



Facultad de Ciencias Económicas y Empresariales  
ICADE

# **LEADING INDICATOR DEL SECTOR BANCARIO EN ESTADOS UNIDOS**

Autor: Pablo Silvano Andreu Matas  
Director: Peter Guenther Antoon Claeys

MADRID | Marzo 2024

## Resumen

Este trabajo presenta un modelo de regresión logística para predecir la probabilidad de que un banco estadounidense caiga en insolvencia. Se analiza la información de 48 bancos entre 2018 y 2023, incluyendo 11 que entraron en *default*, 5 de ellos en 2023.

Para desarrollar el modelo, se seleccionaron 15 ratios financieros y 2 variables externas (tipos de interés y *quantitative easing*) como variables independientes. Se utilizó un análisis de componentes principales o PCA para reducir la dimensionalidad del conjunto de datos y se construyó el modelo de regresión logística con las variables relevantes. El modelo tiene una alta capacidad predictiva, con un AUC de 0.944. Las variables que más afectan a la probabilidad de *default* son: *Return on Average Assets*, *Net Charge-Offs / Average Loans*, *Common Equity Tier 1 Capital Ratio*, *Tangible Common Equity / Tangible Assets*, *Interest Rate* y *Quantitative Easing*. El modelo identifica correctamente 7 de los 9 bancos que entraron en *default* en 2023.

A partir de los resultados, se identifican dos bancos que podrían estar en riesgo de *default*: US Bancorp y Columbia Bank. Se recomienda actualizar el modelo con datos nuevos y utilizarlo junto con otras herramientas de análisis para una mejor toma de decisiones.

Es importante mencionar las limitaciones del estudio, como la disponibilidad de datos y la generalización del modelo a otros países o contextos económicos. Se recomienda ampliar la base de datos con información de más bancos y variables, e implementar el modelo como un *leading indicator* para la identificación de bancos en riesgo.

## Tabla de Contenidos

<b>1. Introducción</b>	<b>4</b>
<b>2. Resumen de los <i>Defaults</i> del 2023</b>	<b>6</b>
2.1 Silicon Valley Bank (SVB) el 10 de Marzo	6
2.2 Signature Bank el 12 de Marzo	7
2.3 First Republic el 1 de Mayo	7
2.4 Heartland Tri-State Bank el 28 de Julio	8
2.5 Citizens Bank (of Sac City) el 3 de Noviembre	8
<b>3. Recolección de Datos y Explicación del Dataset</b>	<b>9</b>
3.1 Variable Dependiente	9
3.2 Ratios Financieros	10
3.3 Variables Externas	11
3.4 Limpieza Manual del <i>Dataset</i>	12
<b>4. El Modelo</b>	<b>15</b>
4.1 Preparación Previa	15
4.2 PCA	16
4.3 <i>Training</i> y <i>Test Sets</i>	18
4.4 Explicación del Modelo	19
<b>5. Interpretación del Modelo</b>	<b>20</b>
5.1 Resumen Estadístico	20
5.2 Curva ROC y AUC	22
5.3 Matriz de Confusión	22
5.4 Variables que Más Afectan al Modelo	23
<b>6. Análisis de Bancos en Riesgo</b>	<b>26</b>
6.1 Falsos Negativos (Error Tipo II)	26
6.2 Falsos Positivos (Error Tipo I)	27
6.2.1 US Bancorp	28
6.2.2 Columbia Bank	29
<b>7. Conclusiones</b>	<b>30</b>
<b>8. Declaración de Inteligencia Artificial</b>	<b>31</b>
<b>9. Bibliografía</b>	<b>32</b>
<b>10. Anexo</b>	<b>37</b>
10.1 Anexo 1: Lista Inicial de Todos los Bancos Utilizados	37
10.2 Anexo 2: Antiguos Estudios de la Probabilidad de Default de los Bancos	38

## 1. Introducción

En el ámbito financiero, la predicción de la insolvencia bancaria se ha vuelto esencial para anticipar y gestionar posibles crisis en el sector y, es por ello por lo que cada vez las instituciones financieras prestan más atención a la gestión del riesgo, estimando que, para el año 2025, las funciones del *risk management* sean el elemento crucial del éxito de los bancos (Härle et al, 2015). Este trabajo de fin de grado se centrará en el ámbito bancario de Estados Unidos con el objetivo de construir un modelo o *leading indicator* capaz de identificar relaciones o patrones entre los bancos que hayan caído en insolvencia y prever con precisión caídas potenciales o riesgos inminentes en bancos que hoy en día siguen funcionando de manera normal.

En este trabajo se va a realizar un análisis exhaustivo del sector bancario estadounidense, utilizando un conjunto de datos extraídos de Factset, Refinitiv, CapitalIQ y Bloomberg que abarca a más de 30 bancos en pleno funcionamiento hasta la fecha. La selección de estos bancos se ha realizado considerando su relevancia en el panorama financiero actual. Además, se ha incorporado a la muestra más de 10 bancos que han caído en *default* en los últimos años con el objetivo de que el modelo pueda identificar qué características o variables son la causa de la inestabilidad financiera en Estados Unidos. Se han extraído datos específicamente de la solvencia, liquidez, rentabilidad y calidad de activos de estos bancos, reconociendo la importancia de estos ratios como indicadores de la salud bancaria.

La determinación de si un banco ha caído en *default* no es siempre un asunto claro y sencillo. En ciertos casos, cuando los depositantes comienzan a retirar fondos masivamente y el banco experimenta clara insolvencia, la Reserva Federal puede optar por permitir que el banco se desplome por completo o incluso cerrarlo para evitar mayores pérdidas. En estas situaciones, la señal de *default* es clara: el banco cesa sus operaciones. Sin embargo, surgen debates en otros escenarios donde la situación es más ambigua. Por ejemplo, cuando un banco es adquirido por otra entidad o la Fed decide intervenir y rescatarlo. En el caso de Credit Suisse, por ejemplo, no se han pagado todavía los CDS o *credit default swaps* porque no ha ocurrido un *bankruptcy event* (Lash, 2023). En este trabajo, no se discutirá sobre si un banco ha colapsado o no, sino que adoptaremos el criterio de la FDIC (*Federal Deposit Insurance Corporation*). Utilizaremos los bancos que la FDIC ha identificado y publicado como *default*, ofreciendo así un enfoque claro y específico para nuestro análisis.

También considero que, antes de entrar en los detalles del estudio, es clave entender los casos más frecuentes de bancarrota en Estados Unidos. Aunque nuestro modelo cuente, por razones obvias, con más casos de *default* que los que ocurrieron en 2023, es realmente importante entender lo que pasó el año pasado con Silicon Valley Bank, First Republic, Signature Bank, Citizens Bank y Heartland Tri-State, para poder hacer una buena interpretación del modelo y llegar a conclusiones sólidas. Cada uno tuvo motivos diferentes, más allá del típico "efecto contagio" y el caos que se desata en el mundo financiero cuando uno de ellos flaquea. Cada banco tiene su propia historia

detrás de su caída, y entender estas historias es muy importante para interpretar los resultados del modelo.

La aplicación de modelos de *machine learning* en esta tarea se ha vuelto muy importante, especialmente al comparar su efectividad con los métodos estadísticos tradicionales. Para procesar las más de 15 variables extraídas de los más de 50 bancos estadounidenses (activos y en *default*), utilizaremos la técnica del PCA (*principal component analysis*) encontrando una forma de reducir la complejidad del modelo sin perder información o lo que llamamos variabilidad. Posteriormente, este estudio se centrará en un enfoque de regresión *logit*. La elección de este modelo no es arbitraria; aporta unas ventajas particulares y consideramos que es un buen balance entre la simplicidad y efectividad del modelo (Eghian, 2021).

Es fundamental reconocer que partimos de la premisa de que cada uno de estos bancos cuenta con equipos de riesgo altamente capacitados: llevan a cabo un análisis exhaustivo de cada una de las métricas financieras utilizando modelos más precisos, herramientas más avanzadas, equipos con mayor conocimiento y mucho más presupuesto de lo que se cuenta en este trabajo. Asumimos que estas instituciones no solo están equipadas con mayores recursos, sino que también están sujetas a un seguimiento con lupa continuo por parte de reguladores, esta vez más estrictos tras la crisis del 2008 y el Comité de Basilea III (Le et al, 2020).

Después de Basilea III, los ratios financieros utilizados en este modelo están siendo supervisados y controlados mucho más de cerca. La lógica detrás de esto es que un nivel de capital más alto hace que los bancos sean más seguros, especialmente durante crisis financieras. Se establecieron una serie de medidas internacionalmente acordadas que requieren que los bancos mejoren la regulación, supervisión y gestión del riesgo, con implementación gradual desde 2014 hasta 2018 (Barrios & Blanco, 2003). Entre los ejemplos de mayor control cuantitativo y cualitativo del capital, se obliga a los bancos a mantener un 4.5% de *risk weighted assets*, que es más del doble de lo requerido en Basilea II (Vázquez & Federico, 2015).

Por esto, pensamos que el mayor control de las métricas financieras de los bancos, aparte de ayudar a que menos instituciones caigan en insolvencia, también va a hacer que los motivos por los que caigan en *default*, estén menos relacionados con los ratios ya controlados, y más por factores externos. No obstante, utilizaremos los resultados de nuestro modelo para (i) identificar los bancos que tienen una mayor probabilidad de caer en *default* y (ii) analizarlos individualmente para ver exactamente qué ratios están perjudicando el *output* del modelo.

## 2. Resumen de los *Defaults* del 2023

Estudios previos han identificado diversas causas de crisis bancarias, desde deficiente gestión de riesgos, regulación y supervisión ineficaces, altas tasas de *non-performing loans*, bajos ratios de *capital adequacy* y políticas monetarias. Se destaca especialmente el papel de la política monetaria, comúnmente empleada para controlar la inflación mediante ajustes de los tipos y el *money supply*. Los economistas argumentan que esta política ha sido una causa indirecta de crisis financieras pasadas (Taylor, J. B. 2018). Por ejemplo, el periodo de bajos tipos en Estados Unidos desde 2004 conllevó al auge de préstamos hipotecarios *subprime*, situación que culminó en el colapso de 2007 y 2008 (Christiano et al, 2004).

El año pasado, los economistas también apuntaron hacia la política monetaria como uno de los detonantes de la crisis bancaria en Estados Unidos en 2023. Los bancos centrales de todo el mundo decidieron subir los tipos de interés. Este movimiento desencadenó una crisis bancaria en Estados Unidos, resultando en el colapso de varios bancos regionales en el país y el desplome de un banco sistémico en Europa (Credit Suisse). El control de los tipos, diseñado para gestionar la economía, terminó desencadenando una serie de eventos que afectaron tanto a nivel local como global en el sistema bancario (Yan et al, 2014). Este trabajo aporta a la literatura de las crisis bancarias (por ej. Laeven & Valencia, 2020) un enfoque distinto, razonando que la política monetaria sí ha influido a la situación bancaria del 2023, pero de manera indirecta. Vamos a analizar brevemente cada caso particular.

### 2.1 Silicon Valley Bank (SVB) el 10 de Marzo

Durante el *peak* tecnológico de 2020/21, SVB, que financiaba las operaciones de start-ups tecnológicas y *ventures capital*, recibió mucho dinero de depósitos. Poco tiempo después, ocurrieron 2 factores fundamentales que le llevaron al *default*: (i) colocaron una gran parte de todo este dinero que habían recibido de depósitos en letras del tesoro a largo plazo con bajos tipos, cometiendo un error típico en banca que es recibir préstamos en el corto plazo e invertir a largo plazo y (ii) la decisión de la Fed de aumentar las tasas hizo que el valor de mercado de los bonos de SVB cayera, obligándolo a vender y materializar esas pérdidas. Esto, y el escaso número de depósitos que tenían asegurados (ver Figura 1) conllevó a retiros masivos de depósitos y falta de liquidez, llevando a su colapso. Por esto, entendemos que los altos tipos no afectaron directamente, ya que el problema estuvo en la decisión de colocar sus depósitos en bonos a largo plazo.

