



Revista de Educación Estadística

ISSN: (Impreso) 1069-1898 (En línea) Página principal de la revista: <https://amstat.tandfonline.com/loi/ujse20>

"¿Debería aprobarse o denegarse este préstamo?": Un gran conjunto de datos con pautas para la asignación de clases

Min Li, Amy Mickel y Stanley Taylor


Para citar este artículo: Min Li, Amy Mickel y Stanley Taylor (2018) "¿Debería aprobarse o denegarse este préstamo?": Un gran conjunto de datos con pautas para la asignación de clases, Journal of Statistics Education, 26:1, 55-66, DOI: [10.1080/10691898.2018.1434342](https://doi.org/10.1080/10691898.2018.1434342)

Para enlazar a este artículo: <https://doi.org/10.1080/10691898.2018.1434342>



© Min Li, Amy Mickel y Stanley Taylor©
Min Li, Amy Mickel y Stanley Taylor



Ver material complementario 



Publicado en línea: 05 de abril de 2018.



Envíe su artículo a esta revista



Vistas del artículo: 5351



Ver datos de Crossmark 

"¿Debería aprobarse o denegarse este préstamo?": Un gran conjunto de datos con pautas para la asignación de clases

Min Li, Amy Mickel y Stanley Taylor

Facultad de Administración de Empresas, Universidad Estatal de California, Sacramento, CA

ABSTRACTO

En este artículo, se presenta un conjunto de datos grande y rico de la Administración de Pequeñas Empresas (SBA) de EE. UU. y una tarea adjunta diseñada para enseñar estadísticas como un proceso de investigación de toma de decisiones. Se proporcionan pautas para la tarea titulada "¿Debería aprobarse o denegarse este préstamo?", junto con un subconjunto del conjunto de datos más grande. Para esta tarea de estudio de caso, los estudiantes asumen el papel de oficial de crédito en un banco y se les pide que aprueben o nieguen un préstamo evaluando su riesgo de incumplimiento utilizando la regresión logística. Dado que esta tarea está diseñada para cursos introductorios de estadística empresarial, también se sugieren métodos adicionales para cursos de análisis de datos más avanzados.

PALABRAS CLAVE

Caso de estudio; Clasificación;
Regla de decisión;
Regresión logística; datos reales;
Indicador de riesgo

1. Introducción

En las Directrices para la Evaluación y la Instrucción en Educación Estadística (GAISE) de la Asociación Estadounidense de Estadística (ASA).

College Report (GAISE College Report ASA Revision Committee 2016), se hicieron las siguientes recomendaciones para enseñar estadística introductoria:

(a) Enseñar pensamiento

estadístico. Enseñar estadística como un proceso investigativo de resolución de problemas y toma de decisiones.

Proporcione a los estudiantes experiencia con el pensamiento multivariable. (b) Centrarse en la comprensión conceptual. (c) Integrar datos reales con un contexto y propósito. (d) Fomentar el

aprendizaje activo. (e) Usar la tecnología para explorar conceptos y analizar datos. (f) Usar evaluaciones para mejorar y evaluar a los estudiantes aprendiendo.

En este artículo, tomamos en cuenta estas recomendaciones al proporcionar un conjunto de datos rico y grande que en sí mismo es una contribución significativa, ya que los educadores pueden utilizarlo para crear oportunidades de aprendizaje que estén alineadas con las recomendaciones de GAISE de 2016. Junto con el conjunto de datos, también se describe un conjunto de pautas para una tarea de estudio de caso diseñada con las recomendaciones antes mencionadas en mente.

El conjunto de datos que acompaña a este artículo es un conjunto de datos real de la Administración de Pequeñas Empresas (SBA) de EE. UU. La tarea del estudio de caso, titulada "¿Debe aprobarse o denegarse este préstamo?" está diseñado para enseñar el pensamiento estadístico centrándose en cómo usar datos reales para tomar decisiones informadas para un propósito particular. Para esta tarea, los estudiantes asumen el rol de un oficial de préstamos que decide si aprobar un préstamo para una pequeña empresa.

Al analizar datos reales, los estudiantes experimentan la estadística como un proceso investigativo de toma de decisiones, ya que el estudiante es

obligado a responder la siguiente pregunta: ¿Como representante del banco, debo otorgar un préstamo a una pequeña empresa en particular (Empresa X)? ¿Por qué o por qué no? El estudiante toma esta decisión evaluando el riesgo de un préstamo.

La evaluación se logra estimando la probabilidad de incumplimiento del préstamo a través del análisis de este conjunto de datos históricos y luego clasificando el préstamo en una de dos categorías: (a) mayor riesgo: probabilidad de incumplimiento del préstamo (es decir, ser cancelado/no pagar en su totalidad)) o (b) menor riesgo: es probable que pague el préstamo en su totalidad. El proceso de hacer esta determinación requiere que los estudiantes comprendan conceptualmente los conceptos estadísticos y cómo aplicarlos.

Hemos utilizado una versión adaptada de esta asignación de estudio de caso en cursos de análisis de datos para estudiantes de negocios de pregrado y posgrado. Estos cursos cubren temas que van desde regresión y análisis de varianza en el curso de pregrado hasta minería de datos en el curso de posgrado. Para todos los cursos, la regresión logística se incluye en la tarea, mientras que las redes neuronales y las máquinas de vectores de soporte (SVM) se presentan solo en el curso de posgrado.

Para ambos cursos, inicialmente presentamos esto como una tarea interactiva en clase. Pasamos dos o tres periodos de clase de 75 minutos en laboratorios de computación guiando a los estudiantes a través de pasos específicos sobre cómo analizar este gran conjunto de datos para ayudar a informar sus procesos de toma de decisiones. Para fomentar un ambiente de aprendizaje activo, alentamos la discusión y las preguntas durante estos periodos de clase y, por lo general, dividimos a los estudiantes en grupos para discutir ciertos pasos y luego les pedimos que presenten sus ideas y fundamentos. Para evaluar el pensamiento estadístico de los estudiantes, se les presenta un caso similar y se les pide que escriban un informe que describa sus decisiones de préstamo y la justificación detrás de tales decisiones.

Esta asignación es ideal para cursos de análisis de datos para varios razones.

- (a) El estudio de caso incorpora todo el GAISE 2016 recomendaciones
- (b) El tema en sí capta el interés de los estudiantes, ya que es una aplicación de datos reales relacionados con decisiones financieras de la vida real.
- (c) Los estudiantes están expuestos a administrar un gran conjunto de datos y comprender cómo se pueden usar los datos históricos para tomar decisiones informadas.
- d)Se promueva el pensamiento crítico; análisis, síntesis y Se utilizan habilidades para la toma de decisiones.
- (e) Los estudiantes son introducidos a la regresión logística y otros métodos más avanzados para la clasificación. (f) La importancia de identificar variables explicativas razonables (p. ej., indicadores de riesgo de impago de préstamos) para incorporarlos en los modelos estadísticos genera discusiones animadas y atractivas.

Además, los instructores de estadística empresarial han informado que el uso de asignaciones de estudios de casos ha dado como resultado una mayor motivación y participación de los estudiantes, una mayor conciencia de los estudiantes sobre la relevancia de las estadísticas para la toma de decisiones comerciales y experiencias de clase más positivas para el instructor (p. ej., Bryant 1999; Nolan y Speed 1999; Parr y Smith 1998; Smith y Bryant 2009). Hemos experimentado beneficios similares con esta asignación de estudio de caso.

2. Antecedentes y descripción de los conjuntos de datos

La SBA de EE. UU. se fundó en 1953 con el principio de promover y ayudar a las pequeñas empresas en el mercado crediticio de EE. UU. (SBA Overview and History, US Small Business Administration (2015)). Las pequeñas empresas han sido una fuente principal de creación de empleo en los Estados Unidos; por lo tanto, fomentar la formación y el crecimiento de pequeñas empresas tiene beneficios sociales al crear oportunidades de trabajo y reducir el desempleo. Una forma en que la SBA ayuda a estas pequeñas empresas es a través de un programa de garantía de préstamo que está diseñado para alentar a los bancos a otorgar préstamos a las pequeñas empresas. La SBA actúa como un proveedor de seguros para reducir el riesgo de un banco al asumir parte del riesgo garantizando una parte del préstamo. En el caso de que un préstamo entre en incumplimiento, la SBA cubre el monto garantizado.

Ha habido muchas historias de éxito de empresas emergentes que recibieron garantías de préstamos de la SBA, como FedEx y Apple Computer. Sin embargo, también ha habido historias de pequeñas empresas y/o nuevas empresas que han incumplido con sus préstamos garantizados por la SBA. La tasa de morosidad de estos préstamos ha sido motivo de controversia durante décadas. Los economistas conservadores creen que los mercados de crédito funcionan de manera eficiente sin la participación del gobierno. Los partidarios de los préstamos garantizados por la SBA argumentan que los beneficios sociales de la creación de empleo por parte de las pequeñas empresas que reciben préstamos garantizados por el gobierno superan con creces los costos incurridos por los préstamos en mora.

Dado que los préstamos de la SBA solo garantizan una parte del saldo total del préstamo, los bancos incurrirán en algunas pérdidas si una pequeña empresa no cumple con su préstamo garantizado por la SBA. Por lo tanto, los bancos todavía se enfrentan

Tabla 1(a). Descripción de 27 variables en ambos conjuntos de datos.

