Evaluación de modelo de regresión logística para la predicción de bancarrota en bancos norte americanos

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

12 de agosto de 2022

Resumen

En el presente trabajo se implementó un modelo de Regresión logística en python. Utilizando la función sigmoide, se buscó predecir el fenómeno de quiebra en los bancos norteamericanos. La información usada 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 la utilidad y eficiencia del modelo para que los inversionistass colombianos tengan más seguridad.

Abstract

In the present work a logistic regression model is implemented in Python. Using the sigmoid function, we sought to predict the phenomenon of bankruptcy 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 logística, razones financieras.

1. Introducción

La economía colombiana es dependiente en gran parte de la estadounidense, por lo cual las inversiones de muchos colombianos se verían perjudicadas si alguno de los bancos estadounidenses cae en bancarrota. Por esto, el objetivo de este trabajo es contribuir a la economía colombiana desarrollando un modelo de regresión logística que sirva a los inversores de manera que eviten pérdidas.

En este trabajo se desarrolla un modelo de clasificación, que permite etiquetar los bancos norteamericanos como "quebradosz "no quebrados". Este modelo fué entrenado y validado con variables financieras como liquidez, productividad, rentabilidad, crecimiento en ventas, índice de apalancamiento, margen operacional, entre otras. Estas variables fueron seleccionadas de manera que el modelo funcionara correctamente.

Se espera que el proyecto sea continuado y mejorado para ser utilizado en otras empresas del sector financiero o incluso en otros sectores económicos.

1.1. Problemática

La economía norteaméricana es una de las más fuertes del mundo que tiene influencia sobre casi todas las economías del planeta; cuando sucedió la crisis del 2008 en Estados Unidos, afectó prácticamente la economía de todos los países. A tal nivel que muchas empresas quebraron y muchos inversionistas perdieron su dinero.

1.2. Objetivos

1.2.1. Objetivo general

Evaluar la eficacia del modelo de regresión logistica en el fenómeno de bancarrota en los bancos estadounidenses para que pueda ser utilizado por los inversionistas colombianos en la toma de decisiones de inversión.

1.2.2. Objetivos específicos

- Definir y conocer claramente los conceptos financieros y de programación para poder implementar el modelo probailístico de manera eficaz.
- Entender el funcionamiento de otros modelos que se relacionan al modelo probabilístico utilizado para poder mejorar y perfeccionar la eficiencia del modelo a la hora de aplicarlo.
- Hacer una investigación sobre un evento ampliamente conocido para poder utilizar modelos, ejemplos y conceptos potentes que ayuden a obtener resultados óptimos.

2. Justificación

La Economía de los Estados Unidos de América, por su cercanía geográfica y política, afecta directamente la economía de la gran mayoría de los países de américa latina, incluyendo la colombiana. Cerca de la cuarta parte de las importaciones colombianas provienen de Estados Unidos, e igualmente alrededor el 26 % del total de exportaciones de Colombia tienen como destino dicho país. Es por ello que el motivo de esta investigación y aplicación tuvo como objeto analizar la bancarrota de los bancos estadounidenses durante el período 2007-2017.

2.1. Antecedentes

La invesigación hecha por 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 esta metodología, el cual consistía en un modelo logit para predecir bancarrota pero según 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 el objeto de estudio, que es analizar la crisis financiera del año 2008 enfocada a la decisión de inversión. Además, se utilizaron conceptos 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 fueron utilizados en la siguiente investigación:

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

2.2. Conceptos

Los conceptos que se definen a continuación, definen parte de los indicadores utilzados para el modelo. Estos fueron elegidos porque representan la salud financiera de un banco o una empresa constiuda.

Liquidez.
Es la velocidad con la que un 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.

Eficiencia.
Es sacar el máximo provecho de los recursos, una empresa es eficiente si máximiza sus ganancias y minimiza los costos.

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.

Productividad.
Calcula cuántos bienes y servicios se han producido por cada factor utilizado.

Razón de aplacamiento.
Es la razón entre el capital de la empresa y las deudas

Rotación de activos.
Mide el nivel de eficiencia con la que una empresa utiliza sus activos para generar ingresos.

Margen operativo.
Cuantifica el porcentaje de ingresos por ventas que la empresa convierte en beneficios, antes de descontar impuestos e intereses.

2.3. La matriz de confusión y sus indicadores

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]

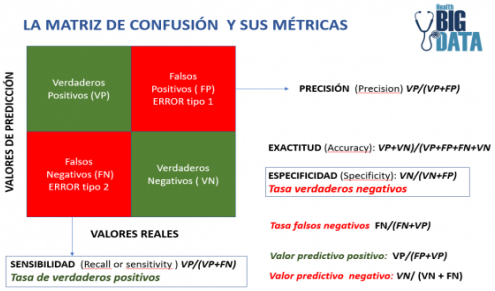

Figura 1 Matriz de confusión

Accuracy
Es una métrica que determina la cantidad de predicciones positivas que fueron acertadas.