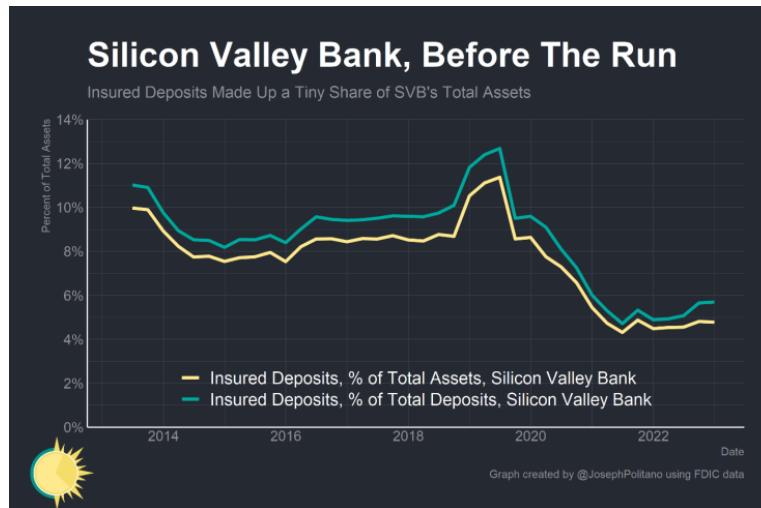


Figura 1. Depósitos Asegurados en SVB Antes de la Crisis (Politano, 2023)

## 2.2 Signature Bank el 12 de Marzo

Signature Bank, con una gran exposición a las criptos y también una proporción significativa de depósitos no asegurados, llegando a los \$79.5 Bn en diciembre de 2022, cayó en *default* no por una mala estrategia de inversión ni por la subida de tipos de la Fed, sino como consecuencia del pánico generado por la caída de SVB – los depositantes retiraron más de \$10 Bn de Signature, llevando a su cierre. La FDIC actuó para prevenir el “efecto contagio” y así una crisis bancaria sistémica. Los colapsos de SVB y Signature nos hacen plantearnos muchas dudas sobre los límites del seguro de depósitos; tal vez sea interesante para interpretar el modelo.

## 2.3 First Republic el 1 de Mayo

La caída de First Republic Bank se debió al aumento de las tasas de interés y sobre todo al colapso de los otros bancos ya mencionados. Esto provocó una fuga masiva de depósitos y que una enorme parte de los inversores vendieran sus acciones, perdiendo c. 85% de todo su valor (ver Figura 2). Además, la venta de bonos en pérdidas agravó su situación financiera.

En contraste con los casos anteriores, el colapso del First Republic fue manejado de manera más proactiva. En lugar de cerrar el banco y establecer un banco puente respaldado por el gobierno federal, lo cual habría requerido más fondos de la FDIC, los reguladores actuaron rápidamente, tomando el control del banco y facilitando su venta a JP Morgan durante el fin de semana.

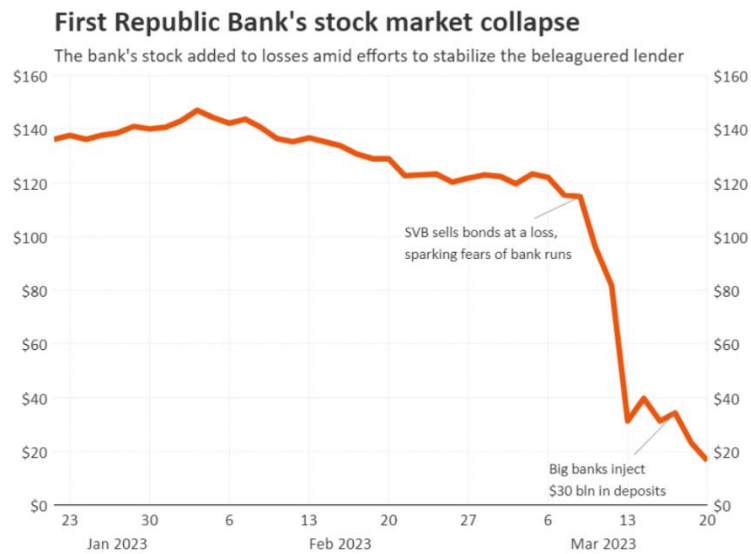


Figura 2. Evolución del Precio de Acción de FRB (Refinitiv Eikon, 2023)

## 2.4 Heartland Tri-State Bank el 28 de Julio

El cierre del banco comunitario Heartland Tri-State fue causado por una estafa, según el comisionado bancario de Kansas, David Herndon. Aunque los detalles exactos son escasos, parece que la actividad fraudulenta interna podría haber jugado un papel importante, con fraudes en préstamos siendo un culpable probable (Bailey, 2023). Es por esto, por lo que para este estudio vamos a excluir el caso de del banco de Kansas, considerándolo un *outlier* y asumiendo que el análisis de sus ratios de liquidez, solvencia y calidad de activos puede ser engañoso por (i) puede ser que los datos reportados no se correspondan con la realidad o (ii) que no haya ningún tipo de relación entre la probabilidad de *default* y estas variables, evitando que el modelo aprenda de este ejemplo.

Por dar un poco de contexto, Heartland Tri-State, que tenía activos totales de \$139.4 millones y depósitos totales de \$130.1 millones a 31 de marzo, fue adquirido por Dream First Bank, por lo que la FDIC estimó un coste de \$54.2 millones al Fondo de Seguro de Depósitos, o el 39% de los activos totales del banco (Graf, 2023).

## 2.5 Citizens Bank (of Sac City) el 3 de Noviembre

Citizens Bank, de Iowa, cerró sus puertas en noviembre debido a grandes pérdidas por préstamos que tenía concebidos en la industria de transporte. Las autoridades y reguladores, FDIC y IDOB, emitieron una orden en agosto que obligaba al banco a involucrar asesores externos que estudiaran su *portfolio* de transporte e hicieran un buen *risk assessment* del mismo. Poco después, la crisis en la industria del transporte, motivada principalmente por cambios en los hábitos de consumo, fue el detonante del colapso del banco. Iowa Trust & Savings Bank adquirieron sus activos el 3 de noviembre (Wack, 2023).



### 3. Recolección de Datos y Explicación del Dataset

En este estudio, la primera consulta de las entidades bancarias se ha obtenido de la base de datos de la *Federal Insurance Deposit Corporation* (FDIC), una agencia independiente del gobierno de los Estados Unidos que vela por la estabilidad del sistema financiero.

El conjunto de datos se compone de información sobre los siguientes dos tipos de entidades:

- Bancos actualmente activos: Aquellos que cumplen con los requisitos regulatorios, no presentan riesgos de solvencia y que actualmente desarrollan sus actividades diarias con normalidad
- Bancos en *default*: Aquellos que han incumplido con sus obligaciones financieras y han sido intervenidos por la FDIC

Es importante destacar que, la definición de *default* utilizada en este estudio incluye cuando (i) un banco ha recibido grandes ayudas financieras para evitar su *default*, es declarado insolvente por la FDIC y finalmente lo cierra para evitar más pérdidas o el efecto contagio entre bancos o (ii) es necesario liquidar los activos para pagar a los depositantes y se vende la compañía a otra entidad bancaria (Petropoulos et al, 2020).

Para el análisis, se han seleccionado los bancos más grandes de Estados Unidos en términos de tamaño de balance (volumen de activos). Posteriormente, se han eliminado los bancos que no son de origen estadounidense, como Toronto-Dominion Bank, UBS, BMO o la subsidiaria americana del Santander, apuntando un total de 37 bancos activos.

Adicionalmente, se incluyeron en el conjunto de datos los 4 bancos que cayeron en 2023, excluyendo como ya hemos mencionado, Heartland Tri-State, y los otros 7 bancos que cayeron entre 2018-2022, a excepción de Enloe State Bank, debido a falta de información en las bases de datos. Con esto, reunimos de primeras y antes de una limpieza, un total de 48 bancos estadounidenses (se puede consultar la lista completa en el Anexo 1).

#### 3.1 Variable Dependiente

Lo primero que hay que identificar de nuestro modelo es la variable dependiente o *response*, que es la que estamos intentando averiguar. En este caso, se trata de predecir la probabilidad de que un banco caiga en *default* en un trimestre determinado.

Para crear esta variable, hemos extraído, como ya hemos comentado antes, información de la FDIC sobre los bancos y los trimestres en los que han caído en *default* desde enero de 2018 hasta diciembre de 2023. Con esta información, vamos a marcar con un 1 aquellos trimestres en los que un banco específico cayó, y con un 0 aquellos trimestres en los que el banco continuó funcionando como siempre sin

problemas de insolvencia. De esta manera, estaremos utilizando como variable dependiente una variable binaria, también conocida como *dummy* o *factor*.

En nuestro conjunto de datos, asignaremos un valor de 1 en la columna *default* para los siguientes bancos y trimestres. Es importante aclarar que el trimestre que marcamos con un 1, es el último trimestre en el que el banco reportó métricas financieras. Para ver la fecha concreta en la que cayó en *default*, es mejor referirse al Anexo 1.

- Citizens Sac City en septiembre de 2023
- First Republic Bank en marzo de 2023
- Silicon Valley Bank en diciembre de 2022
- Signature Bank en diciembre de 2022
- Almena State Bank en septiembre de 2020
- First City Bank of Florida en septiembre de 2020
- The First State Bank en marzo de 2020
- Ericson State Bank en diciembre de 2019
- City National Bank of New Jersey en septiembre de 2019
- Rolute Bank en septiembre de 2019
- Louisa Community Bank en septiembre de 2019

Estos valores nos permitirán identificar y clasificar los trimestres en los que los bancos experimentaron problemas de insolvencia, de manera que, si un banco ha caído en noviembre, como es el caso de Citizens Sac City, apuntaremos el 1 en el último trimestre que reportó; septiembre.

Al incluir esta variable en nuestro modelo, podremos evaluar cómo las diferentes variables independientes de las que hablaremos ahora afectan la probabilidad de que un banco entre en *default* y ver cómo evoluciona esta probabilidad a lo largo del tiempo para posteriormente, hacer un análisis más profundo.

### **3.2 Ratios Financieros**

Los ratios financieros son uno de los puntos de partida más importantes cuando se intenta predecir la quiebra de un banco. Beaver (1966) demostró la utilidad de los ratios financieros para predecir la quiebra bancaria. Una de las conclusiones más importantes de su trabajo es cómo se puede medir la utilidad de estos datos a través de su capacidad de predicción. Aunque Beaver (1996) solo podía utilizar un único ratio a la vez, comentó que, un enfoque con muchos de ellos podría conducir a resultados más precisos, y eso es lo que vamos a realizar en este trabajo.

Una vez apuntados los 48 bancos con los que inicialmente se iba a realizar el estudio, teníamos que ver qué ratios había que seleccionar concretamente. Fijándonos en el trabajo de Le & Viviani (2018), en el que se utilizaron más de 25 ratios financieros diferentes para hacer las predicciones, observamos que muchos de ellos se encontraban en Factset y, en caso de que faltara alguno, se podría extraer de manera

manual de CapitalIQ. Por ello, exportamos todos los ratios que consideramos que podían ser útiles para el modelo, con frecuencia trimestral, desde marzo de 2016, clasificados en:

- *Profitability Ratios:*
  - i. *Return on Equity*
  - ii. *Return on Assets*
  
- *Liquidity Ratios:*
  - iii. *Loans / Deposits (%)*
  - iv. *Avg Loans / Deposits*
  
- *Asset Quality:*
  - i. *Loan Loss Reserves / Total Loans (x)*
  - ii. *Net Charge-Offs / Average Loans*
  - iii. *Loan Loss Reserves / Non-Performing Loans*
  - iv. *Loan Loss Reserves / Non-Performing Assets*
  - v. *Non-Performing Loans / Total Loans*
  - vi. *NPAs / Total Assets*
  - vii. *NPAs / Total Equity*
  - viii. *Texas Ratio*
  
- *Capital Adequacy:*
  - i. *Common Equity Tier 1 Capital Ratio (x)*
  - ii. *Tier 1 Risk Based Capital Ratio (%)*
  - iii. *Leverage Ratio*
  - iv. *Total Capital Adequacy Ratio (%)*
  - v. *Tangible Common Equity / Tangible Assets (%)*

### **3.3 Variables Externas**

Si bien los ratios financieros son herramientas muy útiles para predecir la quiebra bancaria, no son los únicos factores que influyen. El hecho de que 5 bancos hayan quebrado en 2023, 3 de ellos en el primer semestre, no es casualidad. Existen variables externas a los bancos que también pueden afectar su salud financiera.

