definidos en el análisis bivariante.
El tramedo asociado directamente a frecuencias que se realizó para el análisis bivariante de las variables numéricas, es el que se utilizará. Este es un trameado que divide en 5 subidiciones a las variables.

i. Métricas

- Area Under Curve

El área debajo de la curva ROC es 0.82, el cual indica que el modelo tiene una buena capacidad para distinguir entre las clases de default (1) y no default (0).. Un AUC superior a 0.8 se considera excelente, lo que refleja que el modelo es eficaz en la predicción de incumplimientos de pago.

Cabe recalcar que en las variables categóricas la cuota de mercado (Variable 200) tiene un IV de 0.01, considerándose no predictiva. La relación con proveedores (Variable 204), con un IV de 0.02, está en el límite de no predictiva y débilmente predictiva. La relación con clientes (Variable 205) muestra un IV de 0.00, también no predictiva. En contraste, el control accionario (Variable 210) tiene el IV más alto entre las variables categóricas, con un valor de 0.07, lo que indica una predicción débil.

- Coeficiente Gini

El coeficiente de Gini de 0.63 muestra una buena capacidad discriminativa del modelo. Este valor, que varía entre 0 y 1, indica que el modelo tiene un rendimiento significativamente mejor que el azar. Un Gini de 0.63 sugiere que el modelo puede identificar de manera efectiva los riesgos de default.

- Matriz de Confusión

La matriz de confusión muestra que el modelo identificó correctamente 17,489 casos de no default y 47 casos de default. Sin embargo, hay 809 falsos negativos, lo que indica que el modelo tiene dificultades para identificar los casos de default. Esto resalta un desafío importante en su capacidad de predicción.

- Otras métricas

La precisión del modelo para la clase 0 es de 0.96, indicando que el 96% de las predicciones de no default son correctas, mientras que la precisión para la clase 1 es de 0.53, lo que sugiere que solo la mitad de los casos de default fueron identificados correctamente; esto resalta la necesidad de mejorar la identificación de la clase positiva. En cuanto a la recuperación, la clase 0 presenta un valor de 1.00, lo que significa que

todos los verdaderos casos de no default fueron detectados, pero la clase 1 muestra una recuperación de apenas 0.05, evidenciando la debilidad del modelo para identificar riesgos de incumplimiento. Finalmente, el F1-score, que combina precisión y recuperación, muestra un valor de 0.98 para la clase 0, reflejando un excelente rendimiento, mientras que para la clase 1, el F1-score es de solo 0.10, lo que pone de manifiesto un rendimiento deficiente en la identificación de defaults y subraya la necesidad de mejorar la detección de la clase positiva.

c. Modelado 2:

En el segundo modelado, combinaremos regresión logística, árboles de decisión y redes neuronales. Nos enfocaremos en las variables con mejor rendimiento, sin limitarnos a los tramos del análisis bivariante. Entonces en este caso se trabajan con las variables originales, no con las variables trameadas.

Con respecto a la selección de variables se han elegido las siguientes:

- _000170 - Ratio de EBITDA a Gastos Financieros
  Este ratio indica la capacidad de una empresa para cubrir sus gastos financieros con sus ganancias antes de intereses, impuestos, depreciación y amortización. Un mayor ratio puede indicar una menor probabilidad de mora, haciendo de esta variable una buena candidata para el modelo.
- age - Años de la empresa en operaciones
  La edad de la empresa está asociada con su estabilidad financiera. A medida que aumenta la edad de una empresa, disminuye el riesgo de incumplimiento de pagos, lo que la convierte en una variable predictiva moderada.
- _000111 - Ratio de Obligaciones Financieras Estructurales (Largo Plazo) y Obligaciones Financieras Totales
  Este ratio proporciona información sobre la estructura del financiamiento de la empresa. Un ratio más alto podría reflejar un mayor riesgo financiero, lo que aumenta la probabilidad de mora.
- _000155 - Flujo de Efectivo
  El flujo de efectivo es crucial para el funcionamiento de la empresa. Un flujo de efectivo positivo puede reducir la probabilidad de incumplimientos, mientras que un flujo negativo puede aumentarla. Su inclusión puede mejorar la capacidad predictiva del modelo.
- _000173 - Variación Anual de las Ventas
  La variación en las ventas puede ser un indicador de la salud financiera de una empresa. Una disminución en las ventas puede estar correlacionada con un mayor riesgo de mora, lo que justifica su selección.
- _000124 - Periodo de Pago a Proveedores (días)
  Esta variable proporciona información sobre la gestión del capital de trabajo de la empresa. Un periodo de pago más largo puede indicar problemas de liquidez, aumentando el riesgo de incumplimiento.
- _000113 - Capital de Trabajo
  El capital de trabajo es esencial para las operaciones diarias de la empresa. Un capital de trabajo positivo sugiere que la empresa tiene suficientes activos a corto plazo para cubrir sus pasivos a corto plazo, lo que puede reducir la probabilidad de mora.