Nombre de la variable	Tipo de datos	Descripción de la variable
PréstamoNr_ChkDgt	Texto	Identificador – Clave principal
Nombre	Texto	nombre del prestatario
Ciudad	Texto	ciudad prestataria
Estado	Texto	Estado prestatario
Código postal	Texto	Código postal del prestatario
Banco	Texto	Nombre del banco
BankState	Texto	estado del banco
NAICS	Texto	Código del sistema de clasificación de la industria de América del Norte
Fecha de aprobación	Fecha y hora	Fecha de emisión del compromiso de la SBA
AprobaciónFY	Texto	Ejercicio fiscal del compromiso
Término	Número	Plazo del préstamo en meses
NoEmp	Número	Número de empleados de la empresa
NuevoExiste	Texto	1 D Negocio existente, 2 D Nuevo negocio
creartrabajo	Número	Número de puestos de trabajo creados
Trabajo retenido	Número	Número de trabajos retenidos
FranquiciaCódigo	Texto	Código de franquicia, (00000 o 00001) D Sin franquicia
Urbano rural	Texto	1 D Urbano, 2 D rural, 0 D indefinido
RevLineCr	Texto	Línea de crédito renovable: YD Sí, ND No
bajodoc	Texto	Programa de préstamos LowDoc: YD Sí, ND No
FechaCambiar	Fecha/Hora	La fecha en que se declara un préstamo como en defecto
Fecha de desembolso	Fecha y hora	Fecha de desembolso
DesembolsoBruto	Divisa	Monto desembolsado
SaldoBruto	Divisa	Importe bruto pendiente
MIS_Estado	Texto	Estado del préstamo cancelado D CHGOFF, Pagado en su totalidad D PIF
ChgOffPrintGr	Divisa	Importe cancelado
GrAppv	Divisa	Importe bruto del préstamo aprobado por el banco
SBA_Appv	Divisa	la cantidad garantizada de SBA de préstamo

con una elección difícil en cuanto a si deben otorgar dicho préstamo debido al alto riesgo de incumplimiento. Una forma de informar su toma de decisiones es mediante el análisis de datos históricos relevantes, como los conjuntos de datos proporcionados aquí.

Se proporcionan dos conjuntos de datos: (a) Conjunto de datos "National SBA" (denominado SBAnational.csv) de la SBA de EE. UU. que incluye datos históricos desde 1987 hasta 2014 (899,164 observaciones)¹ y (b) Conjunto de datos "SBA Case" (denominado SBACase. csv) que se utiliza en la tarea descrita en este documento (2102 observaciones). El conjunto de datos del "Caso de la SBA" es un subconjunto de la "SBA nacional".²

El nombre de la variable, el tipo de datos y una breve descripción de cada variable se proporcionan para las 27 variables en los dos conjuntos de datos (ver [Tabla 1\(a\)](#)). Para el conjunto de datos del "Caso de la SBA", los autores generaron ocho variables adicionales como parte de la asignación (consulte la [Tabla 1\(b\)](#)) y se describen en las Secciones [4.1.4](#), [4.1.5](#), [4.1.6](#), [4.1.7](#), y [4.3.1](#). Para la mayoría de las variables, la descripción es evidente. Las variables que necesitan más explicación incluyen: NAICS, NewExist, LowDoc y MIS_Status y se describen a continuación.

NAICS (Sistema de Clasificación de la Industria de América del Norte): Este es un sistema de clasificación jerárquico de 2 a 6 dígitos utilizado por las agencias estadísticas federales para clasificar los establecimientos comerciales para la recopilación, el análisis y la presentación de información.

¹ Tenga en cuenta que el conjunto de datos que proporcionamos aquí está restringido a préstamos que se originan en los 50 estados de los Estados Unidos y Washington DC (se excluyeron los territorios de los EE. UU.) y para los cuales se conoce el resultado (pago total o cancelado/incumplimiento); para enseñar la regresión logística, se requiere una variable dependiente binaria.
² El código SAS utilizado para crear el subconjunto de datos se encuentra en el archivo de documentación de datos del "Caso SBA" adjunto.

Tabla 1(b). Descripción de 8 variables adicionales en el conjunto de datos de casos de la SBA.

Nombre de variable	Tipo de datos	Descripción de la variable
Nuevo	Número	D1 si NewExistD2 (Negocio nuevo), D0 si NewExistD1 (negocio existente)
Parte	Número	Proporción del monto bruto garantizado por la SBA
Bienes raíces	Número	D1 si el préstamo está respaldado por bienes inmuebles, D0 en caso contrario
Recesión	Número	D1 si el préstamo está activo durante la Gran Recesión, D0 de lo contrario
Seleccionado	Número	D1 si los datos se seleccionan como datos de entrenamiento para construir el modelo para la asignación, D0 si los datos se seleccionan como datos de prueba para validar el modelo
Periodo de días predeterminado	Número	D1 si MIS_StatusDCHGOFF, D0 si MIS_StatusDPIF
XX	Número	Variable adicional generada al crear "Recesión" en la Sección 4.1.6
	Número	Variable adicional generada al crear "Recesión" en la Sección 4.1.6

datos estadísticos que describen la economía estadounidense. Los dos primeros dígitos de la clasificación NAICS representan el sector económico. La Tabla 2 muestra los sectores de 2 dígitos y una descripción correspondiente para cada sector.

Nota didáctica: La tabla de códigos NAICS de dos dígitos publicada por la Oficina del Censo de EE. UU. (<http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chartD2012>) fusiona algunos sectores (ver Manufactura, Comercio al por menor, Transporte y Almacenamiento). Para ser coherente con la publicación de la Oficina del Censo de EE. UU., también hacemos las mismas fusiones. Sin embargo, es posible que los instructores deseen examinar los sectores individuales de fabricación, comercio minorista, transporte y almacenamiento.

NewExist (1 D Negocio existente, 2 D Nuevo negocio): Esto representa si el negocio es un negocio existente (en existencia por más de 2 años) o un nuevo negocio (en existencia por menos de o igual a 2 años).

LowDoc (YD Sí, ND No): Para poder procesar préstamos de manera más eficiente, se implementó un programa de "Préstamo LowDoc" en el que se pueden procesar préstamos de menos de \$150,000 mediante una solicitud de una página. "Sí" indica préstamos con una solicitud de una página y "No" indica préstamos con más información adjunta a la solicitud. En este conjunto de datos, el 87,31 % está codificado como N (No) y el 12,31 % como Y (Si), para un total de 99,62 %. Cabe resaltar que

Tabla 2. Descripción de los dos primeros dígitos del NAICS.

Sector	Descripción
11	Agricultura, silvicultura, pesca y caza
21	Minería, canteras y extracción de petróleo y gas
22	Utilidades
23	Construcción
31–33	Fabricación
42	Comercio al por mayor
44–45	Comercio al por menor
48–49	Transporte y almacenamiento
51	Información
52	Finanzas y Seguros
53	Inmobiliaria y alquiler y arrendamiento
54	Servicios profesionales, científicos y técnicos.
55	Gestión de empresas y empresas.
56	Servicios administrativos y de apoyo y gestión y remediación de residuos
61	Servicios educativos
62	Asistencia sanitaria y asistencia social
71	Artes, entretenimiento y recreación
72	Servicios de alojamiento y alimentación
81	Otros servicios (excepto administración pública)
92	Administración Pública

el 0,38% tiene otros valores (0, 1, A, C, R, S); estos son errores de entrada de datos. También hay 2582 valores perdidos para esta variable, excluidos al calcular estas proporciones. Hemos optado por dejar estas entradas "como están" para brindarles a los estudiantes la oportunidad de aprender cómo manejar conjuntos de datos con tales errores.

MIS_Status: Esta variable indica el estado del préstamo: en mora/caído (CHGOFF) o bien pagado en su totalidad (PIF).

3. Consideraciones sobre la creación previa a la tarea

Antes de la asignación del estudio de caso, se sugiere que los educadores consideren: (a) desarrollar objetivos de aprendizaje para la asignación; (b) usar paquetes de software de análisis estadístico que sean de fácil acceso para los estudiantes para el análisis; (c) determinar un período de tiempo a ser incluido en los análisis; y (d) decidir cómo integrar la tarea del estudio de caso en una clase y las formas de evaluar el aprendizaje.

3.1. Objetivos de aprendizaje

Podría decirse que este es el paso más importante antes de la creación de la tarea. Es necesaria una comprensión clara y una explicación de lo que la tarea está diseñada para enseñar. Para la sección "¿Debe aprobarse o denegarse este préstamo?" tarea, queremos que nuestros estudiantes:

- 1. Analizar un gran conjunto de datos para promover el pensamiento estadístico;
- 2. Identificar qué variables explicativas pueden ser buenos "predictores" o indicadores de riesgo del nivel de riesgo asociado con un préstamo;
- 3. Trabaja a través de las etapas en la construcción de modelos y validación;
- 4. Aplicar la regresión logística (y otros métodos más avanzados para estudiantes de posgrado) para clasificar un préstamo según el riesgo previsto de incumplimiento; y
- 5. Tomar una decisión basada en escenarios informada por análisis de datos (es decir, si financiar el préstamo).

3.2. Paquetes de software de análisis estadístico Los

conjuntos de datos se preparan para el análisis en la mayoría de los paquetes de software de análisis estadístico disponibles. Se sugiere que los educadores elijan un paquete de software al que los estudiantes puedan acceder fácilmente y pagar. Usamos productos Microsoft Excel, R y SAS (JMP, University Edition) porque están disponibles para nuestros estudiantes sin cargo.

