Modelo de Regresión logística para la predicción de Bancarrota en Bancos Norte Americanos

Julieth Stefanny Escobar Ramírez, Sara Gallego Villada, Isabella Navarro Múnera y Luisa García Salazar
Asesor: Henry Laniado Rojas
Departamento de ciencias matemáticas,Ingeniería Matemática, Universidad EAFIT

15 de mayo de 2022

Resumen

En el presente trabajo decidimos implementar un modelo de Regresión logística en python. Implementando la función sigmoide, se busco predecir el fenómeno de quiebra en los bancos norteamericanos. La información que usaron fue extraída de los estados financieros de bancos aleatoriamente seleccionados, motivado por la situación económica norteamericana en el 2008, la cual provocó problemas económicos internos y externos, incrementando la desconfianza de los consumidores, inversionistas y gobiernos extranjeros. Se desea probar si el modelo tiene utilidad y su eficiencia para que los inversionistass colombianos tengan más seguridad.

Abstract

In this work we decided to implement a logistic regression model in Python, implementing the sigmoid function. By using this method, we sought to predict the phenomenon of bankruptcy of bank failures in North American banks. The information used was extracted from the financial statements of randomly selected banks, motivated by the situation of the American economy in 2008 causing internal and external economic problems, increasing the distrust of consumers, investors and foreign governments. It is desired to test the usefulness and efficiency of the model for Colombian investors to have more security while investing.

Palabras clave: Bancarrota, Modelos de predicción regresión, razones financieras.

1. Introducción

La economía Colombiana es dependiente en gran parte de la Estadounidense, por lo cual las inversones de muchos Colombianos se verían perjudicadas si alguno de los bancos estadounidenses cae en bancarrota. Por lo cual, con el objetivo de contrribuir a la economía colombiana los inversores buscan diversas maneras para evitar perdidas, especialmente mediante modelos probabilísticos.

Este proyecto busca aportar a que los inversores tengan más tranquilidad a la hora de invertir en bancos estadounidenses, mediante la predicción de bancarrota de bancos, para esto se realiza un modelo de clasificación, según la regresión logística, la cual nos permite etiquetar los bancos como quebrados y no quebrados.

El modelo fue entrenado y validado con los datos, estaban variables como liquidez, productividad, rentabilidad, entre otras, para determinar si el modelo está correcto.

Se espera que el proyecto sea continuado y mejorado para usarlo en otros sectores financeros y compañías.

1.1. Problemática

Norteamérica es uno de los países más poderosos en el mundo, teniendo una de las mejores economías mundiales. Cuando sucedio la crisis del 2008 no sólo afecto la economía estaunidense sino a nivel internacional, pues muchas empresas quebraron generando muchos problemas, uno de ellos siendo que los inversores de los bancos que se quebraron perdieran su dinero.

1.2. Objetivos

1.2.1. Objetivo general

Ayudar a determinar la probabilidad de bancarrota en los bancos estaunidences, previniendo a los inversores Colombianos y ayudandolos a reducir sus probabilidades de perdida de inversión. Utilizando un modelo de regresión logística, deseamos predecir el fenómeno de bancarrota en los bancos norteamericanos.

1.2.2. Objetivos específicos

- Definir y conocer claramente los conceptos financieros y de programación para poder hacer el modelo probailístico de manera eficaz.
- Entender el funcionamiento de otros modelos que se relacionan al modelo probabilístico para poder mejorar/perfeccionar la eficiencia del modelo a la hora de aplicarlo.
- Tener una investigación clara para poder utilizar modelos, ejemplos y conceptos competentes que nos ayuden a tener resultados óptimos.