Las variables seleccionadas presentan un buen desempeño tanto en el análisis de WOE como en el cálculo del IV, lo que indica su capacidad predictiva para modelar el riesgo de mora en empresas. Estos indicadores refuerzan la relevancia de estas variables en la evaluación de la estabilidad financiera y la capacidad de pago, mejorando la precisión y efectividad del modelo propuesto.

i. Métricas

Model	AUC Score	Gini Score	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0.65	0.30	0.45	0.01	0.01	0.95

Decision Tree	0.95	0.91	0.84	0.81	0.82	0.98
Neural Network	0.75	0.49	0.85	0.01	0.03	0.95

- Logistic Regression

El modelo de regresión logística muestra un AUC Score de 0.65, lo que indica una capacidad moderada para distinguir entre las clases. A pesar de tener una alta precisión global (0.95), su baja tasa de recall (0.01) sugiere que el modelo tiene problemas significativos para identificar correctamente las instancias positivas, lo que resulta en una alta tasa de falsos positivos (precisión de 0.45). Este bajo desempeño en la identificación de moras implica que, aunque el modelo puede parecer efectivo en general, no es confiable para detectar los casos de incumplimiento.

- Decision Tree

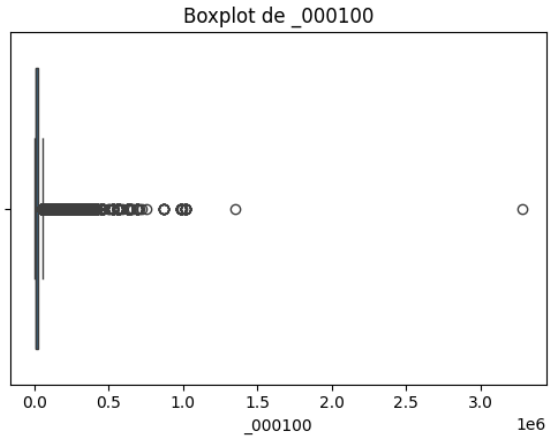
El modelo de árbol de decisión presenta un AUC Score excepcional de 0.95, lo que indica una excelente capacidad para discriminar entre las clases. Con un Gini Score de 0.91, este modelo demuestra una fuerte capacidad predictiva. Además, tiene buenos resultados en precisión (0.84) y recall (0.81), lo que significa que logra identificar correctamente la mayoría de las instancias positivas y negativas. En términos de exactitud, el modelo alcanza un 0.98, lo que refuerza su efectividad en la predicción y su utilidad para la clasificación de moras.

- Neural Network

El modelo de red neuronal muestra un AUC Score de 0.75, lo que indica una capacidad moderada para distinguir entre las clases, mejor que la regresión logística, pero inferior al árbol de decisión. Su precisión es alta (0.85), pero su recall es alarmantemente bajo (0.01), lo que refleja dificultades similares a las de la regresión logística para identificar instancias positivas. A pesar de tener una buena exactitud global (0.95), el bajo F1 Score (0.03) sugiere que el modelo no es confiable para la detección de moras, lo que limita su aplicabilidad en escenarios donde es crítico identificar correctamente los incumplimientos.