Para nuestros estudiantes, exportamos los datos en los siguientes formatos: datos permanentes SAS (.sas7bdat) y valores separados por comas (.csv). Hacemos que nuestros estudiantes de pregrado usen JMP para abrir el archivo de datos SAS para realizar una regresión logística y otros análisis. La interfaz de apuntar y hacer clic fácil de usar de JMP es perfecta para nuestro curso de análisis de datos de pregrado. Nuestros estudiantes de MBA utilizan R para abrir el archivo de datos de valores separados por comas y realizar análisis que incluyen regresión logística, redes neuronales y SVM.



3.3. Periodo de tiempo

Los educadores también pueden querer considerar qué período de tiempo incluir en los análisis. Por ejemplo, en nuestra tarea, se pone énfasis en las tasas de incumplimiento de los préstamos con una fecha de desembolso hasta 2010.³ Elegimos este período de tiempo por dos razones. Nosotros quiere tener en cuenta la variación debido a la Gran Recesión (diciembre de 2007 a junio de 2009)⁴; por lo tanto, se necesitan préstamos desembolsados antes, durante y después de este período. En segundo lugar, restringimos el marco temporal a los préstamos al excluir los desembolsados después de 2010 debido a que el plazo de un préstamo suele ser de 5 años o más.⁵

Creemos que la inclusión de préstamos con fechas de desembolso

posteriores a 2010 daría mayor peso a aquellos préstamos que se cancelan frente a los que se pagan en su totalidad. Más específicamente, los préstamos cancelados lo harán antes de la fecha de vencimiento del préstamo, mientras que los préstamos que probablemente se pagarán en su totalidad lo harán en la fecha de vencimiento del préstamo (que se extendería más allá del conjunto de datos que finaliza en 2014). . Dado que este conjunto de datos se ha restringido a los préstamos para los que se conoce el resultado, existe una mayor probabilidad de que los préstamos cancelados antes de la fecha de vencimiento se incluyan en el conjunto de datos, mientras que los que podrían pagarse en su totalidad han sido excluidos. Es importante tener en cuenta que cualquier restricción de tiempo sobre los préstamos incluidos en los análisis de datos podría introducir un sesgo de selección, en particular hacia el final del período de tiempo. Esto puede afectar el rendimiento de cualquier modelo predictivo

3.4. Formato de la asignación de estudio de caso

Esta tarea se puede adaptar para cursos presenciales, híbridos y en línea. Si bien describimos cómo se ha aplicado esta tarea en nuestros cursos presenciales, alentamos a los instructores a adaptar las tareas para satisfacer las necesidades de los estudiantes y los diversos modos de entrega.

Tanto para los cursos de pregrado como para los de posgrado, inicialmente presentamos esto como una tarea interactiva en clase. Pasamos dos o tres períodos de clase de 75 minutos para guiar a los estudiantes a través de los diversos pasos que se describen a continuación. Alentamos la discusión y las preguntas durante estos períodos de clase. Para promover el aprendizaje activo, dividimos a los estudiantes en grupos para discutir ciertos pasos y luego les pedimos que presenten sus ideas y fundamentos. Como instructores, facilitamos una discusión de clase más grande después de estas presentaciones para garantizar que los estudiantes comprendan los diversos pasos.

Para evaluar el aprendizaje de los estudiantes, desarrollamos una tarea de estudio de caso calificada que es similar a la presentada en clase. Para los estudiantes universitarios, les dejamos completar la tarea en grupos de tres personas. Para los cursos de posgrado, los estudiantes deben completar la tarea como individuos.

4. Directrices para "¿Debería aprobarse o aprobarse este préstamo?" ¿Denegado?" Asignación de estudio de caso

Esta sección está organizada en torno a los pasos involucrados en el proceso de investigación de analizar estos datos para hacer una

decisión informada sobre si un préstamo debe ser aprobado o denegado, uno de los principales objetivos de aprendizaje de esta asignación.

Los estudiantes son guiados a

- través de: Paso 1: Identificar indicadores de riesgo potencial; Paso 2: Comprender el estudio de caso; Paso 3: Construir el modelo, crear reglas de decisión y validar el modelo de regresión logística; y
- Paso 4: Usar el modelo para tomar decisiones.

4.1. Paso 1: Identificación de variables explicativas (indicadores o predictores) del riesgo potencial

En el primer período de clase, proporcionamos a los estudiantes el conjunto de datos de la "SBA nacional", los antecedentes de la SBA y la tarea con sus objetivos de aprendizaje. Dado que los modelos económicos deben basarse en una teoría económica sólida, involucramos a los estudiantes en una discusión que requiere que identifiquen qué variables explicativas creen que serían buenos indicadores o predictores del riesgo potencial de un préstamo: probabilidad de incumplimiento (mayor riesgo) versus pagado en su totalidad (menor riesgo).

Para cumplir con el siguiente objetivo de aprendizaje, para identificar qué variables explicativas pueden ser buenos predictores o indicadores de riesgo del nivel de riesgo asociado con un préstamo, animamos a los estudiantes a considerar las tasas de morosidad para un grupo que está representado por el porcentaje de préstamos que se clasifican como predeterminados. Para un grupo particular de préstamos, la tasa de morosidad se determina usando la variable "MIS_Status" y calculando el porcentaje del número total de préstamos (CHGOFF C PIF) que se clasifican como morosos (CHGOFF).

Nota didáctica: dividimos a los estudiantes en grupos para la discusión y les pedimos que proporcionen una justificación por escrito para cada variable en cuanto a si sería un buen indicador de riesgo y les pedimos que los presenten brevemente a la clase. Esta actividad refuerza la importancia de tener una teoría sólida al construir modelos y promueve el aprendizaje activo.

Hay una serie de variables que emergen consistentemente como indicadores de riesgo que podrían explicar la variación de las tasas de morosidad. A continuación, se analizan siete variables, junto con algunos análisis exploratorios, que incluyen ubicación (estado), industria, desembolso bruto, negocios nuevos versus establecidos, préstamos respaldados por bienes raíces, recesión económica y la porción garantizada del préstamo aprobado por la SBA. Para varios de estos indicadores, se crean variables ficticias para el análisis y se discuten en la enseñanza.

notas

4.1.1. Ubicación (Estado)

La ubicación por estado (representada como "Estado" en la Tabla 1(a)) es un posible predictor que los estudiantes identifican en sus discusiones. Reconocen que los 50 estados y Washington DC tienen diferentes entornos económicos en los que operan, lo que da como resultado diferentes tasas de incumplimiento. Mostramos este mapa de calor (Figura 1) en clase para apoyar esta discusión.

Nota didáctica: Se alienta a los estudiantes a explorar las razones de las diferencias en las tasas de incumplimiento por estado. Por ejemplo, durante la Gran Recesión, Florida tuvo una gran caída en los precios de bienes raíces que podría contribuir a altas tasas de incumplimiento; estados como Wyoming y Dakota del Norte tenían economías más sólidas (debido a su dependencia de los minerales y el petróleo), lo que puede explicar sus tasas de incumplimiento más bajas. Dado que operamos en California, California

³"DisbursementDate" es la variable utilizada para determinar esta clasificación.

⁴Las fechas declaradas por la Oficina Nacional de Investigación Económica (ver http://money.cnn.com/2010/09/20/news/economy/recession_over/)

⁵La distribución del plazo de los préstamos es tal que la moda es de 7 años (el 27% de los préstamos tienen una duración de 7 años) y el 73% tienen una duración superior a los 5 años. Para aquellos préstamos dispersos a partir de 2010, el 66% tiene plazos mayores a 5 años.

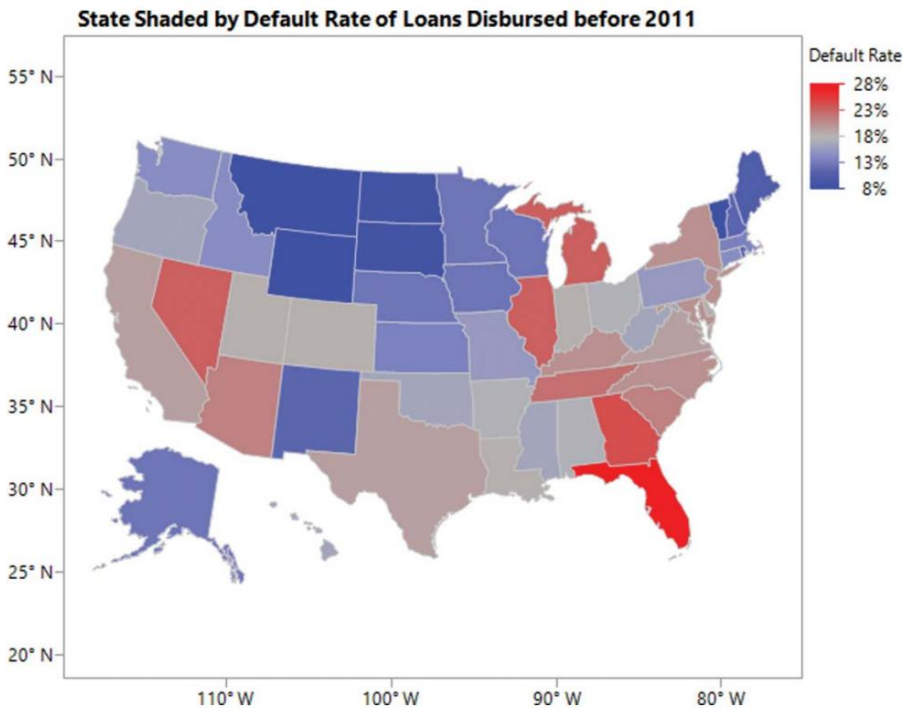


Figura 1. Mapa de calor, tasas predeterminadas por estado (la Figura 1 se creó utilizando JMP).

en relación con otros estados se destaca, ya que "lleva a casa" la discusión. Los profesores pueden optar por centrarse en estados de interés para sus alumnos.

