115G

MÉTRICAS Y PARAMETRIZACIÓN AVANZADAS EN MODELOS DE PREDICCIÓN DE FRACASO EMPRESARIAL OBTENIDOS MEDIANTE PROGRAMACIÓN GENÉTICA.

Autores:

Ángel Beade[1]
Manuel Rodríguez[1]
José Santos[2]

[1] Departamento de Empresa, Universidad de A Coruña – Cátedra AECA-Abanca, IESIDE.

[2] Departamento de Ciencias de la Computación y Tecnologías de la Información – Centro de Investigación en Tecnologías de la Información (CITIC), Universidad de A Coruña,

Área temática:
Nuevas tecnologías y contabilidad.

Palabras clave:
Fracaso empresarial, modelos de predicción, inteligencia artificial, programación genética.

MÉTRICAS Y PARAMETRIZACIÓN AVANZADAS EN MODELOS DE PREDICCIÓN DE FRACASO EMPRESARIAL OBTENIDOS MEDIANTE PROGRAMACIÓN GENÉTICA.

RESUMEN

En este trabajo se pretende obtener modelos de predicción del fracaso empresarial con un horizonte de corto, medio y largo plazo, aplicando técnicas de inteligencia artificial como la Programación Genética. El diseño y análisis de los modelos obtenidos se centra en la elaboración de una estrategia metodológica para evaluar el rendimiento de dichos modelos a fin de obtener mejor poder de clasificación y, consecuentemente, mayor precisión en la capacidad predictiva en todo el horizonte temporal.

ABSTRACT

The aim of this work is to obtain predictive models of business failure with a short, medium and long term horizon, applying artificial intelligence techniques such as Genetic Programming. The design and analysis of the models obtained focuses on the development of a methodological strategy to evaluate the performance of these models in order to obtain better classification power and, consequently, greater accuracy in the predictive capacity over the entire time horizon.

1. INTRODUCCION

En este trabajo se aborda la predicción del fracaso empresarial mediante modelos a corto, medio y largo plazo. El estudio se centra en la metodología de evaluación del rendimiento de dichos modelos (imprescindible a la hora de parametrizar, realizar comparaciones externas, etc.).

La predicción del fracaso empresarial es un problema de clasificación. En esta tipología de modelos es frecuente la utilización del área bajo la curva ROC (Receiver Operating Characteristic) - en lo sucesivo denominada AUC - como forma de medir el rendimiento de una solución. Sin embargo, ciertas peculiaridades, en el caso concreto de la predicción del fracaso empresarial, aconsejan la utilización de métricas más específicas.

A continuación, se reseñan las antedichas peculiaridades más relevantes, que no son sino aspectos básicos del problema, unos que conciernen a la naturaleza del mismo y otros al objetivo perseguido por la modelización de la antedicha predicción. Sin orden de prelación, estos aspectos son los siguientes:

1.1. Aspectos básicos: Naturaleza del problema

Desde el punto de vista de este estudio, el problema de la predicción del fracaso empresarial presenta dos características relevantes.

- Presenta una clase minoritaria (empresas fracasadas) totalmente desequilibrada en tamaño con respecto a la clase mayoritaria (empresas no fracasadas).
- Tomando como clase positiva a las empresas fracasadas, el error de clasificación no tiene la misma relevancia en el caso del porcentaje de falsos positivos (False Positive Rate - FPR) (empresas que se clasifican por el modelo como fracasadas, pero que están etiquetadas como no fracasadas) y en el caso del porcentaje de falsos negativos (False Negative Rate - FNR) (empresas que se clasifican por el modelo como no fracasadas, pero que están etiquetadas como fracasadas). Es FNR el que presenta – generalmente – una relevancia y un coste asociado mayor.

1.2. Aspectos básicos: Objetivo perseguido

El objetivo final de la predicción del fracaso empresarial es la búsqueda de soluciones que sean aplicables en un contexto real y no únicamente en el entorno de aprendizaje. Por lo tanto, se pretende que los distintos modelos sean capaces de generalizar bien,

por lo que se analizarán los resultados obtenidos al aplicar las soluciones al conjunto de test y no al conjunto de entrenamiento.

Entre los propósitos se encuentra, asimismo, el aplicar distintas evaluaciones a una solución. Con ello se pretende sintetizar los distintos enfoques de los usuarios. Concretamente, se evaluarán las soluciones de tres formas diferentes:

- A nivel global.
- En un área delimitada por 2 restricciones (p.ej.: porcentaje mínimo de verdaderos positivos y porcentaje mínimo de verdaderos negativos).
- En un umbral definido por 1 restricción (p.ej.: porcentaje de verdaderos positivos).

Se pretende la evaluación del rendimiento de la solución: 1) en un conjunto de umbrales de clasificación (todos, en el caso de una evaluación global, o un subconjunto, en el caso de una evaluación en un área delimitada por 2 restricciones) y 2) en un único umbral que no tiene por qué coincidir con el umbral con un mayor porcentaje de aciertos.

El nivel global es el que cabría utilizar para comparar alternativas de parametrización y/o comparaciones externas. No se conoce, o no es relevante, el uso particular de un modelo.

Por el contrario, si se piensa en una empresa financiera que desea implantarse en una zona en expansión y crecer por la vía de concesión de crédito a empresas, el modelo global no es relevante. Parece más razonable que el modelo deba ajustarse a unos requisitos mínimos o restricciones:

- De minimización de falsos negativos (el modelo indica que la empresa no fracasa, pero – por el contrario – sí lo hace), que se controlaría por la fijación de un porcentaje mínimo de verdaderos positivos (TPR) que es su complementario.
- Simultáneamente, interesa que el modelo sea capaz de detectar el mayor porcentaje posible de empresas no fracasadas (porcentaje de verdaderos negativos o TNR), mercado potencial al que interesa dirigirse. Con lo que debería fijarse un mínimo exigible al modelo.