2. Justificación

El motivo de esta investigación y aplicación del modelo se debe a que decidimos analizar la bancarrota en los bancos estadounidenses de 2007-2017, ya que el mercado colombiano está directamente relacionado con el americano,y tenemos una grand dependencia de este, cerca del 26 % de exportaciones de Colombia son en dicho país, por lo cual el modelo podría aplicarse en otros campos, no solo en bancarrota de bancos sino en activos, bancarrota de empresas e incluso portafolios. Con ellos decidimos usar el modelo de regrsión logística, el busca predecir la bancarrota estadounidense y que está información sea de utilidad para los colombianos exportadores e inversionistas.

2.1. Antecedentes

La invesigación hecha Dr. Jesús Fernando Isaac García y Dr. Oscar Flores Colbia, "MODELO PROBABILISTICO DE BANCARROTA PARA BANCOS NORTEAMERICANOS ANTE LA RECESION NO RECONOCIDA DEL 2008. UNA HERRAMIENTA PARA LA TOMA DE DECISIONES", fue una motivación para utilizar este modelo, el cual consistia en un modelo de logit para predecir bancarrota pero segun la situcación vivida en la gran depresión. Es decir, el proyecto se basa en esa investigación para profundizar en ella y analizar el modelo de regresión logística adaptado a una situación específica que es nuestro objeto de estudio, el cuál es como la economía colombiana está directamente relacionada con la estadounidense, y como ambas se ven afectadas después de la crisis del 2008 (específicamente inversiones). Además, ayudandonos del artículo realizado por Jorge Ivan Perez G., Karen Lorena Gonzales C. y Mauricio Lopera C., dónde se enuncian dos modelos probabilísticos (logit y probit), con el fin de calcular el riesgo de quiebra en una empresa. Estos estudiantes de economía construyeron indicadores financieros, a partir de los estados financieros reportados en la superintendencia de sociedades, los cuales son:

- Rentabilidad del activo
- Rotación del activo
- Capacidad de endeudamiento

En la sección 2.3 mencionamos máa modelos de regresión, los cuales usamos para entender la aplicabilidad en la probabilidad.

2.2. Conceptos

Los conceptos que se definen a continuación, son parte de los indicadores que usamos para el modelo, estos fueron los elegidos porque representan parte importante de un banco o una empresa constituida, y si tienen cambios drasticos podria afectar las empresas

Liquidez.
Es la velocidad con la que una activo se puede vender, entre más liquidez tenga un activo, menos arriesgamos al venderlo. En una empresa, la liquidez es la capacidad para cumplir con los compromisos a corto y mediano plazo.

Rentabilidad.
Es la capacidad que tiene una inversión de generar una utilidad.

Calidad de crédito.
Es la capacidad de una empresa poder cumplir con sus compromisos y pagos al momento de adquirir una deuda.

Calidad de crédito.
Es sacar el máximo provecho de los recursos, una empresa es eficiente si máximiza sus ganancias y minimiza los costos[2].

Solvencia.
Es la capacidad que tiene una empresa de cumplir con todos sus compromisos de pago, independiente si es de forma inmediata o en un momento posterior.

Matriz de confusión

En el campo de la inteligencia artificial y el aprendizaje automático, una matriz de confusión es una herramienta que permite visualizar el desempeño de un algorítmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. En términos prácticos, nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje con los datos.[7]