4.1.2. Industria En

la Tabla 3, la industria (los primeros dos dígitos de los códigos NAICS) es otro indicador de riesgo que los estudiantes consideran debido a la cantidad significativa de variación en las tasas de incumplimiento. En un extremo del espectro se encuentran industrias con bajas tasas de incumplimiento (8%–10%), como: minería, exploración de petróleo y gas (21), agricultura (11), sociedades de cartera de valores (55) y médicos y dentistas (62). En el extremo opuesto del espectro se encuentran las industrias con tasas de incumplimiento más altas (28 %–29 %), como instituciones financieras como cooperativas de crédito (52) y agencias inmobiliarias (53).

La variación en las tasas de incumplimiento de la industria a menudo se debe a la naturaleza cíclica de la demanda de productos o servicios. Por ejemplo, la industria de la construcción (23) se expande y contrae dramáticamente durante un ciclo económico, mientras que la industria de servicios médicos (62) tiende a ser mucho más estable; en consecuencia, los ingresos y la utilidad neta son mucho menos volátiles para los servicios médicos que para la construcción. Además, a diferencia de la construcción, los servicios médicos tienen requisitos de licencia que crean barreras que las nuevas empresas deben superar.

Dado que no es fácil ingresar a los servicios médicos, quienes ingresan a esta industria toman muy en serio su nueva empresa y esto contribuye aún más al menor riesgo crediticio de la industria médica.

Al igual que la construcción, otra industria que tiene una tasa de morosidad más alta es la industria de alojamiento y servicios de alimentos (es decir, hospitalidad) (72). Con el tiempo, la morosidad de los préstamos hoteleros tiende a ser alta porque los hoteles a menudo construyen en exceso nuevas unidades cuando las tasas de ocupación son altas y luego pueden enfrentar tasas de ocupación bajas por una variedad de razones inesperadas.

Con respecto al servicio de alimentos, el éxito de cualquier restaurante nuevo es muy impredecible, y el éxito continuo de los restaurantes existentes a menudo se ve amenazado por nuevas empresas.

Nota didáctica: En nuestras clases, tendemos a usar los códigos de dos dígitos que se muestran en la Tabla 3. Sin embargo, uno puede hacer que sus alumnos usen dígitos adicionales en el análisis. Por ejemplo, los médicos se codifican como 6211 y los dentistas como 6212. El siguiente enlace proporciona el esquema de codificación con mayor detalle que los que se proporcionan en la Tabla 3: [http://www.census.gov/cgi-bin/sssd/naics/naicsrch? gráficoD201](http://www.census.gov/cgi-bin/sssd/naics/naicsrch?gráficoD201). Estas definiciones, junto con los códigos detallados proporcionados en la variable NAICS, permitirán a los estudiantes analizar industrias más específicas.

Tabla 3. Tasas de incumplimiento de la industria (primeros códigos NAICS de dos dígitos).

código de 2 dígitos	Descripción	Tasa de incumplimiento (%)
21	Minería, canteras y petróleo y gas extracción	8
11	Agricultura, silvicultura, pesca y caza	9
55	Gestión de empresas y empresas.	10
62	Asistencia sanitaria y asistencia social	10
22	Utilidades	14
92	Administración Pública	15
54	Servicios profesionales, científicos y técnicos.	19
42	Comercio al por mayor	19
31–33	Fabricación	19, 16, 14 20
81	Otros servicios (excepto administración pública)	
71	Artes, entretenimiento y recreación	21
72	Servicios de alojamiento y alimentación	22
44–45	Comercio al por menor	22, 23
23	Construcción	23
56	Administrativo/apoyo y residuos	24
	Servicio de gestión/remediación	
61	Servicios educativos	24
51	Información	25
48–49	Transporte y almacenamiento	27, 23
52	Finanzas y Seguros	28
53	Inmobiliaria y alquiler y arrendamiento	29

4.1.3. Desembolso bruto Ei

desembolso bruto (representado como “Desembolso bruto” en el conjunto de datos) es otro indicador de riesgo que muchos estudiantes identifican como una variable clave a considerar. La razón detrás de seleccionar “Desembolso bruto” es que cuanto mayor sea el tamaño del préstamo, más probable será que se establezca y se expanda el negocio subyacente (es decir, comprar activos que tengan cierto valor de reventa), aumentando así la probabilidad de liquidar el préstamo. Este razonamiento se confirma al observar los cuartiles que se muestran en la [Tabla 4](#).

4.1.4. Negocios nuevos versus establecidos Si un

negocio es nuevo o establecido (representado como “NewEx ist” en el conjunto de datos) es otro indicador de riesgo potencial que los estudiantes identifican. Por lo tanto, se creó una variable dummy para la regresión logística: “Nuevo” D 1 si el negocio tiene menos o igual de 2 años y “Nuevo” D 0 si el negocio tiene más de 2 años.

La mayoría de los estudiantes argumentan que las empresas nuevas fracasan a un ritmo mayor que las empresas establecidas. Las empresas establecidas ya tienen un historial probado de éxito y están solicitando un préstamo para ampliar lo que ya hacen con éxito. Considerando que, las nuevas empresas a veces no anticipan los obstáculos que pueden enfrentar y es posible que no puedan superar dichos desafíos con éxito, lo que resulta en el incumplimiento de pago de un préstamo.

Sin embargo, cuando se comparan las tasas de incumplimiento para préstamos a empresas nuevas (menores o iguales a 2 años) y empresas establecidas (más de 2 años) en este conjunto de datos, existe una diferencia relativamente insignificante entre ellas. La tasa de morosidad para los negocios nuevos es del 18,98 % y la tasa para los negocios establecidos es del 17,36 %.

4.1.5. Préstamos respaldados por bienes

inmuebles Si un préstamo está respaldado por bienes inmuebles (posesión de terrenos) es otro indicador de riesgo que se analiza. La justificación de este indicador es que el valor de la tierra suele ser lo suficientemente grande como para cubrir el monto de cualquier principal pendiente, lo que reduce la probabilidad de incumplimiento.

Dado que el plazo del préstamo es una función de la vida útil esperada de los activos, los préstamos respaldados por bienes inmuebles tendrán plazos de 20 años o más (240 meses) y son los únicos préstamos otorgados a un plazo tan largo, mientras que los préstamos no respaldados por inmuebles tendrán plazos inferiores a 20 años (<240 meses). Por lo tanto, los autores crearon una variable ficticia, “Inmuebles”, donde “Inmuebles” D 1 si “Plazo” 240 meses y “Inmuebles” D 0 si “Plazo” <240 meses.

Como se muestra en [la Tabla 5](#), los préstamos respaldados por bienes raíces tienen una tasa de incumplimiento significativamente más baja (1,64%) que los préstamos sin respaldo inmobiliario (21,16%).

4.1.6. Recesión económica Un

indicador de riesgo que surge constantemente en la discusión es cómo la economía puede afectar las tasas de incumplimiento. Los préstamos para pequeñas empresas son

Tabla 5. Préstamos respaldados por bienes inmuebles.

	Por defecto	Pagado
Préstamos devueltos por Bienes Raíces (Plazo 240 meses)	2472 (1,64%)	147.868 (98,36%)
Préstamos no respaldados por bienes raíces (Plazo <240 meses)	153.876 (21,16%)	573.212 (78,84%)

afectados por la economía en general, y más préstamos para pequeñas empresas tienden a incumplir justo antes y durante una recesión económica. Por lo tanto, los autores crearon una variable ficticia, “Recesión”, donde “Recesión” D 1 si los préstamos estaban activos durante la Gran Recesión (diciembre de 2007 a junio de 2009), y “Recesión” D 0 para todos los demás momentos.

Como se ilustra en un gráfico de barras apiladas ([Figura 2](#)), los préstamos activos durante la Gran Recesión tienen una tasa de incumplimiento más alta (31,21 %) que los préstamos que no estaban activos durante la Recesión (16,63 %).

4.1.7. Porción garantizada de la SBA del préstamo aprobado La

porción que es el porcentaje del préstamo garantizado por la SBA (representado como “Porción” en el conjunto de datos) es un indicador de riesgo final que se analiza en nuestros cursos. Esta es una de las variables que los autores generaron al calcular la relación entre el monto de las garantías del préstamo SBA y el monto bruto aprobado por el banco (SBA_Appv/GrAppv). [La Figura 3](#) muestra la distribución de la porción de los préstamos pagados en su totalidad y los préstamos en mora desembolsados entre 2002 y 2010. Estos dos diagramas de caja muestran que, por lo general, los préstamos que se pagan en su totalidad tienen un porcentaje ligeramente mayor garantizado por la SBA, como lo indica la media más alta parte de los préstamos pagados en su totalidad.

Vale la pena señalar que la mediana no se muestra en los diagramas de caja para los préstamos en mora porque el 54% de estos préstamos tienen la mitad del monto del préstamo garantizado por la SBA (porción D 0.5). Como resultado, no hay diferencia en los percentiles 1%, 5%, 10%, 25% y 50% (todos estos percentiles son iguales a 0,5).

Nota didáctica: Además de las variables en el conjunto de datos, preguntamos a nuestros estudiantes si hay otras variables que puedan ser significativas y deban ser consideradas. Por lo general, los estudiantes no pueden encontrar ninguna fuente específica de variación. Sin embargo, cabe señalar que el conjunto de datos no incluye ningún elemento que represente directamente el riesgo de crédito. En los últimos años, la SBA ha recopilado y evaluado la calificación crediticia de los garantes y prestatarios de Fair Issac (FICO). Si un prestatario o garante no es una persona, se obtiene una puntuación de Dun and Bradstreet. Muchas instituciones financieras ahora confían en los puntajes de crédito cuando otorgan préstamos más pequeños. Desafortunadamente, este conjunto de datos no incluye esta información.