Si pensamos en una empresa que quiere ser sumamente cauta en la concesión de crédito a sus clientes, cabría pensar en que desea un modelo con un elevado poder predictivo cuando el modelo dice que la empresa no fracasa. Ello se traduce en un porcentaje mínimo de falsos negativos (FNR) y – consecuentemente – fijará un elevado nivel de TPR al modelo y sobre el mismo decidirá.

1.3. Técnica de modelización: Programación genética

Son múltiples las técnicas que pueden utilizase para la modelización de la predicción del fracaso empresarial. En este estudio, la utilizada será la Programación Genética (PG) mediante clasificación simbólica.

La PG (Koza 1992) es una técnica de computación evolutiva que resuelve problemas automáticamente sin requerir que el usuario conozca o especifique previamente la forma o la estructura de la solución. A nivel más abstracto, es un método sistemático e independiente del dominio para conseguir que los ordenadores resuelvan problemas de forma automática partiendo de un conocimiento de alto nivel sobre lo que "se necesita hacer" (Poli et al. 2008).

Lo distintivo en PG es que se va evolucionando una población de "programas" de ordenador. En otras palabras, generación a generación, PG transforma -estocásticamente - poblaciones de programas en otras nuevas y, posiblemente mejores, poblaciones de programas (Poli et al. 2008).

La PG tiene algunas características relevantes (no necesariamente exclusivas de la PG). Son, sin orden de prelación, las siguientes:

- Permite graduar la complejidad de las soluciones del modelo al poder actuar sobre:
    - Las variables de entrada a utilizar (variables explicativas). Se pueden incluir variables cualitativas binarias y categóricas.
    - El conjunto de funciones que podrá utilizar el modelo. Se pueden incluir todo tipo de funciones matemáticas, trigonométricas, etc. además de funciones condicionales, de evaluación, etc.
    - La extensión de la solución. Se puede limitar la longitud y profundidad de las soluciones aportadas, actuando sobre dichos parámetros.
- La PG no establece hipótesis previas sobre las variables explicativas del modelo. Unido a su baja sensibilidad a la multicolinealidad permite utilizar un sinfín de transformaciones de las variables explicativas sin prácticamente restricciones.
- El proceso automático de selección de variables que conlleva la PG, ajustado siempre a cada modelo en particular. Lo que permite obviar – si se considera pertinente – el proceso previo de selección de variables con otros métodos. La propia PG se encarga de dicha selección.

El presente trabajo constituye la evolución del presentado previamente en la IX Jornada Internacional AECA de Valoración, Financiación y Gestión de Riesgos (Beade, Santos, y Rodríguez López 2022). En dicho trabajo se puede encontrar información adicional a la de este estudio, concretamente en lo que se refiere a población utilizada y criterios de selección y exclusión.

En el trabajo presente ampliamos los métodos de parametrización. Proponemos y utilizamos nuevas métricas adecuadas al ámbito de predicción del fracaso, que incluyen la definición de una zona de interés en el clasificador/predictor, junto a una medida de rendimiento basada en el coeficiente de Gini.

2. METODOLOGIA Y METRICAS

2.1. Métricas habituales

Muchas de las métricas más habituales (léase porcentaje de aciertos o accuracy, F1 score, coeficiente de correlación de Matthews, etc.) se refieren exclusivamente a un único umbral, generalmente el que el software considera mejor solución (p.ej.: con mayor accuracy).

Como se ha indicado, es muy frecuente utilizar como métrica el área bajo la curva ROC (Receiver Operating Characteristic) como forma de medir el rendimiento de una solución en los problemas de clasificación. La curva ROC representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) en los diferentes umbrales de clasificación (a diferencia de otras medidas que se definen para un umbral específico). La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario a medida que se varía el umbral de discriminación. Efectivamente, la sensibilidad es la capacidad de detectar los casos positivos (TPR), mientras la especificidad es la tasa de acierto con respecto a los casos negativos (TNR). Pone en relación la cantidad de casos negativos detectados con la cantidad de casos negativos totales (TNR = TN/(TN+FP)), por lo que FPR = FP/(TN+FP) = 1 – especificidad).

El área bajo la curva ROC (AUC) mide el área bidimensional bajo la curva ROC completa, proporcionando una medida agregada del rendimiento en todos los umbrales de clasificación posibles.

Nótese que cada umbral determina TPR y FPR, pero, al mismo tiempo, está determinando las tasas de la clase negativa: la tasa de verdaderos negativos (TNR),

complementaria de FPR y la tasa de falsos negativos (FNR) complementaria de TPR. Por ello, a efectos de interpretación puede optarse por hacerlo desde el punto de vista de la clase positiva o de la clase negativa.

El AUC, como forma para evaluar las prestaciones globales de un clasificador, sería una métrica en principio aceptable en este trabajo. De cualquier forma, existen algunos aspectos concretos que caracterizan a AUC y que es preciso no olvidar en su utilización - en el problema del fracaso empresarial - como métrica básica en las comparaciones que se puedan realizar. Son los siguientes:

- El AUC no es lo mismo que la que podría denominarse área de interés de una solución (Figura 1). Cuando se evalúa una curva ROC hay dos áreas que describen el comportamiento de dicha solución en situaciones "extremas", son:
    - El área a la derecha, donde el porcentaje de verdaderos positivos (True Positive Rate - TPR) es elevado, acompañado por un elevado porcentaje de falsos positivos (False Positive Rate - FPR).
    - El área a la izquierda, donde conviven un bajo TPR con un bajo FPR.