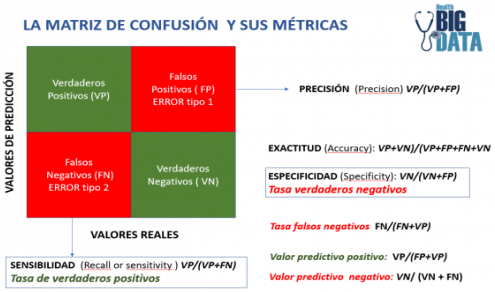

Figura 1 Matriz de confusión

2.2.1. Regresión Logística binaria

La regresión logística binaria es un método estadístico, el cual se utiliza cuando se desea conocer la relación entre una variable dependiente cualitativa y una o más variables independientes o explicativas, que pueden ser cualitativas y/o cuantitativas, con el objetivo de obtener una estimación ajustada de la probabilidad de ocurrencia de un evento a partir de una o más variables independientes.[7]
En nuestro modelo, podriamos decir que esta regresión se hace como un modelo de caja negra en el que el computador nos devukve unos resultados, sin embargo a continuación explicamos como se ve Matemáticamente.
Sea $\boldsymbol{y}$ la variable independeinte, esta variable es categórica y sirve como variable predictoria que tendrá como posibles valores 0 y 1, es decir, es una variable binaria. Si introducimos un valor a la función sigmoide esta nos dara valores entre 0 y 1.
Sean $x_1, x_2, ..., x_i$ las variables independientes que van a representar los indicadores del banco que se describen en las variables (Sección 3.2).
En general las x_i representan las posibles condiciones que puedan incidir en la variable dependiente. $\boldsymbol{y}$.
También se tienen unos parámetros del modelo, los cuales son $w_1, w_2, ..., w_i$.
La variable categórica $\boldsymbol{y}$ se obtiene al hacer combinaciones lineales entre los parámetros del modelo y las características x_i. Para la construcción de dichas combinaciones lineales se tendrán dos vectores:
1. El vector con las características de los bancos:

$$\mathrm{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \tag{1}$$

2. El vector transpuesto al vector de los parámetros, para hacer posible la combinación:

$$w^t = \begin{bmatrix} w_0 & w_1 & \dots & w_n \end{bmatrix} \tag{2}$$

se necesita evaluar la logística estándar acumulada con la combinación lineal entre los parámetros del modelo y las variables caracterısticas , de la siguiente manera[5]:

$$F(w^t x) = 1 - \frac{1}{1 + e^{w^t x}} \tag{3}$$

Luego las probabilidades del modelo estan dadas por:

$$p = \frac{e^{w^t x}}{1 + e^{w^t x}} \tag{4}$$

Como se mencionó anteriormente, la regresión logística está entre 0 y 1 dado por la función sigmoide.

Función sigmoide.

Es una función matemática que tiene una curva característica en forma de "S", que transforma los valores entre el rango 0 y 1. La función sigmoide también se llama curva sigmoidea o función logística. Es una de las funciones de activación no lineal más utilizadas[3].Esta función está definida por la fórmula:

$$S(t) = \frac{1}{1 + e^{-t}} \tag{5}$$

La combinación lineal en la función sigmoide resultante es:

$$S(w^t x) = \frac{1}{1 + e^{-w^t x}} \tag{6}$$

La gráfica muestra lo descrito anteriormente:

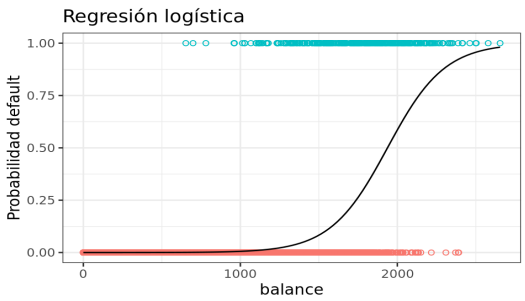

Figura 2 Regresión logística

Ahora, esta permitirá que la probabilidad esté correctamente calibrada y que arroje la probabilidad de que el modelo este entre esos valores, es decir que:

$$0 \text{ si y solo sí } (6) \prec 0{,}5$$

$$1 \text{ si y solo sí } (6) \geq 0{,}5$$

2.3. Estudios previos

Modelos desarrollados usando regresión logística

Un primer estudio corresponde a Marıa Alejandra Velez Clavijo, Valentina Moreno Ramırez, Alejandra Palacio Jaramillo y Juan Jose Wilches Riva quienes realizaron "Regresion logística robusta para la clasificacion de residuos solidos". En este trabajo se busca contribuír a la protección y preservación del entorno, se pretende clasificar los residuos sólidos en orgánicos e inorgánicos mediante la implementación de un modelo de regresión logística, y su versión robusta.