4.2. Paso 2: comprender el estudio de caso y el conjunto de datos

Después de identificar los indicadores de riesgo potencial, se presenta un estudio de caso en el que el estudiante asume el papel de un oficial de crédito que debe determinar si aprobar préstamos a dos pequeñas empresas. Destacamos el hecho de que los bancos intentan

Cuadro 4. Cuartiles de desembolso bruto.

Cuartiles	CHGOFF	FIP
100% máximo	\$4,362,157	\$11,446,325
75% cuartil	\$140,796	\$255,000
50% mediana	\$61,962.5	\$100,000
25% cuartil	\$27,767	\$49,034
Mínimo	\$4000	\$4000

⁶Los préstamos que se codificaron como “Recesión D1” incluyen aquellos que estuvieron activos durante al menos un mes durante el período de tiempo de la Gran Recesión. Esto se calculó sumando la duración del plazo del préstamo en días a la fecha de desembolso del préstamo. La codificación en SAS para esto es: RecessionD0; diasplazoDTérmino 30; xxDFechadesembolsoCdíasplazo; si xx ge ‘1DEC20070 d AND xx le ‘30JUN20090 d entonces RecesiónD1.

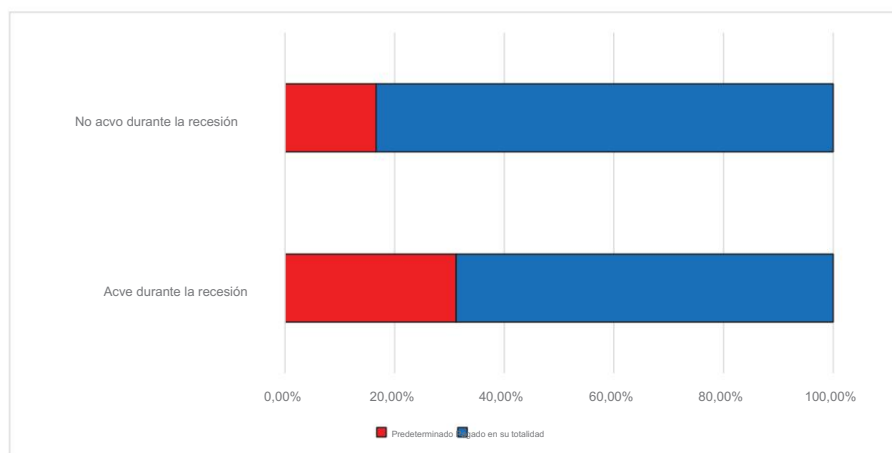


Figura 2. Estado de los préstamos activos o no activos durante la Gran Recesión.

minimizar el riesgo de incumplimiento (cancelación) y solo aprobar préstamos que probablemente se paguen en su totalidad más adelante.

Nota didáctica: para dar cuenta de dos de los indicadores de riesgo, estado e industria, restringimos el estudio de caso a un estado y una industria (código de industria de dos dígitos). Sugerimos que los educadores consideren hacer lo mismo por tres razones: (a) crea un escenario de toma de decisiones más realista; (b) la inclusión de 50 estados (más Washington DC) y 20 clasificaciones de industrias (NAICS de 2 dígitos) daría como resultado una gran cantidad de variables binarias y podría crear problemas de estimación; y (c) el conjunto de datos extraído del conjunto de datos más grande es más manejable para los estudiantes. Describimos este proceso y justificación a los estudiantes en clase.

Para nuestros cursos, hemos optado por limitar el estudio de caso al Estado de California y el código de dos dígitos 53: Bienes Raíces y Alquiler y Arrendamiento. Extraemos los datos relevantes del conjunto de datos más grande, "National SBA", que produce una muestra de 2102 observaciones y se incluye en el documento como los datos del "Caso SBA". Proporcionamos este conjunto de datos a los estudiantes para analizar

en sus roles como oficiales de crédito al decidir si aprobar o denegar dos solicitudes de préstamo.

Nota didáctica: restringimos el escenario para la asignación a California porque aquí es donde estamos ubicados. Los profesores pueden optar por centrarse en estados de interés para sus alumnos. Para el código de la industria, se puede usar cualquier código de dos dígitos o seleccionar un código que use más de dos dígitos.

Estudio de caso basado en California: Usted, un oficial de préstamos de Bank of America, recibió dos solicitudes de préstamo de dos pequeñas empresas: Carmichael Realty (una agencia de bienes raíces comerciales) y SV Consulting (una firma de consultoría de bienes raíces). La información relevante de la aplicación se resume a continuación (consulte la [Tabla 6](#)). Como oficial de préstamos, debe determinar si debe otorgar o denegar estas dos solicitudes de préstamo y proporcionar una explicación de "por qué o por qué no". Para tomar esta decisión, deberá evaluar el riesgo del préstamo calculando la probabilidad estimada de incumplimiento utilizando la regresión logística. A continuación, querrá clasificar este préstamo

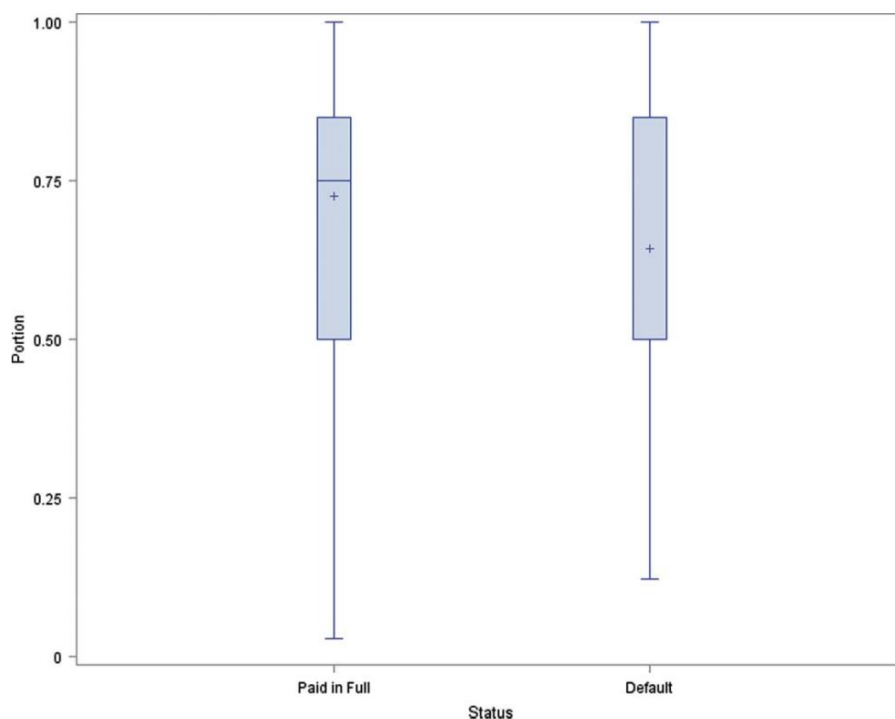


Figura 3. Porciones garantizadas por la SBA para préstamos pagados en su totalidad e incumplidos.

Tabla 6. Estudio de caso basado en California: información para dos solicitudes de préstamo.

Préstamo	Nombre	Ciudad	Fecha	monto del préstamo solicitado	Porción SBA garantizada	¿Asegurado por bienes raíces?
1 2	Bienes Raíces	Carmichael, CA	Actual (no recesión)	\$1,000,000	\$750,000	Sí
	Consultoría SV	San Leandro, California	Actual (no recesión)	\$100,000	\$40,000	No

como: "mayor riesgo: más probabilidades de incumplimiento" o "menor riesgo: más probabilidades de pagar en su totalidad" al tomar su decisión.

Nota didáctica: Pedimos a los estudiantes que proporcionen un resumen escrito de la decisión comercial en cuestión y las posibles limitaciones del conjunto de datos. Nos enfocamos específicamente en el marco de tiempo y el sesgo de selección como se discutió en la Sección 3.3.

4.3. Paso 3: Construir el modelo, elegir una regla de decisión y validar el modelo de regresión logística

Guiamos a nuestros estudiantes a través del proceso de construcción de un modelo de regresión logística para estimar la probabilidad de incumplimiento de las diversas solicitudes de préstamo. Para cumplir con el objetivo de aprendizaje, para comprender las etapas en la construcción y validación de modelos, guiamos a los estudiantes a través de un proceso iterativo de construcción de modelos de tres fases de especificación, estimación y evaluación y luego validamos el modelo.

Para construir el modelo de regresión logística para el estudio de caso basado en California, seleccionamos aleatoriamente la mitad de los datos para que fueran nuestros datos de "entrenamiento" (1051 de las 2102 observaciones originales). En el conjunto de datos "Caso SBA", la variable "Seleccionado" indica qué observaciones son los datos de "entrenamiento" y cuáles son los datos de "prueba" (1 D datos de entrenamiento que se usarán para construir el modelo, 0 D datos de prueba para validar el modelo).