Por este motivo, es importante incluir en el modelo las siguientes variables externas:

1. Tipos de interés: Los bajos tipos de interés, como explicó Mohan (2009), pueden ser un factor clave en las crisis financieras. Entre 2002 y 2004 en Estados Unidos, los tipos estaban muy bajos (también como respuesta de estimulación de la economía a la burbuja del *Dot Com*) haciendo que la gente buscara maneras de obtener más rendimiento de sus inversiones. Esto conllevó a desequilibrios importantes a nivel global, que acabaron resultando en la crisis del 2008 y la caída de una serie de bancos importantes

Por otro lado, como mencionó Sieroñ (2023), la subida de tipos suele ser una respuesta a una alta inflación. Actualmente los bancos centrales se enfrentan a un dilema: ¿deben seguir estimulando el crecimiento económico para la recuperación post-COVID, lo que podría disparar aún más la inflación, o deben subir las tasas para frenar la inflación con la probabilidad de recesión? Aunque subir las tasas es una medida clásica para combatir la inflación, la expectativa de estas subidas puede generar miedo de que los bancos no puedan pagar sus deudas, especialmente aquellos que tienen bonos a largo plazo, causando lo que se vio en 2023 con las retiradas masivas de depositantes (Calvo & Végh, 1995).

En resumen, tanto los bajos como los altos tipos de interés pueden tener un gran impacto en la probabilidad de que un banco caiga en *default*. Por esto, es necesario incluir esta variable en el modelo de predicción.

2. *Quantitative Easing*: Utilizando el estudio de Merton (1974), que evaluó la probabilidad de *default*, se encontró que las políticas de QE de la Reserva Federal de Estados Unidos tuvieron un gran impacto en la posterior emisión de bonos (Duca et al, 2016). Este modelo de Merton, que sigue siendo uno de los más simples y útiles incluso después de cuarenta años, es la base de algunos de los modelos de riesgo crediticio que se utilizan hoy en día (Gheno, 2007).

Durante un fuerte QE, en el que la Fed expande su balance en grandes cantidades y de manera muy rápida, se suele ejercer una presión extrema sobre la emisión de bonos y se reducen sus valores de mercado (Moro Visconti, 2016). Esta expansión masiva del balance de la Fed y sus efectos en el mercado de bonos pueden tener repercusiones significativas en la capacidad de los bancos para gestionar sus riesgos y solvencias y por tanto su probabilidad de *default*. Es por esto por lo que también consideramos importante incluir esta variable en el modelo

### 3.4 Limpieza Manual del *Dataset*

Una vez tenemos los 48 bancos estadounidenses, con las 19 variables (17 ratios financieros + 2 variables externas) asociadas a cada uno de ellos y, a la vez, a cada uno de los trimestres desde marzo de 2016 (horizonte temporal de prácticamente 8 años), es importante hacer un primer análisis por encima antes de utilizarlo para el modelo.

En primer lugar, al hacer el *pull* de datos de Factset, de las variables que no se encontraban había algunas que se exportaban como N/A y otras como 0, por esto, transformamos todos los campos con valor 0 en N/A.

Después, decidimos tratar la gran mayoría de N/A de manera manual, para cuando le demos el archivo a R, y haya que reemplazar estos campos, no se haga de manera “bruta”, perjudicando el tamaño y la variabilidad del *dataset*. Corregimos y limpiamos tanto las variables como las observaciones (bancos):

- De las 19 variables, hemos decidido eliminar 2 ratios financieros porque contaban con un gran número de campos vacíos, estos son el *Avg Loans / Deposits* y el *Leverage Ratio*. Con esto conseguimos que, al hacer la limpieza en R, no se elimine toda la información que corresponde a esos bancos y fechas, sino simplemente omitimos la variable. Además, podemos ver que eran variables que realmente no estaban proporcionando información al *dataset*, porque eran perfectamente sustituibles por el *Loans / Deposits (%)* y el *Tangible Common Equity / Tangible Assets (%)* respectivamente (ver Figuras 3 y 4 para observar la correlación de las variables eliminadas con otros ratios financieros que sí permanecen en el conjunto de datos).

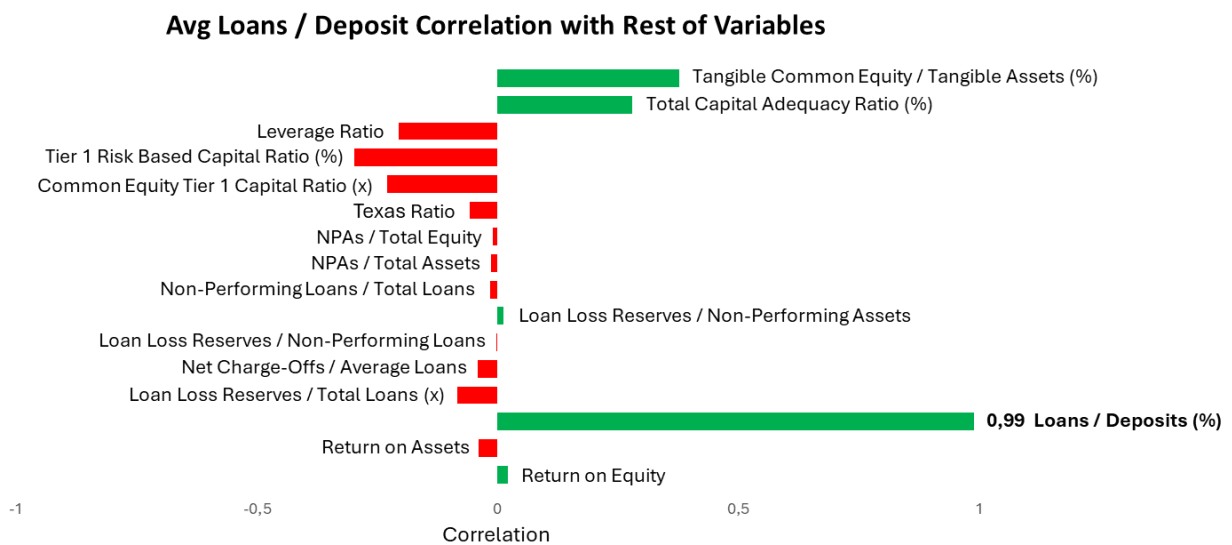


Figura 3. Correlación Entre *Avg Loans / Deposits* y el Resto de Ratios Financieros

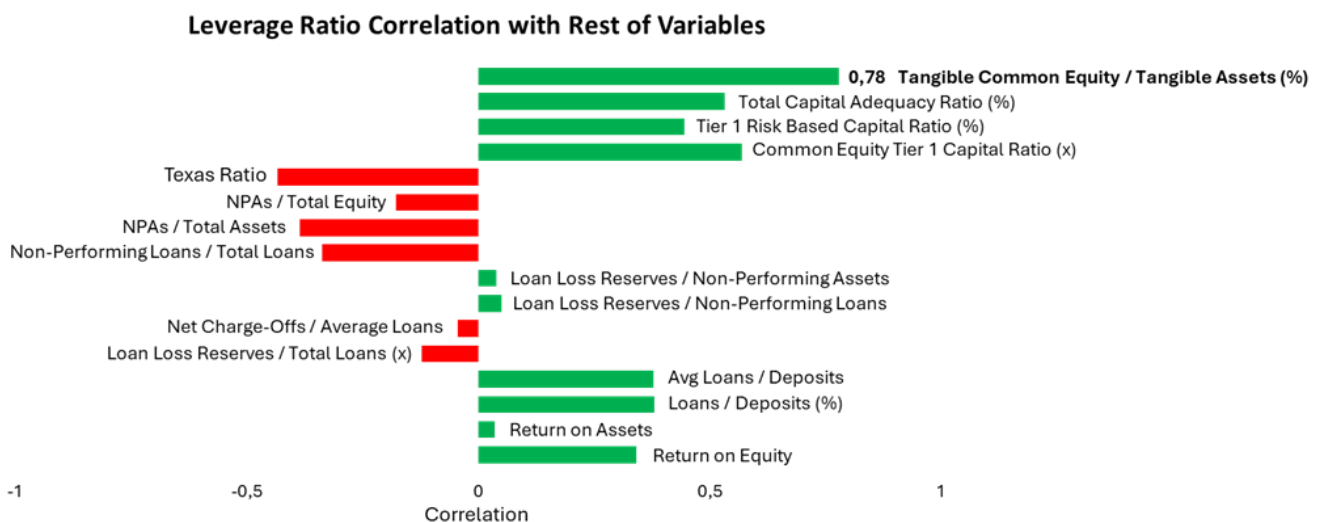


Figura 4. Correlación Entre *Leverage Ratio* y el Resto de Ratios Financieros

Una vez hemos eliminado estos dos ratios por (i) falta de información y (ii) no aportaban variabilidad al conjunto de datos (eran fácilmente sustituibles), nos quedamos con un total de 15 ratios financieros. Siempre es interesante poder ver la matriz de correlaciones del *dataset*, que nos ayudará a identificar asociaciones lineales entre pares de variables, detectar multicolinealidad, seleccionar características relevantes para el modelo y visualizar la estructura de los datos a un nivel general para posteriormente poder llegar a conclusiones (ver Figura 5).

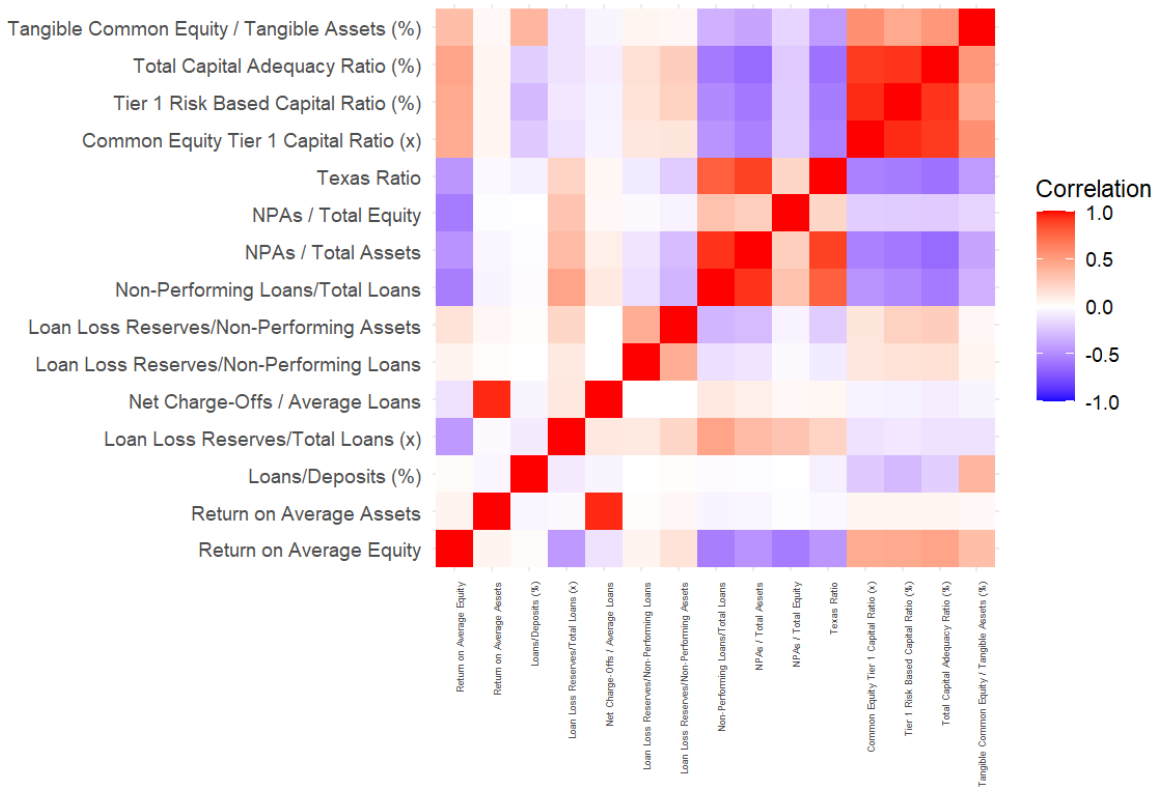


Figura 5. Matriz de Correlaciones

- En cuanto a las observaciones (bancos), tanto State Street Corporation, Comerica Bank y East West Bank cuentan también con una gran cantidad de valores desconocidos, ya sea porque no están obligados a reportar las mismas métricas, o porque Factset no recoge esta información. Además de eliminar los bancos ya mencionados, también hemos considerado importante omitir del *dataset* tanto Goldman Sachs, como Morgan Stanley, al ser únicamente bancos de inversión, y poder desvirtuar el resto de ratios financieros. En total, se han eliminado 5 bancos dejando el *dataset* con 43 instituciones estadounidenses.