Inicialmente se tuvieron 806 imágenes de residuos sólidos de las cuales el 80 % van a ser destinadas al conjunto de entrenamiento y 20 % al conjunto de datos de validacion.
Finalmente, se encontró que de acuerdo con los resultados obtenidos las regresiones logísticas calculadas con los coeficientesde Kendall, Pearson y Spearman son menos sensibles a datos atıpicos respecto a la regresion logıstica clásica. También se encontró que si bien la regresión logística clásica funciona correctamente, al ser más sencible a datos irregulares se pueden obtener conclusiones erroneas o poco precisas, mientras usando la robusta se pueden conseguir mejores resultados. Este trabajo se relaciona con el trabajo en curso puesto que se propone la regresión logística como modelo para clasificación de elementos.

Un segundo estudio corresponde a Mª Visitación García Jiménez, Jesús Mª Alvarado Izquierdo y Amelia Jiménez Blanco de la Universidad Complutense de Madrid quienes realizaron "La predicción del rendimiento académico: regresión lineal versus regresión logística". En este estudio se pretende evaluar la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito/fracaso académico partiendo de variables, como la asistencia y la participación en clase.

La muestra estaba constituida por 175 estudiantes (140 mujeres y 35 hombres) de primer curso de Psicología de la UCM. Los datos fueron tomados durante el curso académico 1997/98.
Finalmente, se encontró que el procedimiento de regresión múltiple no permitió hacer un buen pronóstico del rendimiento académico, mientras que la regresión logística si parece ser un instrumento idóneo para hacer una buena predicción del éxito/fracaso académico.

Como tercer estudio tenemos "Metodología de evaluación del clima organizacional a través de un modelo de regresión logística para una universidad en Bogotá, Colombia".El cual corresponde a Juan Camilo Vega, Edgar Guillermo Rodríguez Díaz y Alexandra Montoya R. El estudio presenta el desarrollo de una metodología de evaluación de clima organizacional entre diferentes grupos de interés dentro de una organización académica, mediante la formulación de un modelo de regresión logística.

El estudio se realizó con el personal administrativo y académico que labora en las diferentes unidades misionales y unidades de gestión en las dos sedes (Central y Sede), ubicadas en la ciudad de Bogotá, Colombia, vinculado durante el primer periodo de 2010 a la universidad. Se entregaron 289 cuestionarios para ser diligenciados de los cuales 169 fueron resueltos. Se concluyó que al utilizar un modelo de regresión logística se complementa la caracterización del clima organizacional y permite evidenciar diferencias significativas con respecto al ambiente laboral en los distintos grupos de interés dentro de la organización, lo cual facilita la toma de decisiones eficientes, la aplicación de un modelo de regresión logística, demuestra que esta herramienta complementa y enriquece el análisis de resultados con la ventaja de permitir comparaciones de carácter dicotómico.

3. Metodología

Se utilizó un modelo de regresión logística binaria, F1 score y matrices de confusión, en los resultados de la regresión 1 es bancarrota y 0 no

bancarrota, además que para transformar el resultado obtenido por la regresión lineal, se utiliza la función sigmoide.

3.1. Muestra

Se tomaron 120 bancos sanos y 71 bancos quebrados desde 2007 hasta 2017, esto fue de manera aleatoria, siendo estos bancos solo norteamericanos. La información fue tomada Kaggle y utilizaremos herrameintas descritas durante la sección.

3.2. Variables

Las x_i serían: Tobin's Q, EPS, Liquidez, rentabilidad, productividad, razon de apalancamiento, rotación de activos, margen operativo, Rentabilidad sobre recursos propios, Market Book Ratio, Crecimiento de activos Crecimiento de ventas, crecimeinto empleados, y nuestra variable dependeinte y Bancarrota.