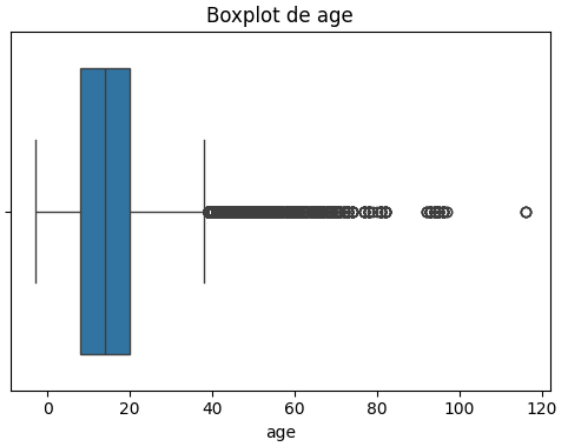
8. Anexo

Anexo 1: Variable 100 - Activo



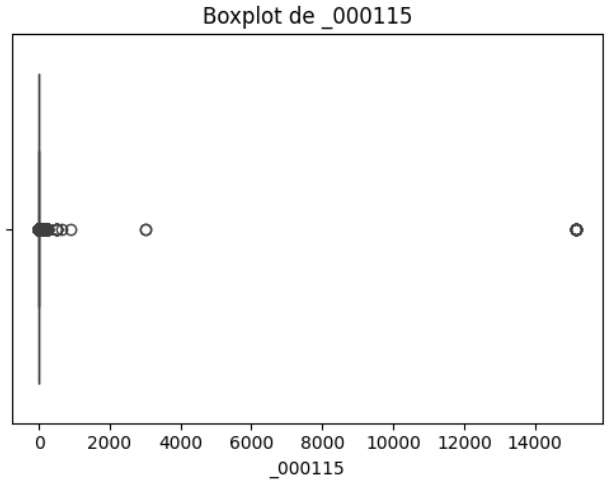
Se muestra una gran dispersión en la distribución.

Anexo 2: Variable Age



Se puede observar que la existencia de valores outliers que se posicionan en la cola de la variable age

