Proyecto Bancarrota

Integrantes:

Brian Winkelman
Ignacio Baratto
Ingrid Rodrigues
Rodrigo Zelaya
Sebastián Salcedo

Coderhouse

22/03/2022

Índice de contenido

Descripción del caso de negocio

El caso de negocio presentado aborda la temática de la clasificación de empresas en dos grandes grupos: aquellas que presentaron bancarrota y aquellas que no. El universo de estudio se reduce a aquellas empresas que cotizaron en la bolsa de valores de Taiwán en el período de años 1999 – 2009. A través de distintas variables financieras, se elabora un algoritmo de clasificación para tratar la temática descripta.

Tabla de versionado

Versión	Fecha	Descripción del versionado
1.0	12/01/2022	Creación del cuaderno de trabajo (Jupyter notebook) conteniendo: adquisición de datos, manejo y limpieza de datos y análisis exploratorio. Elaboración de la presentación ejecutiva.
2.0	13/02/2022	Incorporación del tratamiento de valores atípicos y posterior construcción y evaluación de distintos algoritmos de clasificación: árbol de decisión, bosque aleatorio, regresión logística y máquina de vectores de soporte (SVM, por sus siglas en inglés).
3.0	23/02/2022	Optimización del modelo de clasificación elegido (bosque aleatorio) y evaluación de dicho modelo considerado como final.
4.0	22/03/2022	Elaboración del documento ejecutivo como presentación, respaldo y resumen del trabajo realizado.

Objetivo del modelo

Se busca clasificar a las empresas que cotizaron en la bolsa de Taiwán en el período de años 1999 – 2009 en dos grandes grupos: empresas que declararon bancarrota y empresas que no declararon bancarrota. Por lo tanto, se propone como objetivo construir distintos modelos de clasificación dicotómicos que permitan discriminar a las empresas en cuestión en los dos grandes grupos mencionados.

Una vez construidos los modelos, se pretende compararlos y seleccionar aquel que tenga mejor desempeño para posteriormente realizar la optimización del mismo.

Descripción de los datos

El conjunto de datos fue extraído de la página Kaggle y contiene información financiera de empresas que cotizaron en la bolsa de valores de Taiwán en el período 1999 - 2009. La elección de dichos datos se basó en las preferencias de los integrantes, encontrando un balance entre los gustos personales y el orden y la robustez de los datos. Así, seleccionamos un conjunto de datos "limpio" y completo con el fin de predecir la posibilidad de que una empresa declare bancarrota.

Con el fin de disponer de datos ordenados y útiles, se realizó una limpieza y reducción de los mismos a través del tratamiento de los valores atípicos y un análisis de correlación.

De esta forma, se redujo el conjunto de datos a las variables con mayor correlación (tanto positiva como negativa) respecto a la variable objetivo ('Bankrupt?'), estableciendo un límite absoluto de 10% de correlación. Luego, se excluyeron de estas variables aquellas que tenían una correlación muy alta entre sí (se establece límite absoluto de 80% de correlación).

Luego, se identificó una variable con una muestra altamente desbalanceada ('Liability-AssetsFlag'), la cual se elimina por completo del conjunto de datos.

Por último, se trataron los valores atípicos con el fin de que el conjunto de datos no se vea afectado por los valores extremos de cada variable, por medio del método de *capping* y *flooring*.

Una vez realizado este proceso, quedó configurado el conjunto de datos conteniendo las siguientes variables, seguidas cada una de su descripción correspondiente:

- **'Bankrupt?'**: es la variable objetivo y establece una relación dicotómica de las empresas entre aquellas que declararon bancarrota y aquellas que no lo hicieron.
- **'NetIncometoTotalAssets'**: esta variable captura lo que se conoce como el retorno sobre activos, que se define como el ratio entre el ingreso neto de una compañía y el total de sus activos. Por lo tanto, es un coeficiente de rentabilidad. A mayor (menor) ratio, mayor (menor) retorno sobre activos.
- **'Networth/Assets'**: esta variable se define como la resta entre el total de activos y el total de pasivos (patrimonio) de una empresa, dividido entre el total de activos. De esta forma, captura el porcentaje que representa el patrimonio de una empresa en el total de activos. A mayores (menores) porcentajes, mayor (menor) es el patrimonio de una empresa y menor (mayor) es su deuda en relación al total de sus activos.
- **'PersistentEPSintheLastFourSeasons'**: esta variable captura la magnitud en que las ganancias por acción publicadas resultaron persistentes en las últimas cuatro temporadas/años calendario. A mayor (menor) persistencia, mayor (menor) rendimiento por acción.
- **'RetainedEarningstoTotalAssets'**: esta variable representa el ratio entre las ganancias retenidas de una empresa y el total de sus activos. A mayor (menor) ratio, mayores (menores) son las ganancias retenidas por la empresa en relación a sus activos.
- **'WorkingCapitaltoTotalAssets'**: esta variable representa el ratio entre los activos líquidos netos (capital de trabajo) de una empresa y el total de sus activos. A través de este ratio se evalúa la solvencia a corto plazo de una empresa. A mayor (menor) ratio, mayor (menor) es la solvencia a corto plazo.
- **'NetIncometoStockholder'sEquity'**: esta variable se define como el ratio entre el ingreso neto de una compañía y el capital contable (activos que permanecen una vez que se han liquidado todos los pasivos) de una empresa. Por lo tanto, es un coeficiente que permite evaluar el estado general financiero de una empresa. A mayor (menor) ratio, mayor (menores) es la capaciad de una empresa de cubrir sus obligaciones (pasivos).
- **'CurrentLiabilitytoCurrentAssets'**: esta variable representa el ratio entre los pasivos a corto plazo de una empresa y el total de sus activos a corto plazo. Así,