3.3. Hipotésis de la investigación

Para nuestro modelo decidimos usar el leguaje python, utilizando sus librerias Pandas, Numpy, sklearn, matplotlib, seaborn y imblearn. Utilizando estas librerias construimos el modelo y generamos gráficas para el análisis del mismo, en el cual se espera encontrar según las variables previamente mencionadas y con la evaluación del modelo, su utilidad.

3.4. Modelo desarrollado

El modelo desarrollado es le modelo de regresión logística descrito en la sección 2.2.1, matemáticamente. Tomamos las librerias descritas en nuestra hipótesis de investigación y las variables mencionadas, las cuales sirven para identificar categorías o clases a las que pertenecen las observaciones.[9]

4. Conclusiones

Con los datos obtenidos mostrados en la siguiente gráfica, obtenemos la precisión del algoritmo. Obtuvimos pocos falsos negativos (el valor real es negativo y la prueba predijo que era positivo), y pocos falsos positivos (el valor real es negativo y la prueba predijo que era positivo), lo que quiere decir que el algoritmo tuvo una tasa de fallo baja. Se obtuvo un f1-score mayor a 0.8 lo que significa que el algoritmo tiene buena precisión.

	precision	recall	f1-score	support
0	0.92	0.96	0.94	25
1	0.92	0.86	0.89	14
accuracy			0.92	39
macro avg	0.92	0.91	0.92	39
weighted avg	0.92	0.92	0.92	39

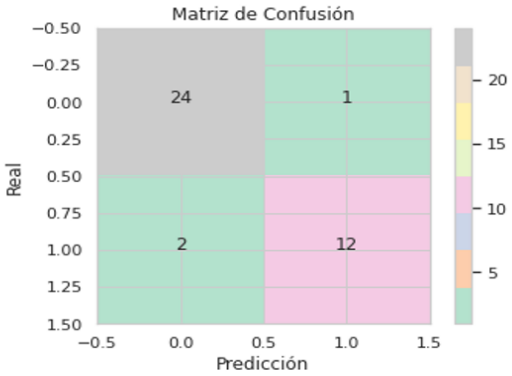

Figura 3 Resultados

4.1. Evaluación del modelo

4.2. Prueba del modelo

4.3. Futuras investigaciones

Después de relaizado el modelo encontramos que, si bien es cierto que las politicas de estados unidos frente a la bancarrota, buscan prevenir esta, y si llega a suceder tratar de retroceder esta situación, tambien se quizo evalauar este modelo en condiciones menos extremas, para ver su fiablidad y como se podría usar en otros campos, ya sea prevenir la bancarrota en empresas, o realizar mejores inversiones, por ello, gracias al machine learning y la probabilidad este proyecto tiene la posibilidad de mejroarse y en mejores escenarios.

5. Referencias

[1] Isaac, J.F., Flores, O."Modelo probabilístico para bancos norteamericanos ante la recesión no reconocida del 2008. Una herramienta para la toma de decisiones, Contribuciones a la Economía". P, UAT, México 2010

[5]Clavijo, Maria A V. 2022. "Regresión Logística Robusta Para La Clasificación de Residuos Sólidos." OSF. April 1. doi:10.17605/OSF.IO/CW6UP.

[6]*et al.* "Modelo de regresión logística para estimar la dependencia según la escala de Lawton y Brody.".Septiembre (2010). Vol. 36. Núm. 7

[**7**] Barrios. I."La matriz de confusión y sus métricas.[en] 2019 URL: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[**8**] Rodrigo 2016.Regresión logística simple y múltiple.[el] URL : https://www.cienciadedatos.net/documentos/27$_{r}egresion_{l}ogistica_{s}imple_{y}_{m}ultiple$.

[**9**]$Programmingfoundation.I$."$Module4Logisticregression$". URL : $https$: $//learn.theprogrammingfoundation.org/gett ing_{s}tarted/intro_{d}ata_{s}cience/module4/$