Nota didáctica: hay varias técnicas de clasificación posibles que se pueden utilizar para modelar estos datos. Dado que nuestro curso de pregrado de estadística empresarial es un curso de servicio para las áreas funcionales de los negocios y un requisito previo para una serie de cursos como finanzas y marketing, los objetivos de aprendizaje de este curso están alineados con los objetivos generales de aprendizaje de nuestra universidad y los objetivos de otros cursos (que incluyen una comprensión de la regresión logística). Por lo tanto, en este artículo presentamos nuestra cobertura de la regresión logística básica para nuestros estudiantes universitarios de negocios. Los estudiantes en cursos de estadística más avanzados pueden explorar interacciones en regresión logística, covariables dependientes del tiempo, así como métodos de clasificación más avanzados.

4.3.1. Especificación y estimación del modelo

Cuando se trata de una respuesta binaria, como es el caso aquí, la regresión logística es una opción de modelo popular para describir la relación entre la respuesta binaria y las variables explicativas (predictores). Los modelos de regresión logística registran las probabilidades como una combinación lineal de variables explicativas (predictores):

$$\text{registro} = \frac{\text{PAG}}{1 + P} \quad D \text{ b0 C b1X1 C b2X2 CC bKXK:}$$

La probabilidad de interés P puede entonces obtenerse como

$$PD = \frac{eb0 \text{ C b1X1 C b2X2 CC bK XK}}{1 \text{ C eb0 C b1X1 C b2X2 CC bK XK}}$$
$$D = \frac{1}{1 \text{ C e}^{-i \delta b0 \text{ C b1X1 C b2X2 CC bK XK P}}}$$

donde b0 C b1 X1 C b2 X2 C $\phi\phi\phi$ C bKXK representa los coeficientes y las variables explicativas de la línea generalizada

estructura del modelo de regresión del oído. La probabilidad de interés P se puede predecir con los coeficientes estimados.

Al construir el modelo, les indicamos a los estudiantes que la variable dependiente es una variable binaria. En nuestro análisis, la variable dependiente binaria es "Predeterminada", que es una variable ficticia creada a partir de la variable "MIS_Status". El valor de "Predeterminado" D 1 si MIS_Status D CHGOFF y "Predeterminado" D 0 si MIS_Status D PIF. Por lo tanto, el modelo de regresión logística para este escenario predice la probabilidad de incumplimiento de pago de un préstamo.

Destacamos por qué se usa el modelo de regresión logística, en lugar de la regresión lineal ordinaria, discutiendo los supuestos de la regresión lineal ordinaria y la violación de algunos de estos supuestos si se hubiera aplicado la regresión lineal ordinaria a este conjunto de datos. Dado que aquí estamos tratando con un resultado dicotómico (es decir, predeterminado o no) en lugar de uno cuantitativo, la regresión de mínimos cuadrados ordinarios no es adecuada.

En su lugar, utilizamos la regresión logística para predecir las razones de probabilidades y las probabilidades.

Para las posibles variables explicativas, revisamos los resultados del Paso 1 donde se identifican siete variables como posibles indicadores de riesgo. Dado que la "ubicación (estado)" y la "industria" ya se tienen en cuenta al restringir los análisis a un estado y una industria, hay cinco variables que deben considerarse para su inclusión en el modelo como variables explicativas: Recesión económica ("Recesión "), Nuevo negocio ("Nuevo"), Préstamos respaldados por bienes inmuebles ("Bienes raíces"), Desembolso bruto ("Desembolso bruto") y Porción garantizada de la SBA del préstamo aprobado ("Porción").

Para ilustrar el proceso de construcción del modelo, guiamos a los estudiantes a través de dos versiones diferentes del modelo utilizando los datos de entrenamiento: (a) modelo inicial con cinco variables explicativas (Tabla 7(a)), incluida la prueba de razón de verosimilitud para el efecto parcial obtenido de un análisis Tipo III de las variables PROC GENMOD de SAS (Tabla 7(b)) (Tabla 8).⁷ ; y (b) modelo re-especificado con tres explicaciones Después de producir el modelo inicial, se produce una discusión sobre las variables significativas y los valores de p. Los estudiantes determinan que los indicadores de riesgo "Nuevo" y "Desembolso bruto" no son estadísticamente significativos y, por lo general, sugieren volver a especificar el modelo sin estas variables. Dado que el objetivo es la predicción, se utilizará el modelo final con las tres variables explicativas "Bienes raíces", "Porción" y "Recesión" para clasificar los préstamos en el caso de estudio utilizando las reglas de decisión descritas en la Sección 4.3.2 .

Vale la pena mencionar que: (a) los autores confirmaron con un empleado de la SBA con más de 30 años de experiencia que tiene sentido económico eliminar "Nuevo" y "Desembolso bruto" del modelo y (b) casi no hay diferencia en el tasas de clasificación errónea calculadas para los datos de prueba, con o sin las dos variables "Nuevo" y "Desembolso bruto". Mientras que la

⁷ El análisis de tipo III prueba la importancia de cada efecto parcial y la importancia de un efecto con todos los demás efectos del modelo.

Cuadro 7(a). Estudio de caso de California: modelo de regresión logística inicial con cinco variables explicativas.

Parámetro	DF	Estimar	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Interceptar	1	1,3537	0,3229	17,5729	<0,0001
Nuevo	1	0,0772	0,2101	0,1349	0,7134
Bienes raíces	1	2,0331	0,3636	31,2663	<0,0001
DesembolsoBruto	1	3,37E-7	3,52E-7	0,9173	0,3382
Parte	1	2,8298	0,5594	25,5909	<0,0001
Recesión	1	0,4971	0,2413	4,2441	0,0394

El modelo no se ajusta a los datos tan bien como cabría esperar, ofrece un rendimiento predictivo razonable que se ilustra en la sección de validación (4.3.3) donde se aplican los "datos de prueba" al modelo.

Como afirma George Box, "Esencialmente, todos los modelos son erróneos, pero algunos son útiles" (Box y Draper 1987, p. 424). Y, Seymour Geisser (1993) afirmó que un modelo es útil siempre que ofrezca un buen rendimiento predictivo.

Nota didáctica: La variable "Seleccionada" indica cuál de los casos son datos de entrenamiento versus datos de prueba (1 para entrenamiento y 0 para prueba). Esta muestra aleatoria se extrajo en SAS utilizando el procedimiento SURVEYSELECT: PROC SURVEYSELECT OUTALL OUT D dataca53 METHOD D SRS SAMPSIZE D 1051 SEED D 18467;

Además de la discusión sobre los valores de p, se describe cómo interpretar las estimaciones de los parámetros del modelo con un enfoque en las probabilidades de incumplimiento. Por ejemplo, dado que Bienes inmuebles es una variable ficticia, podemos interpretar ese coeficiente como: "Dada la misma parte respaldada por la SBA y las mismas consideraciones económicas (recesión o no), la razón de probabilidad estimada de incumplimiento (respaldada por bienes inmuebles versus no respaldada por por inmuebles) es e^{2.1282} D 0.12. Por lo tanto, las probabilidades de incumplimiento cuando están respaldadas por bienes inmuebles son solo del 12 % de las probabilidades de incumplimiento cuando no están respaldadas por bienes raíces. Por lo tanto, como se esperaba, existe un menor riesgo de incumplimiento cuando el préstamo está respaldado por bienes.

Consideramos otras variables explicativas e interacciones entre las variables ficticias "Bienes inmuebles" y "Recesión" y la variable explicativa continua "Porción". Si bien no surgieron variables explicativas significativas adicionales, hubo dos efectos de interacción significativos: "Porción inmobiliaria" y "Porción de recesión"; esto sugiere que "Portion" tuvo una influencia adicional si el préstamo involucraba bienes inmuebles o si se produjo durante la recesión. Dado que la interacción en la regresión logística es un concepto complejo de conceptualizar en estos cursos introductorios, hemos decidido no incluir una discusión de estos efectos de interacción en este artículo.

4.3.2. Elección de una regla de decisión A

continuación, se guía a los estudiantes a través del proceso de elección de una regla de decisión. Discutimos cómo la probabilidad estimada de

el incumplimiento de un préstamo en particular debe compararse con una probabilidad de corte al tomar una decisión, seguido de una discusión sobre cuál podría ser una probabilidad de corte apropiada. Los estudiantes a menudo sugieren 0.5 como el límite, una opción obvia para muchos porque es equivalente a las probabilidades (cargado frente a pagado en su totalidad) de 1.

Hacemos que los estudiantes calculen la tasa de clasificación errónea utilizando diferentes niveles de probabilidad de corte. Los resultados se muestran en la Figura 4.

El nivel de probabilidad de corte que resulta en la tasa de clasificación errónea más baja comienza alrededor de 0,5. La tasa de clasificación errónea comienza a aumentar alrededor de un nivel de probabilidad de corte de 0,6. Por lo tanto, un nivel de probabilidad de corte de 0,5 es una buena elección. Entonces se adoptan las siguientes reglas de decisión: (i) clasificar la solicitud de préstamo

en la categoría de menor riesgo y aprobar el préstamo cuando la probabilidad estimada de incumplimiento sea de 0,5, o (ii) clasificar la solicitud de préstamo en la categoría de

mayor riesgo y denegar el préstamo cuando se estime probabilidad de incumplimiento >0,5.

Nota didáctica: En la Sección 3.3, se discutió el potencial de sesgo de selección debido al período de tiempo utilizado en los análisis. Sin embargo, cabe señalar que aquí hay otra fuente importante de sesgo de selección. Existe una discrepancia crítica entre los datos utilizados para construir el modelo predictivo y los préstamos que se evaluarán utilizando el modelo. Presumiblemente, solo los préstamos que se percibían como de riesgo tolerablemente bajo fueron aprobados en primer lugar. Eso significa que todos los préstamos representados en los datos habrían sido percibidos como de riesgo "bajo" por alguien. Los que se consideran de mayor riesgo (y, por lo tanto, no fueron aprobados) no aparecen en los datos en absoluto. Por lo tanto, la tasa de incumplimiento de la muestra probablemente será más baja que la tasa de incumplimiento real de todas las solicitudes de préstamo que se presentaron en primer lugar.