## 4. El Modelo

Hay precedentes en el pasado, de estudios que han intentado predecir la probabilidad de *default* de los bancos, y lo más frecuente es el uso de un modelo *logit* u otros modelos de *machine learning*. Poon et al (1999), Ohlson (1980), Daily & Dalton (1994), Lee & Yeh (2004), Brédart (2014) y du Jardin (2017), optaron por utilizar un modelo de regresión *logit*. Por otro lado, Liang et al (2016), Lin et al (2011), Hosaka (2019) y Boyacioglu et al (2009), utilizaron otros modelos como *neural networks*, *k-nearest neighbours* o *support vector machines*. En nuestro caso se utilizará la *logistic regression*. Esta decisión se basa en varios factores importantes.

- En primer lugar, la *logistic regression* es una técnica simple y efectiva. Esto permite evitar tener que aprender un modelo completamente nuevo desde cero, lo que habría requerido una curva de aprendizaje significativa y más tiempo de desarrollo.
- Además, es una de las opciones que ofrece mejores resultados en la predicción de insolvencia de las instituciones financieras, como lo demuestran los estudios de du Jardin (2016) y uno de los primeros trabajos de Ohlson (1980). Estos trabajos destacan que el modelo *logit* presenta ventajas sobre otros enfoques, como que no solo te da información del tamaño del coeficiente o predictor, sino que también te relaciona el sentido de la relación (positivo o negativo) y es un modelo que, por norma general, no suele caer en *overfitting*.
- Por último, el modelo *logit* se ajusta a la perfección al objetivo del estudio; que, a través de un conjunto de variables numéricas, se pueda predecir la variable dependiente que es binaria o dicotómica. Se puede observar en el Anexo 2 una tabla de antiguos estudios que se han hecho sobre la probabilidad de *default* de los bancos, destacando cosas interesantes como el modelo utilizado, el tipo de variables involucradas, el rango de tiempo, y el país.

Es importante aclarar que dentro de esta sección no se va a encontrar únicamente el modelo *logit* desarrollado, sino todos los pasos previos desde que se importa el *dataset* en R, para ir viendo las decisiones que se van tomando

### 4.1 Preparación Previa

Lo primero que hay que hacer antes de empezar a desarrollar el modelo es, siguiendo el ejemplo de Canbas et al (2005), un PCA o *principal component analysis*; técnica estadística utilizada para reducir la dimensionalidad de conjuntos de datos de gran tamaño, transformando las variables originales en un nuevo conjunto de variables no correlacionadas. Para poder realizar este PCA, hay una serie de pasos que se deben de hacer previamente para dejar el *dataset* preparado. En primer lugar, el PCA solo funciona con variables numéricas, y siempre que se exportan los datos de Excel, es importante reclasificar el tipo de variables que estamos utilizando. Por esto, se va a crear un nuevo *subset* de datos únicamente con las variables que se desea introducir

en el PCA. Es decir, todas las variables menos los bancos, las fechas, y la probabilidad de *default*. Una vez que tenemos este *subset* de datos, convertimos todas estas variables a tipo numérico. Se podría también incluir los bancos y las fechas en el PCA haciendo *one-hot encoding*, que convierte las variables categóricas en tipo binario, pero no consideramos que aporte información significativa.

En segundo lugar, es importante que gestionar los N/As restantes. Después de hacer la limpieza manual, nos encontramos con un número muy reducido de N/As que no tienen un patrón o una relación clara; son valores desconocidos que, por un motivo u otro, ocupan posiciones en el *dataset* para un banco y fecha específicos, pero ya no son columnas enteras de N/As. La manera en la se van a manejar los valores desconocidos es a través de un algoritmo de Kalman. Este algoritmo utiliza un filtro que proporciona estimaciones de los N/As dados los valores de esa misma variable a lo largo del tiempo (Govaers 2019). Es decir, utiliza la información de las observaciones anteriores para estimar el valor que podría tener esa celda. Los filtros de Kalman son relativamente sencillos de usar, y tanto la capacidad de aprender del pasado, como su flexibilidad en series temporales los hace muy precisos. Se ha utilizado la librería “imputeTS” para poder utilizar la función “na\_kalman” que reemplazará los valores desconocidos.

Por último, como la variable *default* estaba codificada con 1 y 0, la reconocía como variable numérica, y la hemos convertido en *factor* o binaria, de manera que se han cambiado los 1 por *TRUE* y los 0 por *FALSE*. Después de esto, se ha indicado que esta va a ser la variable dependiente y la que vamos a querer predecir posteriormente en el modelo.

## 4.2 PCA

Canbas et al (2005) tenían un objetivo muy parecido al de este estudio; elaborar un *early warning indicator* para el *default* de los bancos, y decidieron que la mejor manera era realizar un PCA. Debido a los buenos resultados y a la similitud entre objetivos, en este trabajo también se partirá de esta base. La elaboración de *principal components* en este estudio es necesaria porque se encuentran una gran cantidad de variables muy relacionadas entre sí, que dan una mayor complejidad al *dataset* sin proporcionar más información. El PCA facilita el manejo de datos complejos y reduce el riesgo de *overfitting* (Howley et al, 2006). En este caso, como estamos trabajando con un conjunto de muchas variables y, entre ellas, muchos ratios financieros con alta correlación entre ellos, el PCA es una de las partes más importantes para que el modelo funcione correctamente. Se espera que el PCA sea capaz de simplificar el tamaño del *dataset* lo máximo posible, pero perdiendo la mínima variabilidad o información.

Realizar el PCA a través de R es muy sencillo, lo único que hay que hacer antes es escalar el *dataset*; convertir todos los valores en números entre 0 y 1 para que ninguna variable predomine sobre otra por tener un rango mayor. Una vez que se ha realizado el PCA, lo más importante es analizar bien todos los *principal components* (nuevas variables) que han salido para elegir el número óptimo. Al seleccionar el número de



componentes principales, es crucial capturar los aspectos más relevantes de los datos, mientras se ignoran aquellos de menor importancia. En la práctica, existen diversas formas para determinar el número óptimo de componentes, ya que observar una tabla con números no suele ser muy eficaz. Por ejemplo, Love & Tucker (1970) aplicaron el *scree-test*, que ayuda a identificar los componentes principales más significativos. Por esto, en este estudio se ha decidido también hacer un *Scree Plot* (ver Figura 5) y realizar la “regla del codo”, en la que hay que mirar la línea de la gráfica e identificar en qué momento hay un cambio significativo en la pendiente, pasando de un descenso (más variabilidad explicada) a una tendencia más horizontal, en la que el efecto marginal de un *principal component* al siguiente es muy pequeño, haciendo el *dataset* más complejo a cambio de, únicamente, un pequeño porcentaje de información explicada (Cattell, 1966).

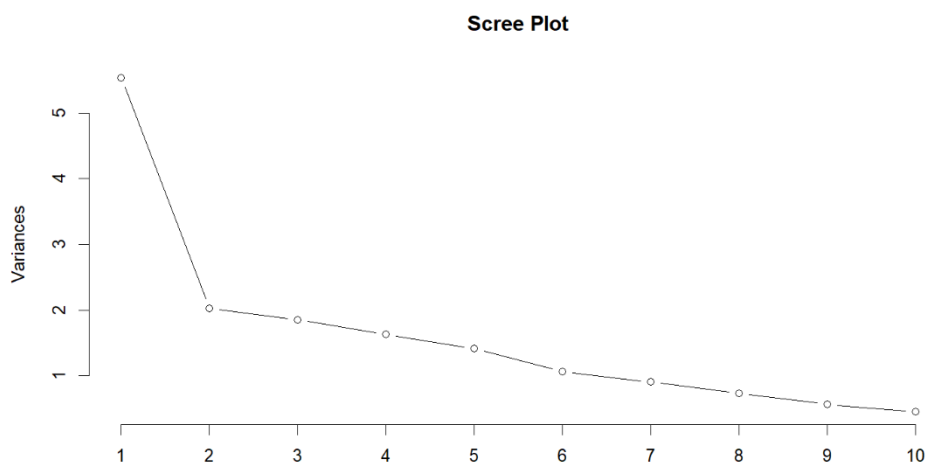


Figura 5. Scree Plot del PCA

En este gráfico podemos observar que el número óptimo de *principal components* para nuestro modelo es 6. Pero si queremos confirmar nuestra decisión o no tenemos claro dónde está exactamente el “codo”, se puede también plotear otro gráfico (ver Figura 6) que explique la variabilidad total explicada según vamos aumentando el número de PCs que elegimos.

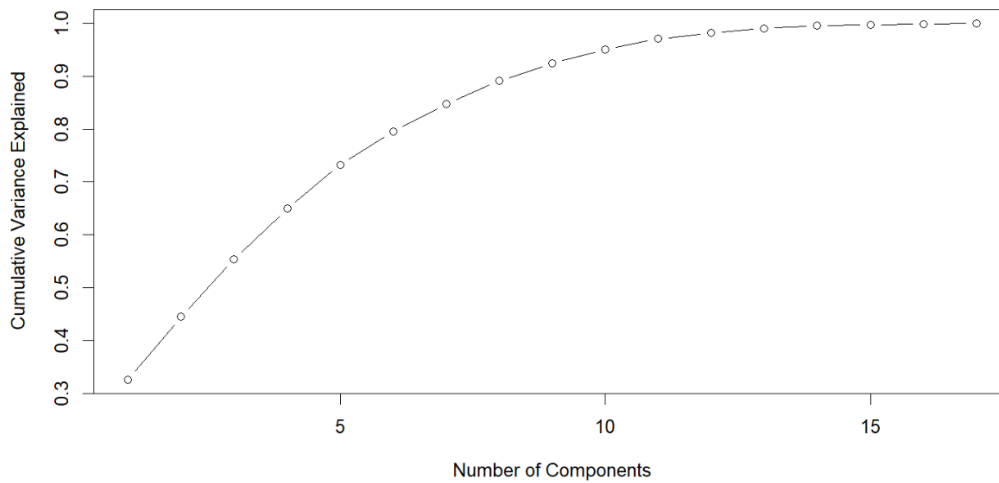


Figura 6. Varianza Acumulada Explicada por los PCs

Al finalizar el PCA, se ha conseguido reducir el número de variables de 17 a 6 y, con ello, la complejidad del conjunto de datos. La ventaja del PCA, es que ha permitido reducir el dataset hasta en un 75%, únicamente sacrificando un 20% del total de la variabilidad. El motivo por el que hemos podido hacer esto es porque las variables anteriores contaban con una alta correlación entre ellas, y a través de este algoritmo hemos podido omitir todos aquellos valores que no aportan información adicional.

Una vez tenemos los 6 *principal components*, los juntamos de vuelta con las variables que habíamos quitado al principio, pero que son necesarias para el modelo, es decir, los bancos, las fechas (trimestres) y la variable binaria *default*.