  El AUC sintetiza toda el área bajo la curva ROC en un número, cuando las antedichas áreas pueden no ser relevantes en un caso dado. Este enfoque no es nuevo ni extraño en el campo médico, (Dodd y Pepe 2003; McClish 1989) y, en el caso concreto de la predicción del fracaso empresarial, parece adecuado pensar que no se debería escoger una solución únicamente por su AUC sin verificar que parte del mismo se debe a estas áreas.

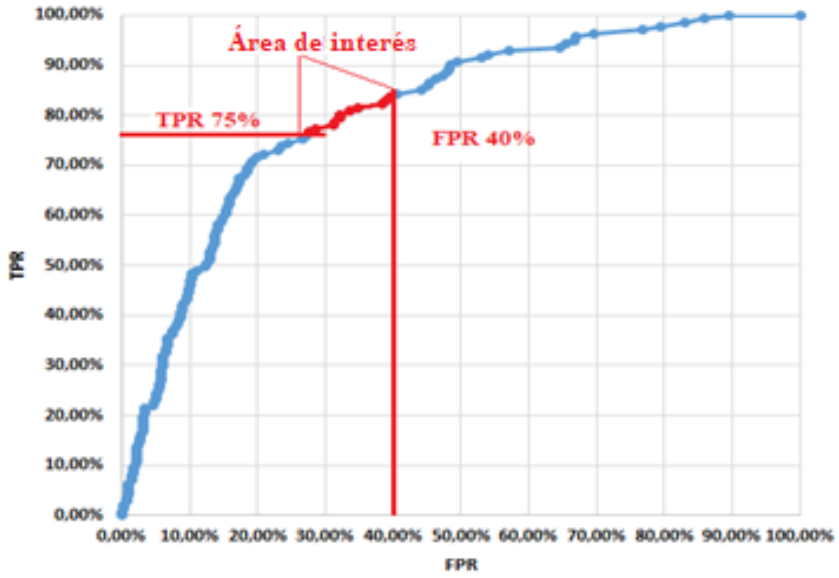

Figura 1: Área de interés de un modelo (mínimo 75% TPR y máximo 40% FPR)

- Para AUC el TPR y el FPR tienen la misma ponderación (Hand 2009). Si nuestro objetivo fuese priorizar TPR, no podrían compararse dos soluciones incluso con similar AUC.
- Relacionado con los costes de los errores de clasificación, en AUC el coste inherente a una correcta clasificación (TPR y TNR) es igual al coste inherente a cada uno los distintos errores de clasificación (FPR y FNR).

  Si se asume que dichos costes son diferentes, como sucede en la realidad con los errores de predicción de fracaso empresarial (p.ej.: si el coste de la correcta clasificación es nulo (TPR y TNR) y el coste de los distintos errores de clasificación (FPR y FNR) es distinto de nulo y distinto entre sí) surgirá una incoherencia en el cálculo de AUC.

  De forma breve, generalmente no se conocen con precisión los costes inherentes a los errores de clasificación, por lo que es habitual recurrir a establecer una ratio entre los distintos errores de clasificación y construir una distribución de probabilidad de dicha ratio de coste de la clasificación errónea. Ello, finalmente, conduce a que AUC evalúa a un clasificador utilizando una métrica que depende del propio clasificador (Hand 2009).

2.2. Software utilizado y sus limitaciones

El software utilizado para implementar el problema en PG ha sido *HeuristicLab* (Wagner et al. 2014).

Los términos solución y programa se utilizan de forma indistinta, y se refiere a la salida de una ejecución de PG. Dicha salida es una función que relaciona las variables independientes y la dependiente en dicha ejecución permitiendo la clasificación.

- *HeuristicLab* (HL) permite dividir el conjunto de entrenamiento (*training*) en dos subconjuntos: aptitud (*fitness*) y validación (*validation*). El primero (aptitud) es el que se usa en la progresión del proceso evolutivo. Esta división en dos subconjuntos va a permitir obtener, en cada ejecución, dos soluciones, una que corresponde al programa que mejor evaluación tiene en el conjunto total de entrenamiento (*best training solution*) y otra que corresponde al mejor programa en el subconjunto de validación (*best validation solution*). Este es el objetivo de dicho subconjunto de validación: permitir detectar soluciones que deberían tener un buen comportamiento al generalizar en el conjunto de test, o, dicho de otro

modo, impedir el sobre-ajuste (*over-fitting*). En este trabajo se ha optado por este enfoque, utilizando inicialmente un conjunto de entrenamiento del cual, en torno al 70% se ha dedicado al subconjunto de aptitud y 30% al subconjunto de validación. Con ello se pretende evitar el sobre-ajuste y obtener soluciones que generalicen bien.

- Asimismo, se le facilita a HL un conjunto de test, compuesto por observaciones diferentes, a fin de que el software aplique a dicho conjunto los modelos obtenidos a partir del conjunto de entrenamiento y el conjunto de validación y obtenga las métricas oportunas.
- HL permite una muy completa y relativamente sencilla parametrización. Las opciones básicas que, por defecto, han sido establecidas inicialmente en la experimentación utilizada en este trabajo, son las siguientes (usando las denominaciones de HL):
    - Clase positiva 1 (empresas fallidas)
    - Evaluator Mean squared error
    - ModelCreator Accuracy Maximizing Thresholds
    - SolutionCreator Probabilistic Tree Creator
    - MaximunDepth 10
    - MaximunLength 100
    - Analyzer Multi Analyzer
    - Crossover Subtree Swapping Crossover - Prob. 90%
    - Elites 1
    - MaximunGeneration 100
    - MutationProbability 15%
    - Mutator Multi Symbolic Expression Tree Manipulator
    - PopulationSize 1500
    - Selector Tournament Group 8

Existen limitaciones propias del software utilizado (HL) y la información que éste suministra.