4.3.3. Validación y clasificación errónea Validamos el

modelo final aplicándolo a la otra mitad de los datos (los datos de "prueba" que incluyen las 1051 observaciones restantes para el ejemplo basado en California) y medimos su rendimiento calculando la tasa de clasificación errónea. Para ello, los estudiantes utilizan el modelo de regresión logística final para generar la tasa de probabilidad estimada de incumplimiento para cada uno de los préstamos en el

Tabla 7(b). Análisis tipo III.

Fuente	DF	chi-cuadrado	Pr > ChiSq
Nuevo	1	0,14	0,7130
Bienes raíces	1	39,96	<0,0001
DesembolsoBruto	1	0,97	0,3258
Parte	1	27,41	<0,0001
Recesión	1	4,27	0,0389

Tabla 8. Estudio de caso basado en California: modelo reespecificado con tres variables explicativas.

Parámetro	Estimación	del DF	Error estándar	Wald	Chi-Cuadrado	Pr > ChiSq
Interceptar	1	1,3931	1	0,3216	18,7670	<0,0001
Bienes raíces	1	2,1282	1	0,3450	38,0529	<0,0001
Parte	1	2,9875	1	0,5041	0,5393	30,6898
Recesión				0,2412	4,3679	0,0366

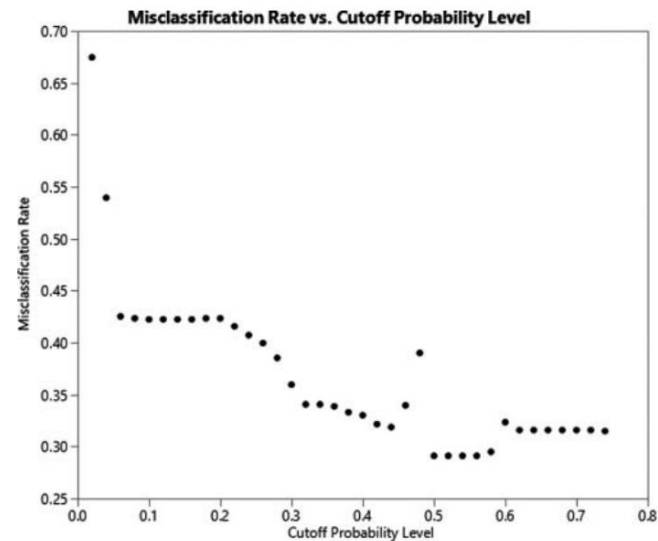


Figura 4. Tasa de clasificación errónea versus nivel de probabilidad de corte.

muestra de datos de "prueba". A continuación, se les pide a los estudiantes que clasifiquen los préstamos en los datos de prueba como "riesgo más alto" o "riesgo más bajo" utilizando las reglas de decisión de la Sección 4.3.2.

Dado que se conocen los verdaderos resultados de los préstamos en los datos de prueba (MIS_Estado de cancelado o pagado en su totalidad), se puede determinar la tasa de clasificación errónea para el escenario basado en California. En el Cuadro 9, las columnas representan la realidad de si un préstamo fue castigado o pagado en su totalidad, y las filas representan la clasificación del préstamo de acuerdo con la regla de decisión (mayor riesgo vs. menor riesgo). A continuación, donde el número de clasificaciones erróneas se representa en negrita, 324 préstamos se clasificaron erróneamente como de "menor riesgo" y 14 préstamos se clasificaron erróneamente como de "mayor riesgo". La tasa general de clasificación errónea es del 32,16 % ((324 C 14)/1051).

Nota didáctica: En clase, discutimos cómo este proceso es parte de la evaluación del desempeño predictivo de un modelo y que este modelo en particular brinda un desempeño predictivo razonable.

Nota didáctica: Dado que se pueden cometer dos tipos diferentes de errores, la clasificación errónea de un préstamo como de "riesgo mayor" o "riesgo menor", alentamos a los estudiantes a discutir las consecuencias de cometer cualquier tipo de error y si tratar los dos tipos de errores. lo mismo es una sabia decisión comercial. Las discusiones generalmente giran en torno al hecho de que el banco perderá capital e intereses si un préstamo se clasifica erróneamente como de "menor riesgo" y luego se cancela, mientras que el banco solo incurrirá en costos de oportunidad en el monto de los intereses si un préstamo se clasifica incorrectamente. clasificado como de "mayor riesgo".

Nota didáctica: En nuestro curso de posgrado, también cubrimos la curva ROC (característica operativa del receptor) para describir la precisión de clasificación por

Tabla 9. Escenario basado en California: Clasificación de préstamos.

Clasificación	Estado de naturaleza: Realidad		
	Préstamos castigados	Préstamos pagados en su totalidad	Total
Mayor riesgo (más probabilidades de ser castigado)	31	14	45
Menor riesgo (más probable que se pague en su totalidad)	324	682	1006
Total	355	696	1051

hacer que los estudiantes vean un video tutorial corto en <http://www.data school.io/roc-curves-and-auc-explained/>. La curva ROC traza la tasa de verdaderos positivos (en el eje y) frente a la tasa de falsos positivos (en el eje x) para cada posible nivel de probabilidad de corte de clasificación, mientras que la tasa de clasificación errónea es solo para un nivel de probabilidad de corte. El área bajo la curva (AUC) es la proporción de la caja (el área de esta caja es 1) debajo de esta curva ROC. El AUC es más alto (más de 0,75) para el modelo más parsimonioso con tres variables explicativas en la Tabla 8, lo que indica un rendimiento de clasificación aceptable (ver Hosmer y Lemeshow 2000, p. 162).

4.4. Paso 4: Uso del modelo para tomar decisiones

Para cumplir con los objetivos de aprendizaje, aprender a aplicar la regresión logística para clasificar un préstamo en función de la probabilidad de incumplimiento y experimentar el proceso de investigación de tomar una decisión basada en escenarios informada por los análisis de datos, el paso final en esta tarea es hacer que los estudiantes responder a la pregunta inicial de aprobar o denegar uno o más préstamos utilizando: (a) el modelo de regresión logística final generado para determinar la probabilidad estimada de incumplimiento de un préstamo específico y (b) las reglas de decisión para clasificar el préstamo. Para el ejemplo basado en California, el modelo final con los indicadores de riesgo en la Tabla 8 se usa para estimar la probabilidad de incumplimiento para las dos solicitudes de préstamo; la probabilidad estimada de incumplimiento para Carmichael Realty (Préstamo 1) es 0,05 y SV Consulting (Préstamo 2) es 0,55. Aplicando las reglas de decisión y la probabilidad de corte de 0,5 de la Sección 4.3, el Préstamo 1 se clasifica como de "menor riesgo" y debe aprobarse, y el Préstamo 2 se clasifica como de "mayor riesgo" y debe denegarse (consulte la Tabla 10).

5. Evaluación del aprendizaje, exploración de métodos de clasificación más avanzados y observaciones finales

5.1. Evaluación del aprendizaje

Anteriormente mencionado, evaluamos el aprendizaje de los estudiantes mediante el desarrollo de un estudio de caso similar al presentado en clase y asignamos esto a los estudiantes para una calificación de letra. Para los estudiantes universitarios, les dejamos completar la tarea calificada en grupos de tres personas. Para los cursos de posgrado, los estudiantes deben completar la tarea como individuos.

Para las tareas calificadas, los estudiantes deben enviar un informe que explique todos los pasos que realizaron (que deben reflejar los pasos descritos anteriormente) y una recomendación final sobre si los préstamos deben aprobarse o denegarse. Sugerimos que el informe tenga tres páginas más cualquier tabla, figura y gráfico que ayude a ilustrar y respaldar su recomendación. Permitimos a los estudiantes dos semanas para completar la tarea después de las sesiones en clase. Al evaluar su aprendizaje, usamos la rúbrica de calificación que se muestra en la Tabla 11.

5.2. Métodos de clasificación avanzada para graduados Estudiantes

Si bien nos enfocamos en la regresión logística en "¿Debería aprobarse o denegarse este préstamo?" asignación, otros métodos de clasificación como redes neuronales (ver Odom y Sharda 1990; Tam y Kiang 1992; Lacher et al. 1995; Zhang et al. 1999) y SVM (ver Chen et al. 2010; Kim y Sohn 2010) podrían ser útiles .



Tabla 10. Resumen del escenario basado en California.

Préstamo	Nombre	Fecha	Monto del préstamo solicitado	Porción SBA garantizada	¿Asegurado por bienes raíces?	Probabilidad estimada de incumplimiento	¿Aprobar?
1 2	Bienes Raíces	Actual (sin recesión)	\$1,000,000	\$750,000	Si	0,05	Si
	Consultoría SV	Actual (sin recesión)	\$100,000	\$40,000	No	0,55	No

enseñado utilizando este conjunto de datos en cursos de análisis de datos de posgrado más avanzados.

