Ciencia de datos para la predicción de insolvencia empresarial

Iván Andrés Trujilllo Abella

trujilloiv@javeriana.edu.co

Insolvencia empresarial

...

El fenómeno de insolvencia (las empresas no pueden cumplir con sus obligaciones financieras) genera:

- Costo directo de capital (Perspectiva de acreedores)
- Costo social, derivado de negar crédito a aquellas que si pueden cumplir con sus obligaciones financieras.

Justificación

Mejorar la capacidad predictiva de los modelos o expertos reduce dichos costos y aumenta el margen de ganancia (reduciendo costos de oportunidad).

Modelo

...

Supongamos que tenemos de alguna u otra forma una solución que involucre analítica, ciencia de datos, o un experto en el área.

El desempeño

Podríamos preguntarnos que también hace su trabajo el modelo o el experto, pero antes debemos especificar la utilidad del ejercicio.

La utilidad

...

Real	Predicción	Resutaldo
Default	Default	0
Default	No-default	$-K$
No-default	Default	$-\pi$
No-default	No-default	π

Función de utilidad

Función de utilidad

$$U(K, \pi) = (\hat{y}_i - 1) y_i K + \pi(1 - y_i)(1 - 2\hat{y}_i) \tag{1}$$

Donde y_i es el suceso real, K capital promedio prestado, π es la ganancia sobre el capital y $\hat{y}_i$ es la predicción del modelo.

(Simulador)

	Model		Profit
A	Specificity	0.73	13507.1
	Sensitivity	0.80	
B	Specificty	0.75	14979.42
	Sensitivity	0.81	

...

La ganancias en la capacidad predictiva del modelo representan aproximadamente un incremento de 11 porciento en la ganancia promedio.

¿Como los construimos?

...

El criterio del experto es importante pero el modelo nos permite conocer bajo diferentes escenarios su desempeño de forma más económica.

...

Desde Beaver con técnicas bivariadas, pasando por el MDA de Altman hasta contemplar el uso de algoritmo híbridos. En Colombia son comúnes:

- Regresión logística
- Modelo probit
- Random Forest
- Máquina de soporte vectorial...

las variables

Ratios

Utilidad

Dos empresas A y B:

- A genera 1000 um al mes
- B genera 1500 um al mes

las variables

Ratios

Utilidad

Dos empresas A y B:

- A genera 1000 um al mes
- B genera 1500 um al mes

Capital

- A tiene 500 um de capital
- B tiene 1000 um de capital

las variables

Ratios

Utilidad

Dos empresas A y B:

- A genera 1000 um al mes
- B genera 1500 um al mes

Capital

- A tiene 500 um de capital
- B tiene 1000 um de capital

...

- A tiene un ratio de 2
- B tiene un ratio 1.5

Las razones financieras

...

Pueden representar mejor la realidad económica de una empresa o las tendencias.

Tipos

- Razones de liquidez
- Razones de endeudamiento
- Razones de rentabilidad....

Ratio	Definition
Gross Profit Margin (GPM)	$\frac{\text{Gross profit}}{\text{Operating revenues}}$
Net Profit Margin (NPM)	$\frac{\text{Profit(Loss)}}{\text{Operating revenues}}$
Return On Equity (ROE)	$\frac{\text{Profit(Loss)}}{\text{Total equity}}$
Return On Assets (ROA)	$\frac{\text{Profit(Loss)}}{\text{Total assets}}$
Indebtedness Ratio (IR)	$\frac{\text{Total liabiliaties}}{\text{Total assets}}$
Debt Equity Ratio (DER)	$\frac{\text{Total liabiliaties}}{\text{Total equity}}$
Current Ratio (CR)	$\frac{\text{Total current assets}}{\text{Total current liabilities}}$
Ratio of Short-Term Liabilities (RSL)	$\frac{\text{Total current liabilities}}{\text{Total liabilities}}$
Altman x_1	$\frac{\text{Total current assets - Total current liabilities}}{\text{Total assets}}$

Ejercico empírico

...

Datos recuperados de la Superintendencia de Sociedades para el periodo (2016-2019) en las Pymes colombianas, definiendo como evento las empresas acogidas en algún proceso de insolvencia.

...

Como variables se usaron las razones financieras expuestas, un año antes del evento, se usó el promedio de razones durante el período de actividad para las empresas que no presentaron el evento.

Criterios de exclusión

- Información financiera presentada una fecha diferente al 31 de diciembre.
- Aquellas empresas que se declararon en una fase preoperativa
- Las empresas que presentaron el evento en el año 2016.
- Firmas que no presentaron información contable completa o no valida.

Modelos

...

- Regresión logística
- Máquina de soporte vectorial
- Red Neuronal ....

Information leakage

Es importante aplicar un flujo analítico adecaudo para no sobrestimar la capacidad predicitiva de los modelos.

- Estandarización de los datos
- Selección de variables
- Train - test datasets
- Undersampling (sobre datos de entrenamiento)
- Validación cruzada para tuneo de hiperparámetros (5 capas).

Lab

(Laboratorio)

Selección de variables

...

La selección de variables es importante porque permite encontrar el mejor conjunto de variables para el problema, aquí utilizamos un algoritmo genético para realizar la tarea, seleccionandose 5 ratios (*ROE, ROA, IR, DER, RSL*), y 8 sectores de un total de 20, usando regresión logística como clasificador.

Selección de variables

Usando logística como clasificador

...

	Genetic selection		Forward selection	
	No-Bankrupt	Bankrupt	No-bankrupt	Bankrupt
precision	0.99	0.05	0.99	0.05
recall	0.73	0.80	0.73	0.78
f1-score	0.84	0.10	0.84	0.10
support	3296	64	3296	64

Desempeño de los modelos

...

	Logistic regression		Backpropagation	
	No-Bankrupt	Bankrupt	No-Bankrupt	Bankrupt
precision	0.99	0.05	1.00	0.06
recall	0.73	0.80	0.75	0.81
f1-score	0.84	0.10	0.86	0.11
support	3296	64	3296	64

Consideraciones

- Flujo analítico
- Selección de variables
- Mejoras "residuales" en los modelos son mejoras sustanciales en la utilidad
- Los modelos híbridos han presentado en la literaura mejoras en la capacidad predictiva.