### 4.3 Training y Test Sets

Uno de los problemas más frecuentes en los modelos es el *overfitting*, que ocurre cuando el modelo se ajusta demasiado a los detalles específicos de los datos de entrenamiento y no es capaz de generalizar bien cuando le presentas nuevos datos; tiene poca capacidad predictiva. Para abordar este problema, se divide el conjunto de datos en dos partes: un conjunto de entrenamiento llamado *training set* o *in-sample* y un conjunto de prueba llamado *test set* o *out-of-sample*.

El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se reserva para evaluar el rendimiento del mismo con datos completamente nuevos. Esto permite estimar cómo se comportará el modelo en situaciones del mundo real, lo que ayuda a detectar si el modelo tiene o no *overfitting* (Eghian, 2021).

Al construir las muestras de *training* y *test sets*, es muy importante tener un conjunto de entrenamiento grande para proporcionar una predicción precisa (Alaka et al, 2018). A pesar de que no tenemos acceso a un gran número de datos para este trabajo, y no vamos a poder contar con un *training set* potente, intentaremos realizar la partición de

la mejor manera posible. Un buen ejemplo fueron las muestras elegidas y seleccionadas por López Iturriaga & Sanz (2015) o las del trabajo de Serrano-Cinca & Gutiérrez-Nieto (2013), en las que dividían el *dataset* en los primeros años como parte del *in-sample*, y únicamente cogían la información de los bancos del último año/últimos dos años para probar la efectividad del modelo, como parte del *out-of-sample*.

En nuestro caso, se va a utilizar como *training set* una partición aleatoria a manos del modelo, que va a representar un 70% del total de la información del conjunto de datos. Por otro lado, para evaluar el modelo, utilizaremos un *training set* que contenga la información no incluida en el aprendizaje, es decir, el 30% restante.

#### 4.4 Explicación del Modelo

En este estudio, dado que la variable dependiente, *default*, es dicotómica, utilizamos un modelo de regresión logística binaria, como es también el caso de muchos estudios relacionados con predecir la probabilidad de insolvencia de un banco (Ohlson, 1980; Mensah, 1984; Zavgren, 1985; Aziz et al, 1988; Burgstahle et al, 1989; Flagg et al, 1991; Platt & Platt, 1991; Bell et al, 1998; Premachandra et al, 2009). Este modelo *logit* describe la relación entre una variable dependiente dicotómica que puede tomar el valor 1 (banco caído) o el valor 0 (saludable), y  $k$  otras variables explicativas  $x_1, x_2, \dots, x_k$ . Estas variables pueden ser cuantitativas o cualitativas, pero en nuestro caso son todas cuantitativas. El modelo de regresión *logit* que vamos a desarrollar exige que la variable endógena sea una combinación lineal de las variables exógenas:  $y_i^* = \beta X_i + \varepsilon_i$  donde  $\varepsilon$  es el error y  $\beta$  es el vector de coeficientes (Brédart, 2014).

Una vez entendido, se aplica el modelo de regresión logística binomial utilizando todas las variables en el conjunto de datos de entrenamiento como predictores, con la variable *default* como la variable de respuesta.

## 5. Interpretación del Modelo

Después de haber desarrollado el modelo con el *training set*, el siguiente paso es evaluar su performance con un conjunto de datos que no haya visto antes. Utilizamos el *test set* para ver las predicciones que hace, y asumimos un *threshold* del 10% (umbral a partir del cual vamos a clasificar a un banco como que ha caído en *default* dándole un valor de 1 o *TRUE*).

### 5.1 Resumen Estadístico

Lo primero que queremos observar es un resumen estadístico del modelo, que se puede ver en la Figura 7.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0558  -0.0443  -0.0173  -0.0079   3.6643

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.43011    1.89785  -4.442 8.92e-06 ***
PC1          0.56413    0.21571   2.615 0.00892 **
PC2          1.08310    1.34791   0.804 0.42166
PC3          0.02464    0.79944   0.031 0.97541
PC4          1.92261    0.80420   2.391 0.01682 *
PC5          0.65956    0.55471   1.189 0.23444
PC6         -0.83042    0.62962  -1.319 0.18720
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.333  on 777  degrees of freedom
Residual deviance: 33.811  on 771  degrees of freedom
AIC: 47.811