Precisión
Es una métrica que determina el porcentaje de casos positivos detectados.

Recall
También conocida como sensibilidad, es una métrica que determina la proporción de verdaderos positivos correctamente identificados.

f1-score
Es una métrica que determina la medida de precisión que tiene un test utilizando la precisión y sensibilidad en una sola métrica.

2.3.1. Regresión Logística binaria

La regresión logística es uno de los métodos estadísticos más usados en machine learning, este sistema de clasificación supervisado pretende modelar cómo influyen las variables en la ocurrencia de una situación particular y así determinar su resultado dicotómico, en otras palabras, estima los valores eperados de $\boldsymbol{y}$, dado las variables.
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, podríamos decir que esta regresión se hace como un modelo de caja negra en el que el computador nos devuelve unos resultados, sin embargo a continuación explicamos como se ve Matemáticamente.
Sea $\boldsymbol{y}$ la variable dependeinte, 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:

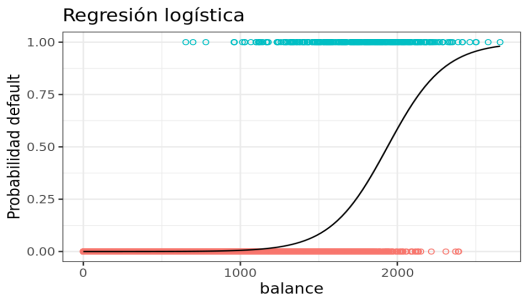

Figura 2 Regresión logística

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:

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

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

2.4. Estudios previos

Modelos desarrollados usando regresión logística

Un primer estudio corresponde a Maria 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.[5]

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.[3]

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.[4]

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

El modelo se implementó en el leguaje python, utilizando sus librerias Pandas, Numpy, sklearn, matplotlib, seaborn y imblearn. Se utilizó un modelo de regresión logística binaria, F1 score y matrices de confusión. En los resultados de la regresión 1 significa bancarrota y 0, no bancarrota. Adicionalmente, para transformar el resultado obtenido por la regresión lineal, se utiliza la función sigmoide.

3.1. Muestra

Los datos utilizados son 120 bancos sanos y 71 bancos quebrados desde 2007 hasta 2017,siendo estos bancos solo norteamericanos. En cada corrida del modelo, se tomaron aleatoriamente el 80 % del total de la información para la etapa de aprendizaje y el 20 % restante para la evaluación. La información fue tomada Kaggle y se utilizaron herrameintas descritas durante la sección.[10]

3.2. Variables

Las x_i serían: Tobin's Q, BPA, 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. 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]

3.4. P values

Esta medida estadística fue utilizada para elegir los indicadores que afectan más al modelo, sin embargo se decidió utilizar todos los indicadores para obtener mejores resultados debido a que cuando eliminamos algunos de los no estadísticamente significativos, habia más error.
De los más importantes para este modelo son: BPA(Beneficio por acción) rentabilidad, productividad, Tobin's Q, margen operativo, rentabilidad de los fondos propios. En la figura 3 podemos observar los P values del modelo

Resultados P: Values:

Dep. Variable:	BK	**No. Observations:**	152
Model:	Logit	**Df Residuals:**	138
Method:	MLE	**Df Model:**	13
Date:	Fri, 27 May 2022	**Pseudo R-squ.:**	0.7191
Time:	14:43:43	**Log-Likelihood:**	-28.245
converged:	True	**LL-Null:**	-100.56

	coef	std err	z	P> |z|	[0.025	0.975]
Data Year - Fiscal	-1.84e-05	0.001	-0.026	0.979	-0.001	0.001
Tobin's Q	-1.5431	0.764	-2.019	0.043	-3.041	-0.045
EPS	0.1427	0.073	1.957	0.050	-0.000	0.286
Liquidity	-1.0581	1.505	-0.703	0.482	-4.008	1.892
Profitability	-0.3310	0.139	-2.374	0.018	-0.604	-0.058
Productivity	8.6644	3.853	2.249	0.025	1.113	16.215
Leverage Ratio	-0.0176	0.068	-0.257	0.797	-0.152	0.117
Asset Turnover	0.3479	0.300	1.159	0.246	-0.240	0.936
Operational Margin	-16.0958	4.944	-3.256	0.001	-25.786	-6.406
Return on Equity	-1.9282	0.858	-2.246	0.025	-3.611	-0.246
Market Book Ratio	-0.0026	0.002	-1.745	0.081	-0.006	0.000
Assets Growth	-2.3049	1.616	-1.426	0.154	-5.473	0.863
Sales Growth	0.8075	1.012	0.798	0.425	-1.176	2.791
Employee Growth	3.4931	1.791	1.951	0.051	-0.017	7.003

Possibly complete quasi-separation: A fraction 0.29 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

Figura 3 P values

4. Conclusiones

4.1. Evaluación del modelo

Con los datos obtenidos mostrados en la siguiente gráfica, obtenemos la precisión del algoritmo. SE obtuvieron 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.