permite evaluar la liquidez a corto plazo de una empresa. A mayor (menor) ratio, menor (mayor) es la liquidez a corto plazo.

- **'NetValuePerShare(A)'**: el valor liquidativo por acción se calcula dividiendo el valor liquidativo por el número total de acciones (de tipo A) en circulación. De esta forma, representa el valor liquidativo por acción (de tipo A). A mayor (menor) valor, mayor (menor) es el valor por acción.
- **'WorkingCapital/Equity'**: esta variable representa el ratio entre los activos líquidos netos (capital de trabajo) de una emrpesa y el total de su patrimonio. Similar a la variable "WorkingCapitaltoTotalAssets", este ratio evalúa la solvencia a largo plazo de una empresa. A mayor (menor) ratio, mayor (menores) es la solvencia a largo plazo.
- **'Total expense/Assets'**: esta variable captura el porcentaje de gastos totales de una empresa en relación al total de activos. A mayor (menor) porcentaje, mayor (menor) es el nivel de gastos de una empresa en relación a sus activos.
- **'CFOtoAssets'**: esta variable captura el porcentaje de flujo de de efectivo dispuesto para operaciones sobre el total de activos de una empresa. Sirve como indicador de la eficiencia de una empresa a la hora de utilizar sus activos para cobrar efectivo. A mayor (menor) ratio, mayor (menor) eficiencia.
- **'Taxrate(A)'**: representa la tasa de impuestos corporativos que paga una empresa por operar.
- **'Cash/TotalAssets'**: esta variable muestra la proporción de los activos de una empresa que se componen de efectivo e inversiones a corto plazo. A mayor (menor) ratio, mayores (menores) son los activos compuestos por efectivo e inversiones a corto plazo.
- **'GrossProfittoSales'**: esta variable representa la proporción de ganancia bruta de una empresa (costo de los bienes vendidos menos la cifra total de ventas netas) sobre sus ventas totales. A mayor (menor) ratio, mayor (menor) es la ganancia bruta.

Conclusiones análisis exploratorio

Por medio del análsis gráfico presentado en el Jupyter notebook, se identiciaron las siguientes tendencias en el conjunto de datos:

- Hay un número muy pequeño de empresas declararon bancarrota en el período 1999 - 2009.
- En promedio, las empresas que declararon bancarrota presentan mayor nivel de liquidez a corto plazo y mayores gastos en relación a sus activos, en comparación con las empresas que no declararon bancarrota y siguen cotizando en bolsa.
- A su vez, en promedio, éstas empresas también presentan menor nivel de ingreso, patrimonio, ganancia retenida, capital de trabajo, ingreso neto, valor liquidativo por acción, flujo de efectivo para operar, impuestos pagos, efectivo y ganancia bruta en relación a su total de activos, en comparación con las empresas que no declararon bancarrota y siguen cotizando en bolsa.

Algoritmo elegido

A la hora de construir el modelo de clasificación, se seleccionaron los siguientes algoritmos: árbol de decisión, bosque aleatorio, regresión logística y máquina de vectores de soporte. A continuación, se presenta una definición y descripción de cada modelo, los cuales están comprendidos dentro de lo que se denomina aprendizaje supervisado.

Árbol de decisión: permite construir un diagrama de decisión con forma de árbol, por medio del cual se construye un flujo de acción (conecta hojas a través de nodos) de acuerdo a una probabilidad (de pertenecer a una determinada clase) asignada en cada nodo de decisión.

Bosque aleatorio: es un conjunto de árboles de decisión, por lo que sigue la misma lógica de construcción del algoritmo descripto anteriormente, pero permitiendo la creación de más de un escenario (más de un árbol) posible de decisión.

Regresión logística: es un algoritmo de regresión que permite predecir la probabilidad de pertenecer a una determinada categoría de la variable objetivo. Tal como su nombre lo indica, el resultado de la variable categórica es predicho a través de la función logística.