Number of Fisher Scoring iterations: 10
```

Figura 7. Resumen del Modelo

Aunque este resumen no es suficiente para poder evaluar el rendimiento del modelo, sí se pueden observar algunas peculiaridades:

- Los valores de los *Deviance Residuals* son bastante pequeños en comparación con el rango general de datos, lo que sugiere que el modelo está bien ajustado. Cuanto menor sea el valor, mejor capacidad tiene el modelo de predecir la variable *response* (Davidson & Gigli, 1989).

- Los coeficientes estimados para PC1 y PC4 son significativos a un nivel de confianza del 5%, lo que indica que estas variables tienen un impacto significativo en la predicción de la probabilidad de *default*.
- La *Null Deviance* mide cuánto de lejos están las probabilidades predichas por un modelo hipotético sin *predictors* de los valores reales. Es decir, qué tan bien puede predecirse la variable *response* mediante un modelo que solo contiene el término *Intercept* (Harris, 2021). En este caso, el valor cambia de 70.3 con el modelo nulo, a 33.8 con el modelo normal (más de la mitad), lo que significa que los *predictors* son capaces de explicar una cantidad significativa de la variabilidad en los datos.
- Por último, el AIC o *Akaike Information Criterion* es un resultado que se utiliza mucho para comparar modelos distintos y seleccionar el óptimo. El modelo con el menor AIC es el que se ajusta mejor. Se fija en qué modelo es capaz de explicar una mayor varianza con el menor número de parámetros posibles; hace un *trade-off* entre estas dos variables. Esto está muy relacionado con seleccionar el número óptimo de *principal components* en el PCA y, por esto, en vez de comparar el AIC con otros modelos, en este caso vamos a compararlo con el mismo modelo, pero utilizando más o menos parámetros. Para confirmar que se cogió el número óptimo, el AIC con 6 componentes principales debe de ser el que tenga un menor valor, y así es (observar Figura 8).

# Components	AIC
8 PCs	50.87
7 PCs	49.08
6 PCs	47.81
5 PCs	50.88
4 PCs	52.05
3 PCs	58.78
2 PCs	58.23
1 PCs	58.37

Figura 8. Valor AIC en Función del Número de PCs

## 5.2 Curva ROC y AUC

La curva ROC muestra cómo la sensibilidad (tasa de verdaderos positivos) cambia con la especificidad (tasa de verdaderos negativos) para diferentes valores en el test o análisis de un modelo *logit*. Un área bajo la curva (AUC) de 1 indica una prueba perfecta, mientras que un AUC de 0.5 sugiere que la prueba no es mejor que el azar. Se elige un punto de corte óptimo en la curva considerando los objetivos asociados con los falsos positivos y falsos negativos (Hoo et al, 2017). En este caso, tenemos un AUC de 0.944, que significa que la habilidad predictiva del modelo es bastante buena, tanto en términos de sensibilidad como de especificidad. El conjunto de datos que estamos utilizando no es suficientemente grande como para que se dibuje una curva perfecta ROC, pero se puede observar el buen resultado en la Figura 9.

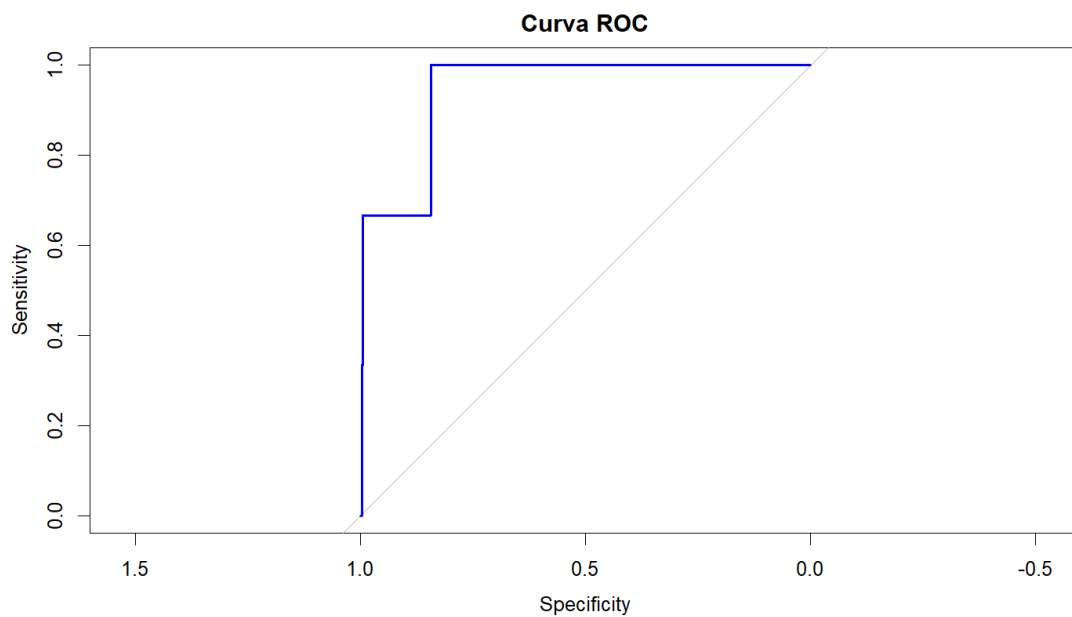


Figura 9. Curva ROC y AUC

## 5.3 Matriz de Confusión

Las matrices de confusión son una herramienta fundamental para evaluar errores en modelos *logit*. Estas matrices codifican las clasificaciones de errores, mostrando el número de elementos clasificados erróneamente para cada par (clase original en la que los elementos deberían ser clasificados, clase incorrecta en la que los elementos son erróneamente clasificados). Para problemas de clasificación binaria, como el nuestro, la configuración de parámetros depende de la tolerancia a los errores de tipo I y II. Las matrices de confusión son realmente una versión simplificada, más fáciles de interpretar gráficamente que las curvas ROC, pero las conclusiones o la información utilizada es exactamente la misma (Beauxis-Aussalet & Hardman, 2014).

El *test set*, que es el que utilizamos para evaluar el *performance* del modelo, cuenta con un total de 331 bancos no han caído en *default* en un trimestre concreto, y 3 bancos que sí han caído en *default* en trimestres concretos. Los errores tipo I (falsos positivos) que encontramos, son 4, que corresponden a un 1.2% del total. Por otro lado, los errores tipo II (falsos negativos) son únicamente 1, pero que corresponden al 33.3% del total (ver Figura 10). Todo esto, como se ha mencionado antes, asumiendo el *threshold* del 10%.

Con estos datos, podemos decir que la sensibilidad (capacidad del modelo de detectar correctamente los casos positivos) es decente, o de un 66%, aunque contando con únicamente 3 casos en el *test set*, es difícil llegar a conclusiones. Por el otro lado, la especificidad del modelo (capacidad de detectar correctamente los negativos), es bastante alta, o de un 98.8%.

		Predicted	
		FALSE	TRUE
Actual	FALSE	327	4
	TRUE	1	2

Figura 10. Matriz de Confusión del *Test Set*

#### 5.4 Variables que Más Afectan al Modelo

Es de especial importancia analizar cuáles son las variables que más están influyendo en la probabilidad de *default* y, por tanto, las que hay que mirar más de cerca.

Tras la aplicación del PCA, aunque se pierde la interpretabilidad directa de las variables originales, aún podemos estudiar cómo estas influyen en la variable dependiente o *response*. La estrategia es la misma que la que aplican Canbas et al (2005); consiste en calcular los coeficientes de correlación entre los componentes principales y las variables originales (ratios financieros + variables externas), permitiéndonos identificar la magnitud y dirección de su relación. Además, con el análisis de carga de los componentes, se obtiene información sobre cómo cada variable contribuye a la estructura de los componentes principales. De esta manera, podemos determinar qué variables independientes originales están más relacionadas con los componentes que explican la variabilidad en la variable dependiente, ayudándonos a mantener un nivel de comprensión e interpretación sobre la influencia de las variables incluso después de la transformación realizada por el PCA. El resultado es el que se observa en la Figura 11.

Para poder interpretar cómo afectan las variables a la probabilidad de *default* de un banco, es necesario saber interpretar los coeficientes:

- Dirección de la relación: El signo del coeficiente (positivo o negativo) indica la dirección de la relación entre la variable independiente correspondiente y la variable dependiente. Un coeficiente positivo sugiere una relación positiva, lo que significa que a medida que el valor de la variable independiente aumenta, la probabilidad de que ocurra el evento de interés también tiende a aumentar. Por el contrario, un coeficiente negativo indica una relación negativa, lo que significa que a medida que el valor de la variable independiente aumenta, la probabilidad del evento de interés tiende a disminuir.
- Magnitud de la relación: La magnitud del coeficiente indica la fuerza de la relación entre la variable independiente y la variable dependiente. Cuanto mayor sea el valor absoluto del coeficiente, más fuerte será la relación entre las variables. Un valor absoluto cercano a 0 sugiere una relación débil, mientras que un valor absoluto cercano a 1 sugiere una relación más fuerte.

Variable Name	Coeficiente
Return on Average Equity	-0,23743545
<b>Return on Average Assets</b>	<b>-0,90873904</b>
Loans/Deposits (%)	0,24940188
Loan Loss Reserves/Total Loans (x)	-0,18495731
<b>Net Charge-Offs / Average Loans</b>	<b>0,90345954</b>
Loan Loss Reserves/Non-Performing Loans	0,03275402
Loan Loss Reserves/Non-Performing Assets	0,48639552
Non-Performing Loans/Total Loans	-0,43926555
NPAs / Total Assets	-0,39948476
NPAs / Total Equity	0,36473185
Texas Ratio	-0,29046507
<b>Common Equity Tier 1 Capital Ratio (x)</b>	<b>-0,84896996</b>
Tier 1 Risk Based Capital Ratio (%)	-0,61144071
Total Capital Adequacy Ratio (%)	-0,63467577
<b>Tangible Common Equity / Tangible Assets (%)</b>	<b>-0,87277573</b>
<b>Interest Rate</b>	<b>0,85874397</b>
<b>QE (trn)</b>	<b>0,84496744</b>

Figura 11. Resumen de Coeficientes con la *Response Variable*

De las 17 variables que tiene nuestro modelo, hay 6 de ellas que influyen especialmente en la probabilidad de *default*:



- *Return on Average Assets* (-0.91): Un retorno más alto sobre los activos indica una mayor eficiencia en la generación de ingresos a partir de los activos del banco. Esto refleja claramente una gestión financiera sólida y una operación rentable, lo que reduce el riesgo de caer en bancarrota.
- *Net Charge-Offs / Average Loans* (0.90): Un ratio más alto de pérdidas netas sobre préstamos indica un mayor riesgo crediticio y una potencial disminución en la calidad de los activos del banco (*asset quality*). Esto puede ser indicativo de problemas con los préstamos y una mayor probabilidad de *default*.
- *Common Equity Tier 1 Capital Ratio* (-0.85): Un CET1 ratio más alto indica una mayor capacidad del banco para absorber pérdidas y resistir periodos de dificultades financieras. Esto refleja una estructura de capital más sólida y una menor probabilidad de insolvencia.
- *Tangible Common Equity / Tangible Assets* (-0.87): Bastante parecido a lo anterior. Una mayor proporción de capital común tangible respecto a activos tangibles indica una mayor solidez financiera y una menor exposición a riesgos. Esto proporciona una base más estable para el banco y menor probabilidad de caída.
- *Interest Rate* (0.85): Aquí está uno de los grandes debates del estudio. A pesar de que los altos tipos de interés les proporcionan mayores beneficios en la cuenta de resultados (véase las subidas en el precio de acción de instituciones financieras en el 2023), para los bancos que tienen una gran exposición a pasivos sensibles a las tasas de interés, la subida de tipos es muy perjudicial. El modelo ha identificado que la subida de tipos afecta enormemente a la probabilidad de *default* de los bancos, aumentándola.
- *Quantitative Easing* (0.84): Aquí también ocurre algo parecido a con los tipos de interés. Si bien un aumento del balance de la Fed puede proporcionar liquidez adicional al sistema financiero, también puede aumentar la competencia y reducir los márgenes de beneficio. Esto puede afectar negativamente la rentabilidad y aumentar el riesgo de incumplimiento para algunos bancos.

## 6. Análisis de Bancos en Riesgo

El primer análisis que vamos a hacer es observar la misma matriz de confusión que hicimos con el *test set*, pero con el conjunto de datos original, para ver el total de valores que es capaz de predecir el modelo y estudiar los errores (ver Figura 12).

		Predicted	
		FALSE	TRUE
Actual	FALSE	1092	11
	TRUE	2	7

Figura 12. Matriz de Confusión del Dataset General

Como se observa en este caso, la sensibilidad del modelo es de un 77.8% y la especificidad es de un 99%. Tiene sentido que ambos valores sean superiores a los del *test set*, porque en este caso el modelo sí está utilizando la misma información con la que ha sido entrenado. Sin embargo, las diferencias no son muy grandes, lo que indica que nuestro modelo no tiene *overfitting*.

### 6.1 Falsos Negativos (Error Tipo II)

Empezamos a analizar los errores por la parte más sencilla, que son aquellos bancos que sí han llegado a bancarrota. Aunque inicialmente introducimos en el *dataset* un total de 11 bancos que han caído en Estados Unidos desde enero de 2018, el modelo ha omitido dos de ellos por falta de datos, SVB y First Republic. De los 9 bancos restantes, se predicen con acierto 7. Con un porcentaje de éxito de +75%, los únicos bancos que no es capaz de reconocer el modelo son The First State Bank, que cae en abril del 2020 y Citizens Bank, que cae en noviembre del 2023. Se puede observar en la Figura 13 los 9 bancos que han caído con la evolución de la probabilidad de cada uno a lo largo del tiempo, marcando con un círculo los 7 que el modelo ha acertado como *TRUE* bajo el *threshold* indicado del 10%.