Con respecto a los propósitos de este trabajo, además de TPR y TNR, la información más relevante de la que se dispone en HL es el coeficiente normalizado de Gini, calculado sobre los valores estimados acotados de la solución (disponible para cada solución – “best training solution” y “best validation solution” - y evaluada en los conjuntos de training y de test). La relevancia viene dada por el hecho de poder utilizarse dicho valor como aproximación al valor que tomará AUC sobre los valores estimados sin acotar. Recuérdese la relación existente entre AUC y el coeficiente normalizado de Gini: AUC = (1 + Gini) / 2.

El acotar los valores estimados por la solución es casi una obligación para evitar caer en errores producidos por valores muy grandes o muy pequeños (p.ej.: divisiones por 0, infinito, etc.). Sin embargo, la acotación de valores genera un problema cuando se

calcula sobre los mismos el coeficiente normalizado de Gini: en los límites de acotación se produce una acumulación de valor (TPR y FPR) que distorsiona dicho coeficiente (aumentando o disminuyendo su valor) y le resta capacidad para ser indicador fiable del coeficiente normalizado de Gini sobre valores no acotados. Dado que este último es el coeficiente que interesa a efectos de análisis más profundo de áreas y/o umbrales e incluso a nivel global del AUC, el coeficiente normalizado de Gini sobre valores acotados es un indicador aceptable, pero, a medida que aumenten los valores estimados que hayan sido acotados, será menos significativo.

La solución propuesta es establecer Evaluator = MSE (Mean Squared Error), ya que así, PG tenderá a obtener soluciones que no precisarán acotación en un elevadísimo porcentaje de casos (100% en la inmensa mayoría de casos). Con ello, el coeficiente normalizado de Gini calculado sobre los valores estimados acotados, obtenido en HL, es un indicador totalmente fiable del mismo coeficiente normalizado de Gini calculado sobre los valores estimados sin acotar y que es preciso para las evaluaciones antedichas.

2.3. Metodología y métrica propuestas

Con estas limitaciones, la metodología que se proponga debe conducir el proceso de parametrización en sus distintos niveles hacia la obtención de las mejores soluciones o programas, tanto si éstos se evalúan a nivel global de la solución, en un área de la misma delimitada por 2 restricciones o en un umbral definido por 1 restricción.

Sobre la base de lo antedicho, la metodología a utilizar será básicamente la siguiente:

- La métrica o métricas utilizadas en cada caso se obtendrán evaluando cada solución (obtenida con los datos del conjunto de training) cuando se aplica al conjunto de test.
- Se establecerá, para cada experimento (conjunto de ejecuciones, cada una de las cuales presenta una o dos soluciones: una “Best training solution” y, adicionalmente, puede presentar una “Best Validation solution”), una zona de interés en la que se desea que se encuentren las soluciones. Esa zona se fijará, de forma totalmente arbitraria, de la siguiente forma:
    - Estableciendo un punto en términos de TPR y porcentaje de verdaderos negativos (True Negative Rate - TNR). Para fijar dicho punto (TNR_0, TPR_0), se podría tomar como referencia los porcentajes de otros estudios para horizontes temporales similares.

- Pasando por el antedicho punto se calcula una recta con pendiente -1. La ecuación de esta recta será: $TPR-TPR_0 = -1 *(TNR-TNR_0)$. Dicha recta cumplirá que TPR + TNR será constante.
- La zona a la derecha de la recta trazada es lo que se denominará zona de interés del experimento (ver figura 2 con explicación más detallada).

Nótese que la pendiente indica la relación exigida entre variaciones porcentuales positivas de una variable, p.ej.: TPR, respecto al punto de referencia y variaciones porcentuales negativas de la otra, en este caso TNR, para mantenerse en la zona de interés.

El fundamento para utilizar una zona de interés en el experimento es que se fuerza a la curva ROC de las soluciones a pasar por una zona determinada (al menos en uno de sus puntos). Con ello se pretende descartar – al menos parcialmente – aquellas soluciones con AUC elevadas debido a "zonas extremas" con elevado peso en las mismas. El objetivo final es utilizar la zona de interés del experimento como filtro para detectar – desde el punto de vista del trabajo - las mejores soluciones de cada experimento.

- Como métricas básicas se utilizarán:
  - El número y rendimiento de todas las soluciones en la zona de interés del experimento.
  - Las X soluciones con mejor rendimiento estimado dentro de la zona de interés del experimento.

  El rendimiento de una solución será su coeficiente normalizado de Gini, calculado sobre los valores estimados acotados, que suponen la aproximación al valor que tomará AUC sobre los valores estimados sin acotar.
- La puntuación de síntesis para la comparación de soluciones se hará con base en una media ponderada de las antedichas métricas.

El carácter estocástico de la PG hace que se pueda obtener – con una cierta parametrización - una muy buena solución por azar. La metodología busca conducir la parametrización hacia la obtención de buenas soluciones de una forma más consistente y más confiable. De ahí que se focalice tanto en el número de soluciones en la zona de interés como en el rendimiento de las X mejores, frente a una focalización exclusiva en el rendimiento de la mejor solución, que sería factible sin el carácter estocástico antedicho de la PG.

3. CONJUNTO DE PRIMITIVAS

Las variables explicativas, variables independientes o terminales, forman, junto a las funciones que pueden utilizarse en la solución, el conjunto de primitivas.

El conjunto de primitivas ha de cumplir dos requisitos (Koza 1992): *Closure y sufficiency.*

De forma muy sucinta, *closure* significa que cada función del conjunto de funciones debería estar bien definida y cerrada para aceptar cualquier argumento que la combinación de primitivas pueda producir.