Anexo 3 : Variable 115 - Ratio de liquidez general



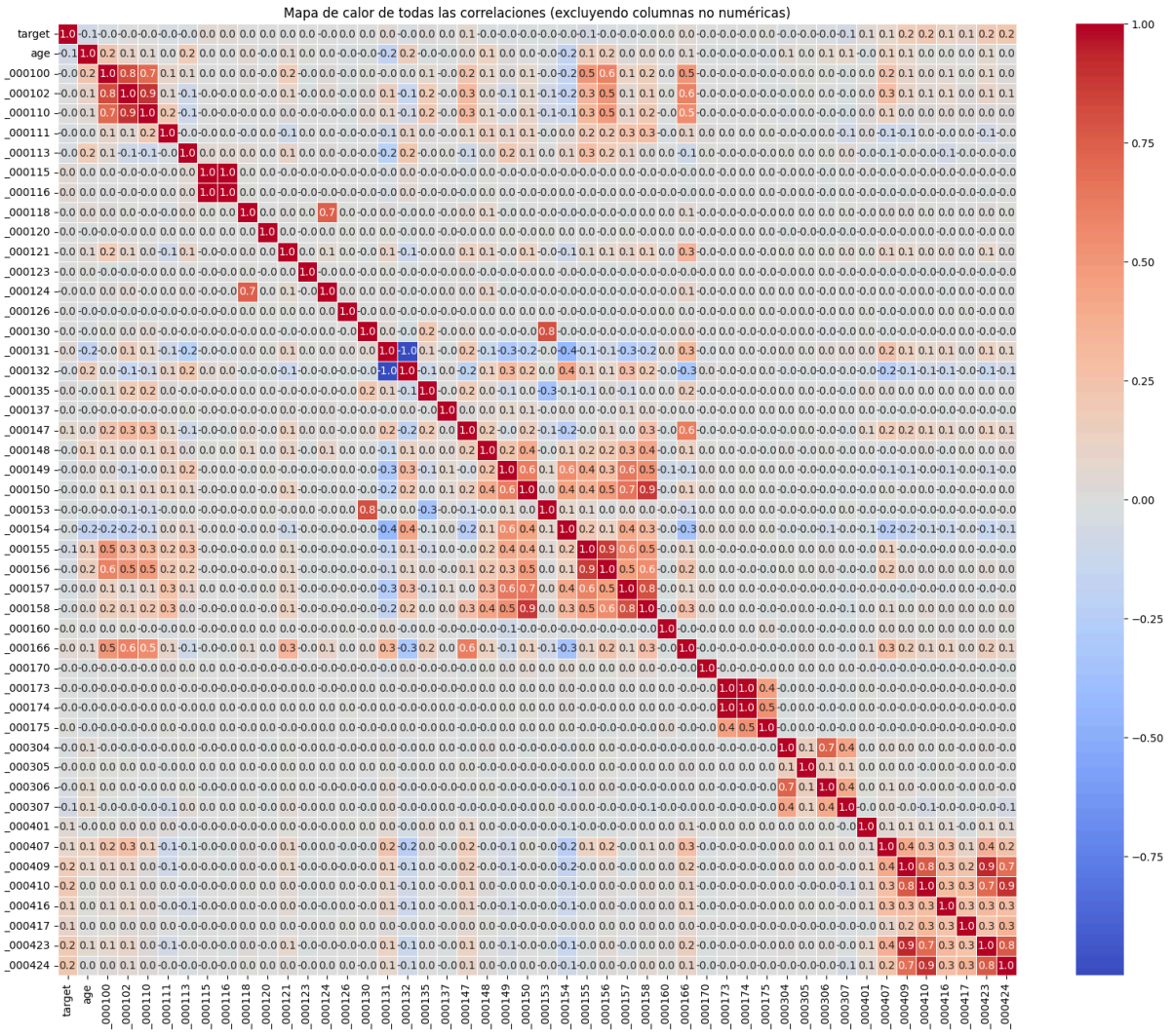
Se puede observar la presencia de outliers que se encuentran fuera de la distribución.

Anexo 4: Tabla de Bloque 1

BLOQUE 1: ESTADOS FINANCIEROS
age: Años de la empresa en operaciones
_000100: Activo
_000102: Obligaciones Financieras Totales
_000110: Obligaciones Financiera Estructurales (Largo Plazo)
_000111: Ratio de Obligaciones Financiera Estructurales (Largo Plazo) y Obligaciones Financieras Totales
_000113: Capital de Trabajo
_000115: Ratio de Liquidez General = Activo Corriente / Pasivo Corriente (Liquidez General)
_000116: Ratio de Prueba Ácida = (Activo Corriente - Existencias) / Pasivo Corriente

_000118: Periodo de Inventario (días)
_000120: Variación Anual del Periodo de Inventario (días)
_000121: Periodo de Cobro de Clientes (días)
_000123: Variación Anual del Periodo de Cobro de Clientes (días)
_000124: Periodo de Pago a Proveedores (días)
_000126: Variación Anual del Periodo de Pago a Proveedores (días)
_000130: Ratio de Endeudamiento = Patrimonial Pasivo / Patrimonio
_000131: Ratio de Endeudamiento = Pasivo / Activo
_000132: Ratio de Propiedad sobre los Activos = Patrimonio / Activo
_000135: Ratio de Obligaciones Financieras Totales respecto al Patrimonio
_000137: Variación Anual del Ratio de Obligaciones Financieras Totales respecto al Patrimonio
_000147: Ratio de Gastos Financieros a Ventas
_000148: Ratio de Utilidad Bruta a Ventas
_000149: Ratio de Utilidad Neta a Ventas
_000150: Ratio de Utilidad Operativa a Ventas
_000153: Rentabilidad sobre el Patrimonio (ROE) = Utilidad Neta / Patrimonio
_000154: Rentabilidad sobre los Activos (ROA) = Utilidad Neta / Activo
_000155: Flujo de Efectivo
_000156: Beneficio Antes de Intereses, Impuestos, Depreciación y Amortización (EBITDA)
_000157: Ratio de Flujo de Efectivo a Ventas
_000158: Ratio de EBITDA a Ventas
_000160: Ratio de Obligaciones Financieras a Flujo de Efectivo
_000166: Ratio de Obligaciones Financieras y Comerciales a Ventas
_000170: Ratio de EBITDA a Gastos Financieros
_000173: Variación Anual de las Ventas
_000174: Variación Anual de los Activos
_000175: Variación Anual de los Activos No Corrientes