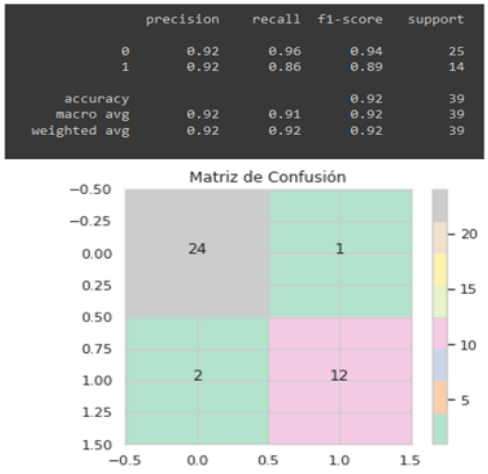

Figura 4 Resultados

Después de evaluar varias veces el modelo, como se ve en la siguiente figura.

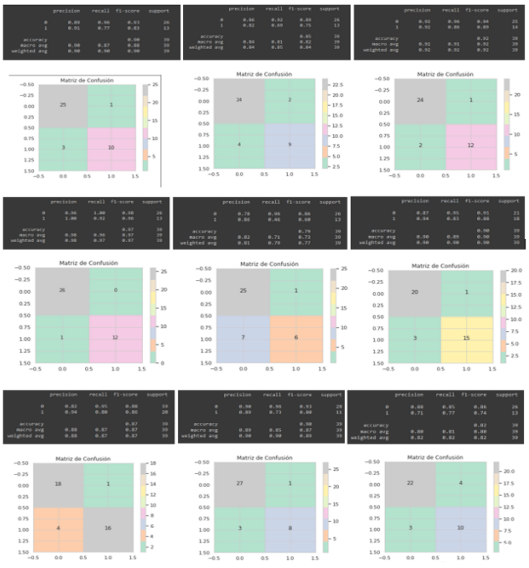

Figura 5 Evaluación

Se obtuvieron varios datos importantes. Es importante racalcar que cada clase tiene su precisión, recall, f1-score, weighted average, macro average, y support distintos ya que se desea evaluar cómo el modelo funciona en cada una de las clases por separado, con esto se puede saber si el modelo es más preciso en una clase que en otra. En la clase 0 (bancos sin bancarrota) se obtuvo en promedio una precisión de 0.87, un Recall de 0.94 y un F1-score de 0.9. En la clase 1 (Bancos en bancarrota) se obtuvo una precisión de 0.85, un recall de 0.75 y un f1-score de 0.81. Como se puede ver, el algoritmo tiene una mejor precisión en determinar los bancos que no caerán en bancarrota. Cabe aclarar que el número de veces que se corrió el algoritmo para hacer el promedio de cada métrica fue 9. La precisión de la clase 0 se debe a que hay más datos sobre los bancos sin bancarrota ya que un banco quiebre no es muy común. Por los resultados dados anteriormente podemos concluir que el modelo es efectivo.

El aporte hecho en este trabajo consitió en seleccionar los datos desde el año 2007 hasta el 2017, haciendo que el modelo tuviera un mejor aprendizaje, obteniendo mejores resultados.

4.2. Futuras investigaciones

Los entes gubernamentales de Estados Unidos tratan de prevenir la bancarrota de los bancos e incluso revertir esa situación. Es por ello que este modelo se utilizó con datos desequilibrados, para probar su fiabilidad. Igualmente se podría utilizar para predecir la bancarrota de empresas, personas o incluso en portafolio de inversiones.

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

[2]FINANZAS "¿Qué es y cómo se logra la eficiencia financiera?" .13 de marzo del 2018. URL: https://rpp.pe/campanas/contenido-patrocinado/que-es-y-como-se-logra-la-eficiencia-financiera-noticia-1110108

[3] Mª Visitación García Jiménez, Jesús Mª Alvarado Izquierdo y Amelia Jiménez Blanco "La predicción del rendimiento académico: regresión lineal versus regresión logística" 2000. Vol. 12,

Supl. nº 2, pp. 248-252.

[4]Juan Camilo Vega, Edgar Guillermo Rodríguez Díaz, Alexandra Montoya R. "Metodología de evaluación del clima organizacional a través de un modelo de regresión logística para una universidad en Bogotá, Colombia" Revista CIFE: Lecturas de Economía Social, ISSN-e 2248-4914, Vol. 14, Nº. 21, 2012, págs. 63-88.

[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/$27_regresion_logistica_simple_ymultiple$.

[9]$Programmingfoundation.I."Module4 Logisticregression".URL$: $https$: $//learn.theprogrammingfoundation.org/getting_started/intro_data_science/module4/$

[10]$USBankruptcyPredictionDataset(1971 - 2017).URL : https://www.kaggle.com/datasets/shuvamjoy34/us-bankruptcy-prediction-data-set-19712017$