Sufficiency requiere que el conjunto de primitivas debe ser capaz de expresar una solución al problema.

El requisito de *sufficiency* no es un requisito claro en el caso del problema del fracaso empresarial. Por ello, es preciso contrastar de alguna forma las prestaciones de diferentes conjuntos de primitivas a fin de elegir el más adecuado.

3.1. Variables explicativas

En primer lugar, nos centramos únicamente en las variables explicativas, variables independientes o conjunto de entrada, manteniendo constante el conjunto de funciones utilizado.

Para analizar qué conjunto de variables de entrada es el más adecuado se han diseñado cuatro experimentos, de forma que se pueda disponer de suficiente información del comportamiento de los diferentes conjuntos de variables de entrada en variados entornos temporales de la predicción del fracaso elegidos arbitrariamente. Los experimentos giran en torno a los siguientes horizontes temporales de predicción:

- Modelo m2_2 de predicción 2 años previos al fracaso.
- Modelo m4_4 de predicción 4 años previos al fracaso.
- Modelo m1_3 de predicción de 1 a 3 años previos al fracaso.
- Modelo m4_6 de predicción de 4 a 6 años previos al fracaso.

Los modelos 1_3 y 4_6 son modelos plurianuales. La razón de incorporar estos últimos es la de experimentar con modelos con intervalos más amplios de predicción (p.ej.: modelo que predice el fracaso en el intervalo 1-3 años previos al fracaso). La hipótesis a contrastar es que dichos modelos sacrificarían la posibilidad de predecir cuándo se producirá el evento del fracaso (no se podría determinar si fracasa en el año 1, 2 ó 3

previos al fracaso), a cambio de mejorar las prestaciones del modelo (si fracasa, o no, en el intervalo 1-3 años).

Las condiciones comunes a todos los modelos son las siguientes:

- Se utilizan los datos de 2005-2007. Los datos de 2008 y sucesivos se han reservado para otros fines (persistencia, evaluación en crisis, etc.)
- Se utiliza un único perfil de HeuristicLab para la totalidad de ejecuciones. Este perfil tiene como Evaluator a MSE.
- Cada ejecución (run) evalúa el mismo número de soluciones (100 generaciones y un tamaño de la población de 1.500).
- De cada ejecución se obtienen dos soluciones referidas al conjunto de test: “Best training solution (test)” y “Best validation solution (test)”.
- En todos los casos, las funciones utilizadas son las mismas: aritméticas.
- En el conjunto de entrenamiento se utiliza el emparejamiento aleatorio (e0) denominado así por no existir emparejamiento alguno entre las observaciones de empresas fracasadas y las de empresas no fracasadas. Se eligen las observaciones de empresas fracasadas de forma aleatoria. Por cada observación de empresa fracasada hay una de empresa no fracasada también elegida aleatoriamente.

Con cada uno de los modelos antedichos se han analizado los siguientes conjuntos de variables explicativas (variables de entrada):

- r97_raw: Valores acotados de las 97 ratios sin transformación.
- r97_n1: Valores acotados de las 97 ratios normalizadas con respecto a la media y la desviación estándar de cada ratio en el periodo total de datos.
- r97_log: Valores acotados de las 97 ratios tipificados de acuerdo a la distribución logística con media y desviación estándar de cada ratio en el periodo total de datos.
- r97_n: Valores acotados de las 97 ratios normalizados de acuerdo a la media y la desviación estándar de la ratio en el ejercicio correspondiente.
- r97_nyd: Valores de los numeradores y denominadores utilizados en el cálculo de las ratios y que se expresen en euros, en porcentaje sobre el total de activo.
- r67_var: Variación anual de las ratios que componen ratios 97 y que presentan numerador y denominador en euros y referidos al mismo periodo de tiempo. Estos valores se acotan según un percentil superior y otro inferior, calculados sobre la totalidad de datos utilizables por los modelos (por razones de facilidad

de cálculo y dado que es una acotación, no se ha hecho anualmente, fijándose arbitrariamente dos valores).

En cada experimento se realizan 1000 ejecuciones con cada uno de los seis conjuntos de variables explicativas, por lo que cada experimento consta de 6.000 ejecuciones.

A la hora de analizar los resultados se tienen en cuenta tanto la “Best training solution (test)” como la “Best validation solution (test)” de cada una de las ejecuciones. Se dispone, por lo tanto, de 12.000 soluciones en cada experimento.

Como se explica en el apartado “Métricas”, el análisis se centra en la capacidad de generalización de las soluciones, por lo que los análisis se realizarán sobre el conjunto de test.

Se observa la capacidad para generar soluciones en la denominada zona de interés del experimento (ver “Métricas”). Para ello, después de elegir el punto por el que pasará la recta que delimita la antedicha zona, se siguen los siguientes pasos (Figura 2):

- Las soluciones a la izquierda de la recta trazada se etiquetan como “A”. La idea es que las soluciones a la derecha de la curva – zona de interés – se sitúen en torno al 5% del total.
- Siendo TPR_0 y TNR_0 las coordenadas del punto por el que pasa la recta que delimita la zona de interés, se separan las soluciones que cumplen Y($TPR>TPR_0$; $TNR>TNR_0$). Las soluciones que cumplen este requisito se etiquetan como “B”.
- Se separan las soluciones que cumplen Y($TPR>TPR_0$; $TNR<TNR_0$), que se etiquetan como “C”.
- Por último, se etiquetan como “D” las soluciones restantes, que cumplen Y($TPR<TPR_0$; $TNR>TNR_0$).