Máquina de vectores de soporte: es un algoritmo que se fundamenta en la separación de los datos en el espacio, dividiendo a las clases de la variable objetivo en dos espacios lo más amplio posibles, por medio de un vector de soporte.

Métricas de desempeño

Modelo de predicción	Accuracy		Precision		Recall		F1-score		AUC
	Set de entrenamiento	Set de evaluación	Clase 0	Clase 1	Clase 0	Clase 1	Clase 0	Clase 1	
Regresión Logística	88%	87%	99%	18%	87%	86%	93%	30%	93%
SVM	88%	86%	99%	17%	86%	86%	92%	29%	93%
Árbol de Decisión	100%	93%	98%	17%	95%	34%	96%	23%	64%
Bosque Aleatorio	100%	95%	98%	29%	97%	41%	97%	34%	92%

Al tener una muestra desbalanceada, se establece que la métrica *F1-score* es la más conveniente para evaluar al modelo. Esta métrica permite realizar un promedio entre lo que es la *precision* y el *recall*. Así, se ve claramente que el bosque aleatorio es el modelo de predicción que mejor se desempeña para el problema de clasificación planteado.

En resumen, la métrica *F1-score* toma en cuenta tanto la calidad de la predicción (*precision*) como la cantidad correcta de preddiciones (*recall*) de la clase de interés, siendo 29% y 41% respectivamente en el modelo de predicción del bosque aleatorio. Si se multiplican ambas medidas entre sí (0.29 x 0.41), este resultado luego se multiplica por la suma de ambas medidas (0.7) y, finalmente, dividen estos dos resultados (0.29 * 0.41 / 0.7) y se multiplica por dos (2 x 0.29 x 0.41 / 0.7), obtenemos la medida resumen de *F1- score* del modelo.

Iteraciones de optimización

Por medio del método de la validación cruzada *k-fold* en conjunto con iteraciones del bosque aleatorio empleando distintos hiperparámetros, se optimizó el modelo recorriendo un amplio espectro de posibilidades. Si bien esto requiere una cantidad significativa de tiempo, se puede ver en la siguiente sección que es sumamente beneficioso para mejorar el modelo.

Métricas finales del modelo optimizado

Modelo de predicción	Accuracy		Precision		Recall		F1-score		AUC
	Set de entrenamiento	Set de evaluación	Clase 0	Clase 1	Clase 0	Clase 1	Clase 0	Clase 1	
Bosque Aleatorio (parámetros estándar)	100%	95%	98%	29%	97%	41%	97%	34%	92%
Bosque Aleatorio (hiperparámetros)	100%	95%	98%	32%	97%	45%	97%	38%	92%

Teniendo en cuenta la descripción realizada en el apartado *Métricas de desempeño*, se puede apreciar que luego de las iteraciones mencionadas, la métrica elegida para evaluar el modelo asciende de 34% a 38%. Por lo tanto, hay una mejora sustancial en la capacidad del modelo, tanto para identificar la cantidad de población de la clase de interés como para identificar correctamente a la clase de interés.

Futuras líneas

Una vez desarrollado el modelo, se puede poner en práctica para un conjunto de datos similares con el fin de extender el algoritmo de clasificación para comprender empresas de distintas partes del mundo o de un mercado en particular. Dicho esto, al tratar con conjuntos de datos similares, pero no iguales, puede requerir de un esfuerzo adicional al necesitarse de una limpieza de los datos (cambiar etiquetas, realizar transformaciones, entre otros) más exigente, aunque resultaría provechoso al estar desarrollado el modelo.

Además, también puede pensarse en probar modelos similares o con pequeñas modificaciones que resulten en mejores métricas de desempeño (nuevos modelos y/o nuevos hiperparámetros), así como también implementar otras metodologías para la optimización del modelo (ensamble y *boosting*).

Conclusiones

A modo de conclusión, teniendo en cuenta las particularidades de la muestra y las características del conjunto de datos, se ha podido construir un modelo de predicción aceptable para poder identificar a las empresas que presentaron bancarrota en el mercado de valores de Taiwán en el período 1999 – 2009. Si bien las métricas del modelo no demuestran un desempeño excelente, se puede concluir que el modelo permite identificar y predecir correctamente a casi 4 de cada 10 empresas que entraron en quiebra.

Así, abordando el problema desde la perspectiva del negocio se puede utilizar el modelo para asesorar a un inversionista, indicándole que en algunas ocasiones (4 de cada 10) no le convendrá invertir en empresas, ya que las mismas presentarán bancarrota. En los casos restantes (6 de cada 10), ante la duda se preferirá no invertir y así evitar el peor error (lo que se denomina *error tipo II*), siendo en este caso el peor error invertir cuando no debía invertir (cuando la empresa presenta bancarrota).