Anexo 5: Mapa de Calor



Anexo 6: Tabla de variables correlacionadas (por encima de 0.70) ordenadas de mayor a menor:

Variable 1 (Código = Nombre)	Variable 2 (Código = Nombre)	Correlación
_000131 = Ratio de Endeudamiento = Pasivo / Activo	_000132 = Ratio de Propiedad sobre los Activos = Patrimonio / Activo	-1.0
_000115 = Ratio de Liquidez General = Activo Corriente / Pasivo Corriente	_000116 = Ratio de Prueba Ácida = (Activo Corriente - Existencias) / Pasivo Corriente	1.0
_000173 = Variación Anual de las Ventas	_000174 = Variación Anual de los Activos	1.0
_000409 = Incumplimiento de pagos de 8 días en los últimos meses	_000423 = Incumplimiento de pagos de 8 días en los últimos años	0.9
_000102 = Obligaciones Financieras Totales	_000110 = Obligaciones Financiera Estructurales (Largo Plazo)	0.9
_000410 = Incumplimiento de pagos de 15 días en los últimos meses	_000424 = Incumplimiento de pagos de 15 días en los últimos años	0.9
_000150 = Ratio de Utilidad Operativa a Ventas	_000158 = Ratio de EBITDA a Ventas	0.9
_000155 = Flujo de Efectivo	_000156 = Beneficio Antes de Intereses, Impuestos, Depreciación y Amortización (EBITDA)	0.9
_000100 = Activo	_000102 = Obligaciones Financieras Totales	0.8

_000157 = Ratio de Flujo de Efectivo a Ventas	_000158 = Ratio de EBITDA a Ventas	0.8
_000130 = Ratio de Endeudamiento = Patrimonial Pasivo / Patrimonio	_000153 = Rentabilidad sobre el Patrimonio (ROE) = Utilidad Neta / Patrimonio	0.8
_000409 = Incumplimiento de pagos de 8 días en los últimos meses	_000410 = Incumplimiento de pagos de 15 días en los últimos meses	0.8
_000423 = Incumplimiento de pagos de 8 días en los últimos años	_000424 = Incumplimiento de pagos de 15 días en los últimos años	0.8
_000118 = Periodo de Inventario (días)	_000124 = Periodo de Pago a Proveedores (días)	0.7
_000304 = Variación de cobros por parte de la empresa en el último año	_000306 = Variación de cobros particulares por parte de la empresa en el último año	0.7
_000410 = Incumplimiento de pagos de 15 días en los últimos meses	_000423 = Incumplimiento de pagos de 8 días en los últimos años	0.7
_000100 = Activo	_000110 = Obligaciones Financiera Estructurales (Largo Plazo)	0.7

Anexo 7: Variables preseleccionadas

Nombre de Variable	**Descripción**
age	Años de la empresa en operaciones
_000111	Ratio de Obligaciones Financiera Estructurales (Largo Plazo) y Obligaciones Financieras Totales
_000135	Ratio de Obligaciones Financieras Totales respecto al Patrimonio
_000147	Ratio de Gastos Financieros a Ventas
_000155	Flujo de Efectivo
_000160	Ratio de Obligaciones Financieras a Flujo de Efectivo
_000170	Ratio de EBITDA a Gastos Financieros
_000173	Variación Anual de las Ventas
_000305	Variación de pagos por parte de la empresa en el último año
_000306	Variación de cobros particulares por parte de la empresa en el último año
_000307	Variación de pagos particulares por parte de la empresa en el último año
_000401	Grado de riesgo crediticio
_000407	Número de entidades acreedoras

9. Bibliografía

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.