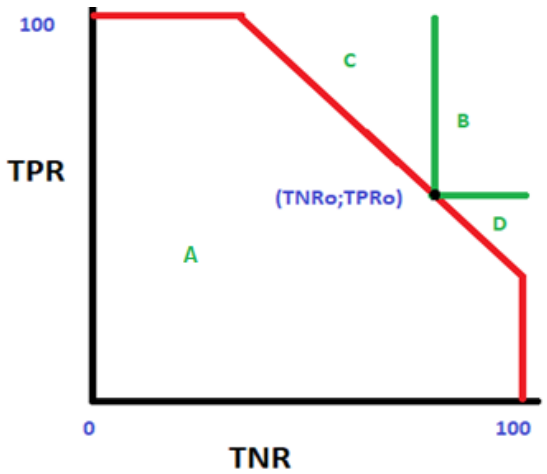

Figura 2: Determinación de la zona de interés de un experimento

Asimismo, para las soluciones etiquetadas como B, C ó D (dentro de la zona de interés), se calcula el rendimiento de la solución, calculado como el coeficiente normalizado de Gini sobre valores estimados acotados.

Nótese que la pendiente de la recta que determina la zona de interés (-1) está favoreciendo la aparición de muchas soluciones en la zona C (cumplen: Y(TPR>TPR_0; TNR<TNR_0)), mientras aparece 1 en la zona D (cumple: Y(TPR<TPR_0; TNR>TNR_0)). Recuérdese que los porcentajes TPR y TNR son los correspondientes al conjunto de test, donde las observaciones correspondientes a empresas fracasadas son totalmente minoritarias (menos del 5% en el mejor de los casos y en algún modelo menos del 1%) respecto a las correspondientes a empresas no fracasadas. La zona D "exige" que un descenso de un X% en TPR, venga acompañado por un aumento en TNR del X% o superior, lo cual en el conjunto de test es desproporcionado por la propia estructura de dicho conjunto. Dado que el estudio prioriza TPR, se mantiene – de forma arbitraria - dicha pendiente en la determinación de la denominada zona de interés del experimento.

Los puntos iniciales, por los que pasarán las rectas de delimitación de las zonas de interés, correspondientes al experimento con cada modelo son los siguientes (Tabla 1):

Tabla 1: Variables explicativas - Puntos de delimitación de zonas de interés de los modelos

modelo	TPR_0	TNR_0	A	B	C	D	Total soluciones	% zona interés
m2_2	88,25%	77,25%	11408	0	592	0	12000	4,93%
m1_3	86,26%	75,26%	11398	0	601	1	12000	5,02%
m4_4	84,07%	69,07%	11392	0	608	0	12000	5,07%
m4_6	82,39%	65,39%	11406	1	593	0	12000	4,95%

Los resultados parciales de cada uno de los modelos son expuestos en la tabla 2 en forma de puntuación:

Tabla 2: Variables explicativas – Puntuaciones simples por modelos (soluciones en zona de interés del experimento)

	m2-2		m1-3		m4-4		m4-6	
	puntuación zona	puntuación 10 mejores soluciones	puntuación zona	puntuación 10 mejores soluciones	puntuación zona	puntuación 10 mejores soluciones	puntuación zona	puntuación 10 mejores soluciones
r97_nyd	0,6524	0,9704	0,0347	0,8429	0,5735	0,9421	0,1104	0,9633
r67_var	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
r97_log	0,9965	0,9981	1,0000	0,9949	0,7225	0,9924	0,9073	0,9976
r97_n	0,9728	1,0000	0,8643	1,0000	0,7029	1,0000	1,0000	1,0000
r97_n1	0,4625	0,9929	0,3120	0,9789	0,5467	0,9696	0,3549	0,9827
r97_raw	1,0000	0,9940	0,1812	0,9729	1,0000	0,9758	0,4199	0,9815

El resumen de puntuaciones ponderadas (ponderando 1 para la puntuación de las soluciones de la zona de interés del experimento y 1 para la puntuación de las 10 mejores soluciones, es decir, calculando la media de ambos valores) para cada una de las opciones y su orden se muestran en la siguiente tabla:

Tabla 3: Variables explicativas – Puntuaciones ponderadas 1:1 por modelos (soluciones en zona de interés del experimento)

modelo	%zona de interés	r97_nyd	r67_var	r97_log	r97_n	r97_n1	r97_raw	Posición 1	Posición 2	Posición 3
m2_2	4,93%	0,8114	0,0000	0,9973	0,9864	0,7277	0,9970	r97_log	r97_raw	r97_n
m1_3	5,02%	0,4388	0,0000	0,9974	0,9322	0,6455	0,5771	r97_log	r97_n	r97_n1
m4_4	5,07%	0,7578	0,0000	0,8575	0,8514	0,7582	0,9879	r97_raw	r97_log	r97_n
m4_6	4,95%	0,5369	0,0000	0,9524	1,0000	0,6688	0,7007	r97_n	r97_log	r97_raw

Los resultados finales (posiciones 1, 2 y 3 para cada modelo) no se alterarían aun si la ponderación fuese 1 para la puntuación de las soluciones de la zona de interés del experimento y 2 para la puntuación de las 10 mejores soluciones.

La transformación más equilibrada en términos de los cuatro modelos analizados y según los criterios antedichos, es r97_log.

3.2. Funciones

Como se ha indicado, en los experimentos anteriores (tendentes a facilitar la decisión sobre qué conjuntos de variables de entrada aparecían como los más prometedores) se utilizó siempre la que se ha denominado gramática reducida, compuesta por las funciones aritméticas (suma, resta, multiplicación, división).

Para decidir acerca del conjunto de primitivas más adecuado al problema, se añaden tres conjuntos de funciones al análisis.