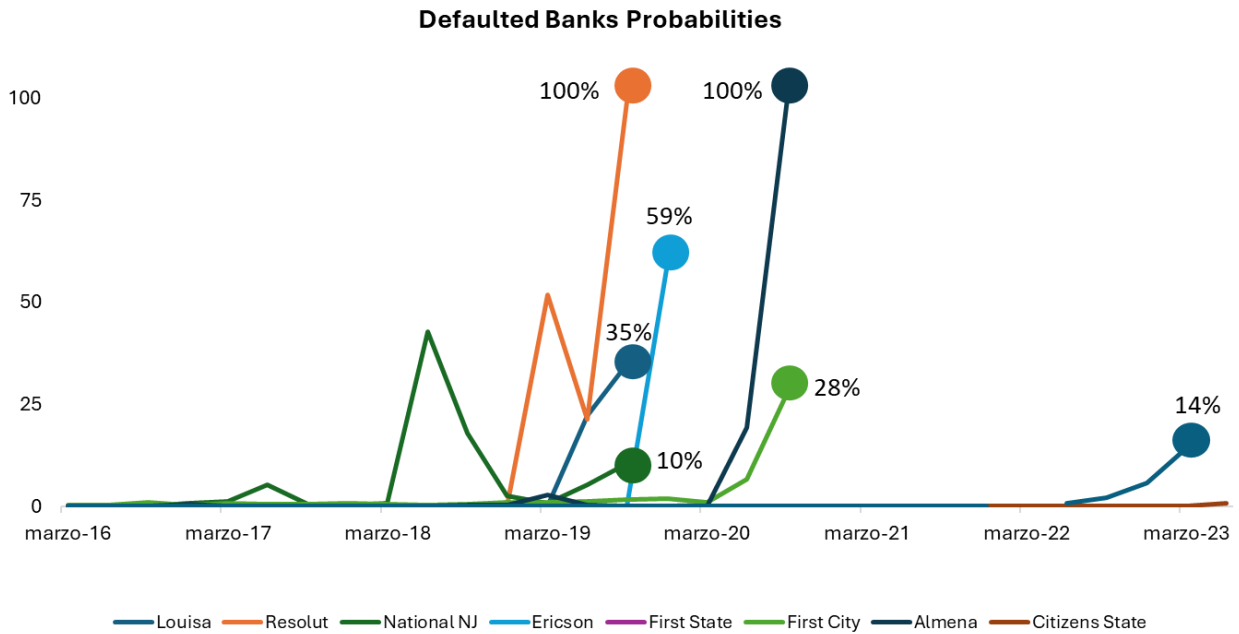


Figura 13. Probabilidades de *Default* de los Bancos que Sí Han Caído

## 6.2 Falsos Positivos (Error Tipo I)

Lo que más nos preocupa es entender aquellos bancos que, para un trimestre en concreto, sí han sido clasificados como *TRUE* o alta probabilidad de *default*, pero que actualmente están funcionando de manera normal. Los 11 valores que han tenido el error tipo I son los siguientes:

- US Bancorp – junio 23 – 15.84%
- US Bancorp – marzo 23 – 20.47%
- Columbia Bank – junio 23 – 23.02%
- Columbia Bank – marzo 23 – 23.07%
- Cullen Bank – diciembre 20 – 86.19%
- Louisa Community Bank – junio 19 – 22.15%
- Resolute Bank – junio 19 – 21.34%
- Resolute Bank – marzo 19 – 51.78%
- City National Bank of New Jersey – septiembre 18 – 18.05%
- City National Bank of New Jersey – junio 18 – 42.73%
- Almena State Bank – junio 20 – 19.33%

De todos estos falsos positivos, podemos omitir los bancos que cayeron después de esas fechas, interpretando los resultados no desde el punto de vista de un fallo de predicción del modelo, sino como un acierto de lo que iba a acontecer en un futuro.

Si quitamos del análisis a Louisa Community Bank, Resolute Bank, City National Bank of New Jersey y Almena State Bank, nos quedan únicamente 3 bancos que o (i) han estado cerca de caer en el pasado o (ii) están asumiendo riesgos excesivos y se encuentran en una situación delicada. Aunque sería interesante entender por qué el modelo le dio a Cullen Bank una probabilidad tan alta de caer en diciembre de 2020, por motivos prácticos vamos a centrarnos únicamente en los bancos que pueden estar cerca del colapso a día de hoy; US Bancorp y Columbia Bank.

### 6.2.1 US Bancorp

De las 6 variables que afectan especialmente a la probabilidad de *default*, hay dos que no cambian entre bancos para el mismo trimestre, porque son variables externas; los tipos de interés y el *quantitative easing*. Por esto, aunque probablemente sí hayan influido en gran medida a la probabilidad de caída de US Bancorp (en marzo y junio de 2023 ambas variables se encontraban en picos), es interesante ver las diferencias de los otros 4 ratios financieros entre US Bancorp, y el resto de los bancos solventes (ver Figura 14).

Comparison Between Active Banks and US Bancorp in June 2023

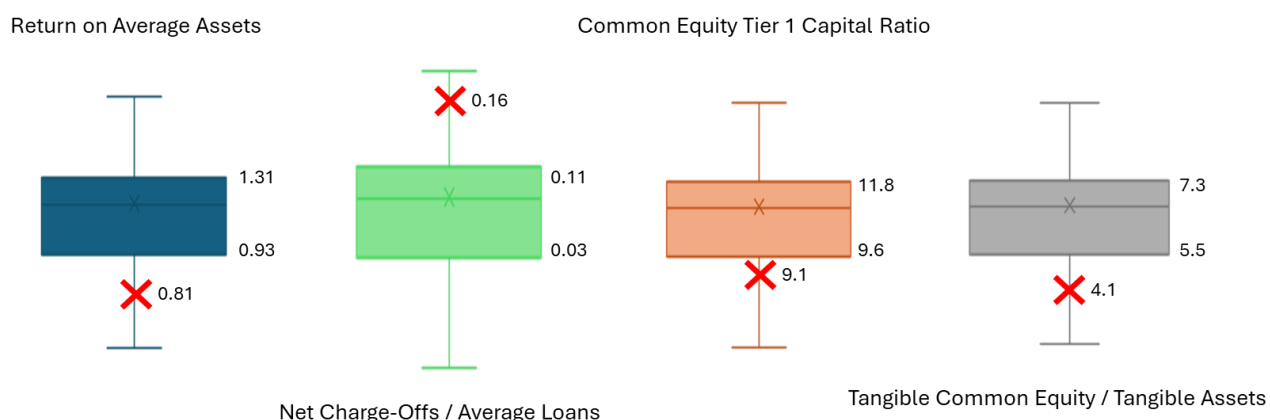


Figura 14. Comparación Entre Bancos Activos y US Bancorp en junio 2023

No es sorprendente cómo US Bancorp se encuentra mucho peor que la media e incluso fuera del percentil 25 en cada una de las variables y en la dirección que corresponde (todas las variables afectan negativamente a la probabilidad de *default* menos *Net Charge-Offs / Average Loans*) para junio de 2023.

A parte de mirar estas métricas financieras, si nos fijamos en el precio de la acción de US Bancorp en el último año, observamos que ha caído un -12%, mientras que el Dow Jones US Banks Index ha crecido un 4%. Todo apunta a que US Bancorp no está pasando por una situación de solvencia estable y el mercado ya se ha dado cuenta repercutiendo en el precio de la acción.

### 6.2.2 Columbia Bank

El caso de Columbia Bank es muy similar al de US Bancorp, en el que el modelo también le da como *default* en marzo y junio de 2023, también provocado por los altos tipo de interés y el alto *quantitative easing*, pero también vale la pena estudiar la comparación de los ratios financieros con el resto de bancos para ver si ocurre lo mismo (ver Figura 15).

Comparison Between Active Banks and Columbia Bank in June 2023

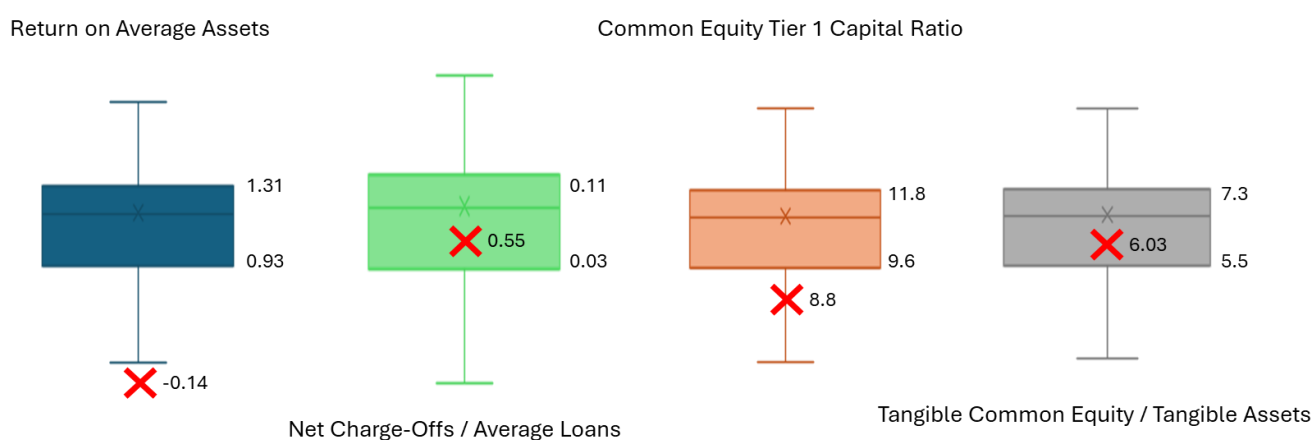


Figura 15. Comparación Entre Bancos Activos y Columbia Bank en junio 2023

Destaca especialmente el *Return on Average Assets* de Columbia, con valor negativo por los retornos negativos de ese año y muy por debajo de la media, siendo un *outlier*. Por otro lado, el *CET1 Capital Ratio* también lo tiene relativamente inferior al resto de bancos en junio de 2023, lo que supone un mayor riesgo de insolvencia, ya que el banco tiene menos capital de alta calidad para absorber pérdidas. Además, puede enfrentarse a restricciones en su capacidad de préstamo, costes de endeudamiento más altos, posibles rebajas en sus calificaciones crediticias, y daño a su reputación en el mercado.

En cuanto al precio de la acción de Columbia Bank en el último año, la situación es todavía peor a la de US Bancorp, ha caído un -38%. Lógicamente, a los inversores no les habrá pasado desapercibidos los retornos negativos durante estos trimestres.

## 7. Conclusiones

El objetivo principal del estudio era desarrollar un modelo que funcionara y fuera capaz de identificar bancos que se encuentran en situaciones parecidas a otros que cayeron en *default*. Considero que es realmente importante entender los precedentes para saber identificar qué banco puede ser un *outlier* y te va a distorsionar el modelo, como puede ser el caso de Heartland Tri-State. El tratamiento de los datos es realmente importante, al igual que intentar recoger el máximo de ellos.

En la literatura, se encuentran muchos ejemplos de análisis de ratios financieros o CAMELs, pero no hemos encontrado precedentes que incluyan variables externas como el *quantitative easing* que, posteriormente, el modelo ha demostrado que afecta enormemente a la probabilidad de *default*. También ha sido uno de los objetivos del estudio ver cómo las variables originales influyen en la probabilidad de *default*, encontrando que *Return on Average Assets*, *Net Charge-Offs / Average Loans*, *Common Equity Tier 1 Capital Ratio*, *Tangible Common Equity / Tangible Assets*, *Interest Rate* y *Quantitative Easing* tienen una correlación muy alta – todas con su interpretación y sentido económico.

Se podría pensar que ha faltado comparar la efectividad del modelo *logit* con la de otros modelos como puede ser el *random forest*, pero con los resultados obtenidos en cuanto a sensibilidad y especificidad (matriz de confusión), podemos concluir que el modelo es bastante satisfactorio. Una vez han salido marcados US Bancorp y Columbia Bank, se ha confirmado que son bancos que no están pasando por una situación sencilla (observando métricas financieras), y se ha contrastado con el precio de la acción en el último año (-12% y -38% respectivamente).

Con estos resultados consideramos que el modelo tiene un uso práctico en la actualidad; únicamente habría que darle las últimas actualizaciones de las variables para ver si un banco puede o no caer, siempre contrastándolo con otras fuentes y estudios como los mencionados en la literatura.

## 8. Declaración de Inteligencia Artificial

Por la presente, yo, Pablo Silvano Andreu Matas, estudiante de Grado en Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas, al presentar mi Trabajo Fin de Grado titulado "LEADING INDICATOR DEL SECTOR BANCARIO EN ESTADOS UNIDOS", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. Interpretador de código: Para realizar análisis de datos preliminares.
3. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
4. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
5. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 20/03/2024

Firma:



## 9. Bibliografía

Alaka H. A., Oyedele L. O., Owolabi H. A., et al. (2018). Systematic review of bankruptcy prediction models: towards a framework for tool selection. *Expert Syst Appl* 94:164–184. <https://doi.org/10.1016/j.eswa.2017.10.040>

Aziz A., Emmanuel D. & Lawson G. (1988). Bankruptcy prediction: an investigation of cash flow-based models. *Journal of Management Studies*, 25(5), 419–437. <http://dx.doi.org/10.1111/j.1467-6486.1988.tb00708.x>

Bailey S. (2023). ‘Huge scam’ in rural Kansas town fells fourth U.S. bank in 2023.

Barrios V. E. & Blanco J. M. (2003). The Effectiveness of bank capital adequacy regulation: A theoretical & empirical approach. *Journal of Banking & Finance*, 27(10), 1935–1958

Beauxis-Aussalet E. & Hardman L. (2014). Visualization of Confusion Matrix for Non-Expert Users.

Beaver W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71. doi:10.2307/2490171

Bell G. G., Mossman C. E., Swartz L. M. & Turtle H. (1998). An empirical comparison of bankruptcy models. *Financial Review*, 33(2), 35–54. <http://dx.doi.org/10.1111/j.1540-6288.1998.tb01367.x>

Bharath S. T. & Shumway T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3), 1339–1369.

Boyacioglu M. A., Kara Y., Baykan Ö. K. (2009) Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Syst Appl* 36:3355–3366

Brédart X. (2014). Financial distress & corporate governance: the impact of board configuration. *Res* 7:72–80. <https://doi.org/10.5539/ibr.v7n3p72>

Burgstahler D. Jiambalvo J. & Noreen E. (1989). Changes in the Probability of Bankruptcy & Equity Value. *Journal of Accounting & Economics*, 11, 207–224. [http://dx.doi.org/10.1016/0165-4101\(89\)90006-2](http://dx.doi.org/10.1016/0165-4101(89)90006-2)

Calvo G. A. & Végh C. A. (1995). Fighting inflation with high interest rates: the small open economy case under flexible prices. *Journal of Money, Credit & Banking*, 27(1), 49-66.

Canbas S. Cabuk A. & Kilic S. B. (2005) Prediction of commercial bank failure via multivariate statistical analysis of financial structures: the Turkish case. *Eur J Oper Res* 166:528–546

Cattell, R. B. (1966). The meaning & strategic use of factor analysis.



Andreu, P.S.

Chang X. Y., Chen S. Q., Wang K. & Zhang W. (2019), "Credit Default Swaps & Corporate Innovation," *Journal of Financial Economics*, 134(2), 474-500.

Christiano L. J., Gust C. & Roldos J. (2004). Monetary policy in a financial crisis. *Journal of Economic theory*, 119(1), 64-103.

Daily C. M., Dalton D. R. (1994) Corporate governance and the bankrupt firm: an empirical assessment. *Strateg Manag J* 15:643–654. <https://doi.org/10.1002/smj.4250150806>

Davidson A. C. & Gigli A. (1989). Deviance residuals & normal scores plots. *Biometrika*, 76(2), pp.211–221.

du Jardin P. (2016). A two-stage classification technique for bankruptcy prediction. *Eur J Oper Res* 254:236–252. <https://doi.org/10.1016/j.ejor.2016.03.008>

Duca M. L., Nicoletti G. & Martinez A. V. (2016). Global corporate bond issuance: What role for US quantitative easing? *Journal of International Money & Finance*, 60(February), 114–150.

Eghian A. (2021). Comparing Machine Learning Techniques for Predicting Bank Failure.

Flagg J. C., Giroux G. A. & Wiggins C. E. (1991). Predicting corporate bankruptcy using failing firms. *Review of Financial Studies*, 1(1), 67–78.

Gheno A. (2007). Corporate valuations & the Merton model. *Applied Financial Economics Letters*, 3(1), 47–50.

Govaers F. (2019). Introduction & Implementations of the Kalman Filter. Google Books. BoD – Books on Demand.

Graf A. (2023). Kansas bank failure reminds industry to be wary of scams, keep guard up. S&P Global.

Härle P., Havas A., Kremer A., Rona D. & Samandari H. (2015). The future of bank risk management McKinsey Working Papers on Risk.

Harris J. K. (2021). Primer on binary logistic regression. *Family Medicine & Community Health*, 9(Suppl 1), p.e001290.

