Proyecto de semestre - Entrega 1

UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

Materia
Introducción a la inteligencia artificial

Docente
Raul Ramos Pollan

Estudiantes
Hernán Javier Aguilar Cruz
Jhonier Raúl Jiménez Acevedo

Universidad de Antioquia
Medellín
2020

Descripción problema predictivo a resolver

Dada las características de una compañía, definir si esta puede enfrentarse a una bancarrota. El problema a resolver mediante el uso de este dataset hace parte de un problema de clasificación, ya que esperamos un Sí o un No.

Descripción dataset a utilizar

El dataset viene de un diario económico de Taiwán desde 1999 hasta 2009. La intención de este entonces es identificar si una compañía puede quebrarse. Este cuenta con un total de 96 columnas o atributos, y 6819 instancias o filas.

```
0   Bankrupt?
1    ROA(C) before interest and depreciation before interest
2    ROA(A) before interest and % after tax
3    ROA(B) before interest and depreciation after tax
4    Operating Gross Margin
5    Realized Sales Gross Margin
6    Operating Profit Rate
7    Pre-tax net Interest Rate
8    After-tax net Interest Rate
9    Non-industry income and expenditure/revenue
10   Continuous interest rate (after tax)
11   Operating Expense Rate
12   Research and development expense rate
13   Cash flow rate
14   Interest-bearing debt interest rate
15   Tax rate (A)
16   Net Value Per Share (B)
17   Net Value Per Share (A)
18   Net Value Per Share (C)
19   Persistent EPS in the Last Four Seasons
20   Cash Flow Per Share
21   Revenue Per Share (Yuan ¥)
22   Operating Profit Per Share (Yuan ¥)
23   Per Share Net profit before tax (Yuan ¥)
24   Realized Sales Gross Profit Growth Rate
25   Operating Profit Growth Rate
26   After-tax Net Profit Growth Rate
27   Regular Net Profit Growth Rate
28   Continuous Net Profit Growth Rate
29   Total Asset Growth Rate
30   Net Value Growth Rate
```

```
30   Net Value Growth Rate
31   Total Asset Return Growth Rate Ratio
32   Cash Reinvestment %
33   Current Ratio
34   Quick Ratio
35   Interest Expense Ratio
36   Total debt/Total net worth
37   Debt ratio %
38   Net worth/Assets
39   Long-term fund suitability ratio (A)
40   Borrowing dependency
41   Contingent liabilities/Net worth
42   Operating profit/Paid-in capital
43   Net profit before tax/Paid-in capital
44   Inventory and accounts receivable/Net value
45   Total Asset Turnover
46   Accounts Receivable Turnover
47   Average Collection Days
48   Inventory Turnover Rate (times)
49   Fixed Assets Turnover Frequency
50   Net Worth Turnover Rate (times)
51   Revenue per person
52   Operating profit per person
53   Allocation rate per person
54   Working Capital to Total Assets
55   Quick Assets/Total Assets
56   Current Assets/Total Assets
57   Cash/Total Assets
58   Quick Assets/Current Liability
59   Cash/Current Liability
```

```
60   Current Liability to Assets
61   Operating Funds to Liability
62   Inventory/Working Capital
63   Inventory/Current Liability
64   Current Liabilities/Liability
65   Working Capital/Equity
66   Current Liabilities/Equity
67   Long-term Liability to Current Assets
68   Retained Earnings to Total Assets
69   Total income/Total expense
70   Total expense/Assets
71   Current Asset Turnover Rate
72   Quick Asset Turnover Rate
73   Working capitcal Turnover Rate
74   Cash Turnover Rate
75   Cash Flow to Sales
76   Fixed Assets to Assets
77   Current Liability to Liability
78   Current Liability to Equity
79   Equity to Long-term Liability
80   Cash Flow to Total Assets
81   Cash Flow to Liability
82   CFO to Assets
83   Cash Flow to Equity
84   Current Liability to Current Assets
85   Liability-Assets Flag
86   Net Income to Total Assets
87   Total assets to GNP price
88   No-credit Interval
89   Gross Profit to Sales
90   Net Income to Stockholder's Equity
```

Nota: Este dataset cumple con los 2 primeros requerimientos, sin embargo no se cumplen con los últimos 2 ya que ninguna columna es categórica (se componen de int64 o float64) y no hay datos faltantes. Para la siguiente entrega se espera trabajar con el dataset de tal manera que cumpla con los requisitos. Así, trabajar con los datos para balancear el dataset.

Este fue obtenido del UCI Machine Learning Repository sin embargo puede ser encontrado en el repositorio de github del proyecto.

Métricas de desempeño requeridas

Para el desarrollo de este proyecto se plantean utilizar las siguientes métricas de Machine Learning: (Estas pueden variar de entrega a entrega).

Logistic Regression

Random Forest Classifier

SVC

KNN Confusion matrix

Naive Bayes

Desempeño deseable en producción

Una vez se ha trabajado con los datos, y se tiene el accuracy de cada modelo, se desea que cada uno cuente con un 90% o más de accuracy. Esto significa entonces que los modelos a utilizar tienen una efectividad considerable, por lo que valdría la pena ponerlos en producción. Caso contrario, de ser menor al 90%, ya existiría un error más grande, por lo que las empresas que quieren saber si es posible que entren a bancarrota no estarían interesados ya que les gustaría más precisión al momento de realizar el análisis, entonces no valdría la pena poner el modelo en producción.