- La denominada gramática normal, que consta de las siguientes funciones:
  - Aritméticas (suma, resta, multiplicación y división)
  - Condicionales:
    - IfThenElse.
    - Comparadores (mayor que, menor que)
    - Booleanas (Y, O, NO)
- La denominada gramática ampliada, que añade a la gramática normal las siguientes funciones:
  - Potencias (x^2, x^3, x^a, $x^{1/2}$, $x^{1/3}$, $x^{1/a}$)

    - Trigonométricas (seno, coseno, tangente y tangente hiperbólica)
    - Exponenciales y logarítmicas (ab^x, $\log_{10} x$)
- La denominada gramática ampliada sin condicionales, con las siguientes funciones: Aritméticas, potencias, trigonométricas, exponenciales y logarítmicas.

Es preciso advertir de la existencia de un problema potencial con las gramáticas normal y ampliada. Con el conjunto de funciones aritméticas es relativamente fácil que una solución genere una amplia colección de valores estimados (dado que cada observación cuenta con 97 valores de entrada que oscilan entre 0 y 1 y las funciones aritméticas utilizadas no facilitan la agrupación de los valores de salida, probablemente se obtenga un valor estimado diferente para cada observación). Por el contrario, tanto la gramática normal como la gramática ampliada incluyen, entre las funciones que pueden utilizarse en la construcción de soluciones, las funciones condicionales. Ello puede dar lugar a que los valores estimados se reduzcan drásticamente. Este hecho no puede ser detectado por ninguna métrica salvo analizando la solución y los valores estimados.

Si, por ejemplo, la clase positiva tiene 170 elementos, lo ideal es que la solución facilite los 170 umbrales que permitan conocer cada punto de la evolución de TPR con su correspondiente FPR. De esta forma los análisis pueden ser completos.

Conscientes de este problema potencial se realiza un primer experimento cuyo objetivo es valorar qué gramáticas se perfilan como más adecuadas en la predicción del fracaso empresarial. Para ello se configura el experimento con las siguientes características básicas:

- Modelo m2_2 de predicción 2 años previos al fracaso.
- Se utilizan los datos de 2005-2007.
- Se utiliza un único perfil de HeuristicLab para la totalidad de ejecuciones. Este perfil tiene como Evaluator a MSE.
- Se analizan los siguientes conjuntos de primitivas (variables explicativas y funciones):
    - r97_log: Valores acotados de las 97 ratios tipificados de acuerdo a la distribución logística con media y desviación estándar de cada ratio en el periodo total de datos.
        - Con gramática reducida
        - Con gramática normal
        - Con gramática ampliada
        - Con gramática ampliada sin condicionales

- Cada ejecución (run) evalúa el mismo número de soluciones (100 generaciones y un tamaño de la población de 1.500).
- De cada ejecución se obtienen dos soluciones referidas al conjunto de test: “Best training solution (test)” y “Best validation solution (test)”.
- En el conjunto de entrenamiento se utiliza el emparejamiento aleatorio (e0). Se eligen las observaciones de empresas fracasadas de forma aleatoria. Por cada observación de empresa fracasada hay una de empresa no fracasada también elegida aleatoriamente.

Los resultados de este experimento son los siguientes (tabla 4):

Tabla 4: Funciones – Puntuación simple modelo 2-2

m2-2	puntuación zona	Puntuación 10 mejores soluciones
gramReducida	1,0000	1,0000
gramNormal	0,1488	0,3453
gramAmpliada	0,0000	0,0000
gramAmpliadaSinCondicional	0,7169	0,8935

Se observa que, con independencia de la ponderación que se otorgue a la puntuación de las soluciones de la zona de interés del experimento y a la puntuación de las 10 mejores soluciones, la gramática normal y la gramática ampliada son las peores alternativas de las contempladas. A la vista de ello se opta por continuar la parametrización evaluando únicamente las opciones de gramática reducida y gramática ampliada sin condicionales.

Un problema similar al antedicho, en el que la solución puede ser no totalmente válida (puesto que tiende a agrupar los valores estimados, imposibilitando el análisis de un área de dicha solución) se produce con la denominada gramática ampliada sin condicionales, debido básicamente a dos circunstancias:

- Que entre las funciones que pueden utilizarse, existen algunas cuyo valor de salida está acotado, (p.ej.: seno x, cuyo valor de salida oscila entre -1 y +1).
- Que entre las funciones que pueden utilizarse, existen algunas que pueden dar como resultado un error. Ello genera un valor estimado de la solución que es la media de los límites de acotamiento de los valores estimados (p.ej.: logaritmo de un número negativo, que se puede haber producido por la simple resta de dos ratios).

Aunque esté problema generalmente es de menor magnitud que el que generan las gramáticas con condicionales, tampoco es detectable salvo con un análisis pormenorizado de la solución.

Para analizar qué conjunto de funciones (de entre los dos antedichos) es el más adecuado se han diseñado cuatro experimentos en los mismos entornos temporales de la predicción del fracaso elegidos en el caso del análisis del conjunto de variables.

- Modelo m2_2 de predicción 2 años previos al fracaso.
- Modelo m4_4 de predicción 4 años previos al fracaso.
- Modelo m1_3 de predicción de 1 a 3 años previos al fracaso.
- Modelo m4_6 de predicción de 4 a 6 años previos al fracaso.

Las condiciones comunes a todos los modelos son los indicados anteriormente.

Con cada uno de los modelos antedichos se han analizado los siguientes conjuntos de primitivas (variables explicativas y funciones):

- r97_log: Valores acotados de las 97 ratios tipificados de acuerdo a la distribución logística con media y desviación estándar de cada ratio en el periodo total de datos.
    - Con gramática reducida
    - Con gramática ampliada sin condicionales

Recuérdese que, tal como se vio en el caso de las variables explicativas, el análisis se centra en la capacidad de generalización de las soluciones, por lo que se realiza sobre el conjunto de test. Los pasos seguidos son los antedichos.