Hoo Z. H., Candlish J. & Teare D. (2017). What is an ROC curve? *Emergency Medicine Journal*, 34(6), pp.357–359. doi:<https://doi.org/10.1136/emered-2017-206735>.

Hosaka T. (2019) Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Syst Appl* 117:287–299. <https://doi.org/10.1016/j.eswa.2018.09.039>

Howley T. Madden, M.G. O'Connell, M.L. Ryder, A.G. (2006). The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. In: Macintosh, A. Ellis, R. Allen, T. (eds) *Applications & Innovations in Intelligent Systems XIII*. SGAI 2005. Springer, London. [https://doi.org/10.1007/1-84628-224-1\\_16](https://doi.org/10.1007/1-84628-224-1_16)

Andreu, P.S.

Laeven L. & Valencia F. (2020). Systemic banking crises database II. *IMF Economic Review*, 68, 307-361.

Lash H. (2023). Credit Suisse bankruptcy event has not occurred - CDS panel. *Reuters Finance*.

Le H. H. & Viviani J. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in International Business & Finance*, 44, 16-25. doi:10.1016/j.ribaf.2017.07.104

Le T. N. L., Nasir M. A. & Huynh T. L. D. (2020). Capital requirements & banks performance under Basel-III: A comparative analysis of Australian & British banks. *The Quarterly Review of Economics & Finance*, 87. doi:https://doi.org/10.1016/j.qref.2020.06.001.

Lee T., Yeh Y. (2004) Corporate governance and financial distress: evidence from Taiwan. *Corp Gov Int. Rev* 12:378–388. https://doi.org/10.1111/j.1467-8683.2004.00379.x

Liang D., Lu C. C., Tsai C. F. & Shih G. A. (2016) Financial ratios and corporate governance indicators in bankruptcy prediction: a comprehensive study. *Eur J Oper Res* 252:561–572. https://doi.org/10.1016/j.ejor.2016.01.012

Lin F., Liang D & Chen E. (2011) Financial ratio selection for business crisis prediction. *Expert Syst Appl* 38:15094–15102. https://doi.org/10.1016/j.eswa.2011.05.035

López Iturriaga F. J. & Sanz I. P. (2015). Bankruptcy visualization & prediction using neural networks: a study of U.S. commercial banks. *Expert Syst Appl* 42:2857–2869. https://doi.org/10.1016/j.eswa.2014.11.025

Love W. D. & Tucker L. R. (1970). A three-mode factor analysis of serial learning. Office of Naval Research Report.

Mensah Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: a methodological study. *Journal of Accounting Research*, 22(1), 380–395. http://dx.doi.org/10.2307/2490719

Merton R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449–470.

Mohan R. (2009). Global financial crisis: Causes, impact, policy responses & lessons. *Reserve Bank of India Bulletin*, 879-904.

Moro Visconti R. (2016). The impact of quantitative easing (QE) on the cost of debt in project finance investments. *Public Money & Management*, 36(2), 129–135.

Ohlson J. A. (1980). financial ratios & the probabilistic prediction of bankruptcy. *J Acc Res* 18:109–131. https://doi.org/10.2307/2490395

Ozili P. K. (2023). Causes & Consequences of the 2023 Banking Crisis.

Andreu, P.S.

Ozili P. K. (2024). Causes & consequences of the 2023 banking crisis. In *Governance & Policy Transformations in Central Banking*. IGI Global.

Park D., Kitwiwattanachai C. & Lee J. (2023). How Credit Default Swaps Affect Risk-Shifting.

Petropoulos A. Siakoulis V. Stavroulakis E. & Vlachogiannakis N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*. doi:<https://doi.org/10.1016/j.ijforecast.2019.11.005>.

Platt H. D. & Platt M. B. (1991). A note on the use of industry-relative ratios in bankruptcy prediction. *Journal of Banking & Finance*, 15(6), 1183–1194. [http://dx.doi.org/10.1016/0378-4266\(91\)90057-S](http://dx.doi.org/10.1016/0378-4266(91)90057-S)

Poon W. P. H., Firth M. & Fung H. G. (1999) A multivariate analysis of the determinants of Moody's bank financial strength ratings. *J Int Financ Mark Inst Money* 9:267–283. [https://doi.org/10.1016/S1042-4431\(99\)00011-6](https://doi.org/10.1016/S1042-4431(99)00011-6)

Premachandra I. M., Bhabra G. S. & Sueyoshi T. (2009). DEA as a tool for bankruptcy assessment: a comparative study with logistic regression technique. *European Journal of Operational Research*, 193(2), 412–424. <http://dx.doi.org/10.1016/j.ejor.2007.11.036>

Serrano-Cinca C. & Gutiérrez-Nieto B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decis Support Syst* 54:1245–1255. <https://doi.org/10.1016/j.dss.2012.11.015>

Sieroñ A. (2023). Monetary Policy After the Coronavirus Crisis. In *The Austrian School of Economics in the 21st Century: Evolution & Impact* (pp. 177-190). Cham: Springer International Publishing.

Subrahmanyam M. G., Tang D. Y. & Wang S.Q. (2014), “Does the Tail Wag the Dog?: The Effect of Credit Default Swaps on Credit Risk,” *Review of Financial Studies*, 27(10), 2927-2960.

Taylor J. B. (2018). Government as a cause of the 2008 financial crisis: A reassessment after 10 years. “Workshop Series on the 2008 Financial Crisis: Causes, The Panic, The Recession, Lessons” *Economics Working Paper* 18116

Vazquez F. & Federico P. (2015). Bank funding structures & risk: Evidence from the Global Financial Crisis. *Journal of Banking & Finance*, 61, 1–14

Wack K. (2023). Dramatic collapses made 2023 the biggest year ever for bank failures. *American Banker*.

Yan X., Skully M. Avram K. & Vu T. (2014). Market Discipline & Deposit Guarantee: Evidence from Australian Banks. *International Review of Finance*, 14(3), 431-457.

Andreu, P.S.

Zavgren C. (1985). Assessing the vulnerability of failure of American industrial firms: a logistic analysis. *Journal of Business Finance & Accounting*, 12(1), 19–45. <http://dx.doi.org/10.1111/j.1468-5957.1985.tb00077>

## 10. Anexo

### 10.1 Anexo 1: Lista Inicial de Todos los Bancos Utilizados

- Almena State Bank – *default* en 23 octubre 2020
- Bank of America
- Bank of New York
- BOK Financial Corporation
- Cadence Bank
- Capital One
- Citigroup
- Citizens
- Citizens State Iowa – *default* en 3 noviembre 2023
- City National Bank of New Jersey – *default* en 1 noviembre 2019
- Columbia Bank
- Comerica Bank
- Cullen Bank
- Discover Bank
- East West Bank
- Ericson State Bank – *default* en 14 febrero 2020
- Fifth Third Bank
- First Citizens
- First City Bank of Florida – *default* en 16 octubre 2020
- First Horizon Bank
- First Republic Bank – *default* en 1 mayo 2023
- First State Bank – *default* en 3 abril 2020
- Goldman Sachs
- Huntington
- JP Morgan
- KeyCorp
- Louisa Community Bank – *default* en 25 octubre 2019
- M&T
- Morgan Stanley
- Northern Trust
- New York Community Bank
- Old National Bank
- PNC
- Popular Bank
- Regions Bank
- Resolute Bank – *default* en 25 octubre 2019
- Signature Bank – *default* en 12 marzo 2023
- Silicon Valley Bank – *default* en 10 marzo 2023

- State Street
- Synovus Bank
- Truist
- US Bancorp
- Valley National Bank
- Webster Bank
- Wells Fargo
- Western Alliance Bank
- Wintrust Bank
- Zions Bank

## 10.2 Anexo 2: Antiguos Estudios de la Probabilidad de Default de los Bancos

Authors	Sector	Country	Sample Size	Model	Period	Time-Horizon of Prediction	Categories
López Iturriaga and Sanz, (2015)	Commercial Banks only	US	Training: 386 failed banks—386 non-failed randomly selected/ Test: 52 failed—52 non-failed	NN: MLP and SOM	training: 2002–2012/ Test: 2012–2013	1,2, and 3 years	5 sets: bank's earning, asset structure, loan portfolio, risk concentration, solvency
Wang et al. (2014)	Financial Institutions	Poland	240 (112 failed companies)  132 (66 risk cases—66 non-risk cases)	Feature Selection Boosting	1997–2001  1970–1982	–  –	Financial ratios  Financial ratios
Serrano-Cinca and Gutiérrez-Nieto (2013)	Banks	US	Training: 140 failed banks, 140 non-failed banks— Test: 180 failed banks, 7833 non-failed banks	Partial Least Square—Discriminant Analysis	training: 2009— Test: 2010–2011	–	Income and expense to asset, profitability, efficiency, assets, capital
Feki et al. (2012)	Commercial Banks	Tunisia	Training: 50. Test: 10	Bayes models and vector machine	2000–2006	–	CAMELS and Size

Authors	Sector	Country	Sample Size	Model	Period	Time-Horizon of Prediction	Categories
Chauhan et al. (2009)	Banks	US, Turkish, Spanish	Turkish: 22 bankrupt, 12 healthy/ Spanish 37 bankruptcy, 29 healthy/ US: 65 bankrupt, 64 healthy	DEWNN	Spanish: 1982, US: 1975–1982	1 year	Financial ratios
Boyacioglu et al. (2009)	Banks	Turkey	21 bankrupts (14 training and 7 test)/ 44 non-failed (29 training and 15 test) (randomly selected double the failed)	T-test, PCA, DA and Artificial NN	1997–2004	–	CAMELS
Liang et al. (2016)	Non-financial firms	Taiwan	239 bankrupt, 239 non-bankrupt	SVM, KNN, NB, CART, and statistics	1999–2009	3 years	7 categories of financial ratios and 5 categories of CG
Lee and Yeh (2004)	all listed firms	Taiwan	45 distressed, 88 healthy (double the failed sample)	logistic regression	1996–1999	1 year	2 categories of CG, profitability, R&D
Wu (2016)	Non-financial firms	US	217 bankrupt, 9,100 non-bankrupt	multi-period logit model	1996–2006	1 and 2 years	3 categories of CG, financial ratios
Daily and Dalton (1994)	listed firms in 1990	US	50 bankrupt, 50 healthy	logistic regression	1990	3 and 5 years	CG and financial ratios
Brédart (2014b) Brédart (2014a)	listed in AMEX, Nasdaq, NYSE	US	312 firms	logit model	2007–2009	–	CG and financial ratios
Jones (2017)	public firms	US	1115 bankrupt	Gradient Boosting Model	1987–2013	3 years	CG, market, accounting, macro-economic
Canbas et al. (2005)	Private banks	Turkey	21 failed, 19 non-failed	PCA, discriminant, logit, probit	1997–2003	1,2, and 3 years	Financial ratios
Kao and Liu (2004)	Commercial Banks	Taiwan	24 banks	Data envelopment analysis	2000	1 year	Financial ratios

Authors	Sector	Country	Sample Size	Model	Period	Time-Horizon of Prediction	Categories
du Jardin (2010)	Retail sector	France	Train: 250 bankrupt- 250 healthy/Test 260 Bankrupt/260 Healthy	Neural Network	Train: 2006–2007/ Test: 2004–2005	–	Liquidity, Financial structure, Profitability, Efficiency, Rotation, Withdrawal, Contribution
Lin et al. (2011)	publicly listed	Taiwan	120 distressed and 120 non-distressed	Support Vector Machine	2000–2008	1,2,3 years	Financial ratios
du Jardin (2017)	firms	France	95,910 non-failed firms and 1920 failed firms	logistic regression, DA, NN, VM, boosting	1997–2003	5 years	liquidity, turnover, profitability, activity, solvency and financial structure
Hosaka (2019)	publicly listed	Japanese	102 bankrupt, 2062 healthy	NN	2002–2016	up to 3 years	balance sheet, income statement
Beaver (1966)	Industrial publicly owned	US	79 failed firms, 79 non-failed firms	Profile Analysis	1949–1963	up to 5 years	6 groups
Altman (1968)	Manufacturers	US	66	Multiple Discriminant Analysis (MDA)	1946–1965		5 groups: liquidity, profitability, leverage, solvency, activity ratio
Ohlson (1980)	industrial	US	105 bankrupt, 2058 data vectors for non-bankrupt	Logit Analysis, MDA	1070–1976	1, 2 years	–
Poon et al., (1999)	Banks	30 countries	130	Logistic	1996–1997	–	Financial ratios
Ioannidis et al. (2010)	Banks	78 countries	944	OLS and NN	2007–2008	–	Bank-level variables, regulatory variables, country-level variables
Demirgüç-Kunt et al. (2008)	Banks	39 countries	304	regression	1999–2003	–	–



Authors	Sector	Country	Sample Size	Model	Period	Time-Horizon of Prediction	Categories
Gasbarro et al. (2002)	Banks	Indonesia	126	GLS	1993–1997	–	Camels
du Jardin (2016)	firms	France	16,240 observations per year	DT, DA, Logistic, NN	training 2002–2011 testing 2003–2012	1, 2, and 3 years	Activity, Financial structure, Profitability, Turnover, liquidity, solvency
Cox and Wang (2014)	banks	US	322 failed banks	DA	2003–2008	1,2,3 and 4	