En nuestro curso de minería de datos para graduados, enfatizamos a los estudiantes que existen suposiciones estrictas para los modelos paramétricos tradicionales, como la regresión logística. Cuando estas suposiciones no se cumplen, los métodos de clasificación no paramétricos no lineales, como las redes neuronales y las SVM, son alternativas poderosas. Las redes neuronales (feed-forward) son modelos de regresión no lineal flexibles con muchos parámetros, que conectan entradas (variables explicativas o predictores) a salidas (la variable dependiente) a través de capas ocultas entre entradas y salidas. La “función de activación” de las unidades de la capa oculta suele ser la función logística (ver Venables y Ripley 2002, sec. 8.10). La regresión logística es equivalente a la red neuronal sin nodo oculto (Zhang et al. 1999), y es natural comparar los resultados de la red neuronal con los de la regresión logística. Si el objetivo de aprendizaje de una tarea es separar los préstamos de los préstamos que probablemente incumplan sin necesidad de la probabilidad prevista de incumplimiento, entonces las redes neuronales y las SVM son buenas opciones.

Nota didáctica: comenzamos nuestra introducción a las redes neuronales demostrando cómo aplicar la función de redes neuronales "nnet" en R para entrenar y probar los datos que los estudiantes habían usado antes para la regresión logística. Nuestros estudiantes probaron dos configuraciones de redes neuronales: sin capa oculta y una capa oculta de 5 unidades. Luego pasamos a discutir algunos aspectos teóricos de las redes neuronales.

Nota didáctica: Venables y Ripley (2002) brindan una breve introducción muy fácil de leer con instrucciones claras para usar el paquete de redes neuronales "nnet" en R. También les pedimos a los estudiantes que revisen el paquete actualizado

documentación para "nnet" con ejemplos en <http://cran.r-project.org/web/packages/nnet/nnet.pdf>. Nuestros estudiantes graduados pueden lograr fácilmente el ajuste de un modelo de red neuronal de este tipo con unas pocas líneas de código en R.

```
Con la misma tarea descrita anteriormente, los estudiantes de posgrado pudieron
ajustar fácilmente el modelo de redes neuronales con las mismas variables explicativas y
datos de entrenamiento usando R y obtener una tasa de clasificación errónea ligeramente
menor del 31,97 % ((324 C 12)/1051) para la prueba. datos. El código R es: #Neural
Networks data <- read.csv(archivo D "C:/SBACase.csv",
encabezado D TRUE,
sep D ",", resumen(datos) adjunto(datos) x1 D RealEstate x2 D ( Porción-
media(Porción))/
raiz
cuadrada(var(Porción))
x3 D Recesión y D as.factor(MIS_Status) dat D data.frame(x1,x2,x3,y)
biblioteca(nnet)
entrenar D (Seleccionado>0 ) nnfit
D nnet(y >., datos D dat[tren,], tamaño
D 5, saltar D
```

```
TRUE, rang D 0.02, decay D 1e-3, maxit D 10000) summary(nnfit) test
D dat[ttrain,]
model_pred D
predict(nnfit, test, type D "class") table(model_pred, test$y)
```

Otro método de clasificación popular para este problema de clasificación binaria (pagado en su totalidad frente a cancelado) son las SVM. SVM es una extensión del clasificador de vectores de soporte, que está estrechamente

Tabla 11. Rúbrica de calificación para la tarea.

Paso (Peso)	no cumple con las expectativas	Se acerca a las expectativas	Cumple con las expectativas	Supera las expectativas
1) Identificación de indicadores de riesgo potencial (30%)	Identifica indicadores de riesgo potencial que no tienen sentido	Identifica indicadores de riesgo potencial, pero no proporciona una justificación razonable	Identifica indicadores de riesgo potencial proporciona una justificación razonable o por qué no como por qué o por qué no se deben considerar las variables y proporciona evidencia de respaldo (análisis)	Identifica indicadores de riesgo potencial y proporciona una justificación razonable como por qué
2) Entender el caso de estudio (10%)	Entiende el caso de estudio, proporciona una sinopsis inexacta y/o una sinopsis confusa de la decisión comercial en cuestión, de la decisión comercial en cuestión, pero no incluye una discusión sobre las limitaciones del conjunto de datos.	Comprende el estudio de caso, pero no crea una regla o regla de decisión apropiada, pero no valida adecuadamente el modelo de decisión apropiada.	Comprende el estudio de caso proporcionando una sinopsis de la decisión comercial en cuestión e incluye una discusión de limitaciones relacionadas con el marco de tiempo o el sesgo de selección	Comprende el estudio de caso proporcionando una sinopsis de la decisión comercial en cuestión e incluye una discusión de las limitaciones relacionadas con el marco de tiempo y el sesgo de selección.
3) Construcción del modelo, creación de reglas de decisión y validación del modelo de regresión logística (50%)	Construye un modelo que no tiene sentido	Construye un modelo que hace	Construye un modelo que tiene sentido, pero no valida adecuadamente el	Construye un modelo que tiene sentido, crea una regla de decisión apropiada y valida adecuadamente el modelo
4) Usar el modelo para tomar decisiones (10%)	Deriva un error probabilidad estimada de incumplimiento para ambos préstamos y toma malas decisiones para ambos préstamos, utilizando su modelo	Deriva la estimación correcta probabilidad de incumplimiento para uno o ambos préstamos, pero toma malas decisiones para ambos préstamos, utilizando su modelo	Deriva la estimación correcta probabilidad de incumplimiento para ambos préstamos y toma una buena decisión para un préstamo (pero no para el otro), utilizando su modelo	Deriva la estimación correcta probabilidad de incumplimiento para ambos préstamos y toma una buena decisión para ambos préstamos, utilizando su modelo



relacionado con la regresión logística (ver James et al. 2013, cap. 9.5). Por lo tanto, es natural pedir a los estudiantes que comparen SVM con la regresión logística. En clase, los estudiantes pueden ajustar fácilmente el SVM a los datos usando la función "SVM" en la biblioteca R e1071. Se encontró que la clasificación errónea era más alta que la de la regresión logística o las redes neuronales.

5.3. Observaciones finales

En conclusión, este rico conjunto de datos brinda a los educadores la oportunidad de crear tareas significativas para enseñar una variedad de conceptos estadísticos y resaltar cómo se pueden usar los datos para informar decisiones comerciales reales. En línea con las recomendaciones de GAISE de 2016, "¿Debe aprobarse o denegarse este préstamo?" La asignación de un estudio de caso es un excelente ejemplo de cómo promover el aprendizaje activo y enseñar el pensamiento estadístico en un contexto empresarial utilizando datos reales.

Alentamos a otros a pensar en formas creativas de incorporar los datos en asignaciones alternativas. Esperamos que los instructores compartan sus tareas con la comunidad de educación estadística para mejorar la eficacia de la enseñanza en todo el campo de la estadística.

Material suplementario

Se puede acceder a la información complementaria de este artículo en el [sitio web del editor](#). Esto incluye los archivos de datos "National SBA" y "SBA Case" y su documentación correspondiente.

Referencias

Bryant, PG (1999), "Discusión, debate y desacuerdo: Enseñanza de la regresión múltiple mediante discusión de casos", Boletín del Instituto Internacional de Estadística, Actas del Instituto Internacional de Estadística, vol. 58, Libro 2, Helsinki: Instituto Internacional, págs. 215–218.

- Box, GEP y Draper, NR (1987), Construcción de modelos empíricos y superficies de respuesta, Nueva York: Wiley, pág. 424.
- Chen, S., H€ardle, WK y Moro, RA (2010), "Modelización del riesgo de incumplimiento con máquinas de vectores de soporte", Quantitative Finance, 11, 135–154.
- Informe universitario de GAISE Comité de revisión de ASA (2016), "Pautas para la evaluación e instrucción en el informe universitario de educación estadística 2016", disponible en <http://www.amstat.org/education/gaise>.
- Geisser, S. (1993), Predictive Inference: An Introduction, Nueva York: Chapman & Hall.
- Hosmer, DW y Lemeshow, S. (2000), Applied Logistic Regression (2.ª ed.), Nueva York: Wiley.
- James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013), Introducción to Statistical Learning, Nueva York: Springer.
- Kim, HS y Sohn, SY (2010), "Máquinas de vectores de soporte para la predicción de incumplimiento de las pymes basadas en el crédito tecnológico", European Journal of Operational Research, 201, 838–846.
- Lacher, RC, Coats, PK, Sharma, SC y Fant, LF (1995), "Una red neuronal para clasificar la salud financiera de una empresa", European Journal of Operational Research, 85, 53–65.
- Nolan, D. y Speed, TP (1999), "Teaching Statistics Theory Through Applications", The American Statistician, 53, 370–375.
- Odom, MD y Sharda R. (1990), "Un modelo de red neuronal para la predicción de bancarota", Actas de la Conferencia internacional IEEE sobre redes neuronales, II, 163–168.
- Parr, WC y Smith, MA (1998), "Desarrollo de cursos de estadísticas comerciales basados en casos", The American Statistician, 52, 330–337.
- Smith, M. y Bryant, P. (2009), "Gestión de discusiones de casos en clases introductorias de estadística empresarial: enfoques prácticos para instructores", The American Statistician, 63, 348–355.
- Tam, KY y Kiang, MY (1992), "Aplicaciones gerenciales de redes neuronales: el caso de las predicciones de fallas bancarias", Management Science, 38, 926–947.
- Administración de Pequeñas Empresas de EE. UU. (2015), Historia recuperada el 22 de agosto de 2015 de <https://www.sba.gov/about-sba/what-we-do/history>.
- Venables, WN y Ripley, BD (2002), Modern Applied Statistics with S (4.ª ed.), Nueva York: Springer.
- Zhang, G., Hu, MY, Patuwo, BE e Indro, DC (1999), "Redes neuronales artificiales en la predicción de quiebras: marco general y análisis de validación cruzada", European Journal of Operational Research, 116, 16–32.