Los resultados obtenidos en cada modelo (número de soluciones en la zona de interés del experimento y rendimiento de las 10 mejores soluciones en dicha zona de interés) muestran que en dos de los modelos (m2_2 y m4_6) el conjunto de funciones denominado gramática reducida presenta los mejores resultados, mientras en los otros dos modelos (m1_3 y m4_4) es el conjunto de funciones denominado gramática ampliada sin condicionales el que presenta los mejores resultados. Los resultados son similares si se emplean ponderaciones similares como si se pondera el doble al rendimiento de las 10 mejores soluciones.

En esta tesitura y ante los potenciales problemas que plantea la gramática ampliada sin condicionales (soluciones no analizables parcialmente en algún tramo de las mismas), se opta por el conjunto de funciones denominado gramática reducida.

Sobre la base de lo antedicho, el conjunto de primitivas que se califica como más idóneo para la predicción del fracaso empresarial planteada en este estudio, es el compuesto de:

- r97_log: Valores acotados de las 97 ratios tipificados de acuerdo a la distribución logística con media y desviación estándar de cada ratio en el periodo total de datos.
- gramReducida: Funciones aritméticas (suma, resta, multiplicación y división).

Por una parte, la transformación r97_log proporciona valores entre 0 y 1 para cada uno de las ratios. De otra parte, la utilización de únicamente las funciones aritméticas tiene dos efectos que se analizarán más adelante:

- Las soluciones son más interpretables, en el sentido de que los resultados de la solución que aporta la inteligencia artificial pueden ser más entendidos por humanos expertos, en línea con la denominada inteligencia artificial explicable (XAI).
- Las soluciones son más sencillas y ello, como señala Finlay (citado por du Jardin y Séverin 2012), afecta directamente a la estabilidad del poder predictivo de una solución a lo largo del tiempo, por cuanto más complejo es un clasificador, más a menudo debe ser reestimado.

4. CONJUNTO DE ENTRENAMIENTO

A la hora de configurar el conjunto de entrenamiento es preciso recordar que, en el problema de clasificación del fracaso empresarial, las poblaciones están totalmente desequilibradas. Existe una clase mayoritaria (observaciones de las empresas no fracasadas) que supera con mucho a la clase minoritaria (observaciones de las empresas fracasadas).

Existen varias alternativas para conformar el conjunto de entrenamiento. A continuación, exponemos las opciones utilizadas en este estudio.

4.1. Emparejamiento en el conjunto de entrenamiento

Alaka y col. (Alaka et al. 2018) concluyen que el 80% de los estudios sobre fracaso empresarial utilizan conjuntos de entrenamiento con porcentajes clase mayoritaria / clase minoritaria entre 50%-50% y 60%-40%. En nuestro caso nos hemos decantado por la proporción 50%-50%, por lo que el conjunto de entrenamiento tendrá el mismo

número de observaciones correspondientes a empresas fracasadas que a empresas no fracasadas.

4.2. Selección de las observaciones del conjunto de entrenamiento

Es frecuente, en estudios de fracaso empresarial, seleccionar las observaciones de empresas fracasadas y no fracasadas de los conjuntos de entrenamiento atendiendo al tamaño, edad, sector, etc. De cualquier forma, tal como señala Palepu (Palepu 1986), el no seleccionar una muestra de forma aleatoria puede presentar al menos dos importantes inconvenientes: Sobreestimar la capacidad predictiva del modelo y dificultar la generalización al resto de la población.

Ante esta disyuntiva, se ha procedido a comparar distintos métodos de selección de las observaciones del conjunto de entrenamiento, incorporando a esta comparación métodos habituales en otros campos de Machine Learning.

4.2.1. Selección aleatoria

En este caso, tanto las observaciones de empresas fracasadas, como las correspondientes a empresas no fracasadas, se eligen de forma aleatoria. Se denomina *e0*.

4.2.2. Selección no aleatoria

En este caso, las observaciones correspondientes a las empresas fracasadas se eligen aleatoriamente. Para seleccionar las observaciones de empresas no fracasadas se manejan las siguientes alternativas:

4.2.2.1. Con base en variables externas

Por cada observación correspondiente a empresas fracasadas, existe una observación de empresa no fracasada con las siguientes características, según la alternativa elegida:

- Del mismo ejercicio económico y misma actividad. Se denomina *e2* (dos variables emparejando observaciones de empresas fracasadas y no fracasadas).
- Del mismo ejercicio económico, misma actividad y mismo grupo de total activo. Se denomina *e3*.

- Del mismo ejercicio económico, misma actividad, mismo grupo de total activo y mismo grupo de ingresos de explotación. Se denomina *e4*.

4.2.2.2. Con base en las variables explicativas

Las variables de las empresas no fracasadas se escogen según las siguientes alternativas:

- NearMiss-1: Se retienen aquellos puntos de la clase mayoritaria (sin problemas) cuya distancia media a los k más cercanos de la clase minoritaria (con problemas) sea menor. Se ha utilizado k=3. Se denomina *e5*.
- NearMiss-2: Se mantienen aquellos puntos de la clase mayoritaria (sin problemas) cuya distancia media a los k más alejados de la clase minoritaria (con problemas) sea menor. Se ha utilizado k=3. Se denomina *e6*.
- NearMiss-3: Selecciona k vecinos más cercanos de la clase mayoritaria (sin problemas) para cada punto de la clase minoritaria (con problemas). Dado que se pretende un conjunto de entrenamiento con igual número de elementos de cada una de las clases, el valor de k será 1. Se denomina *e7*.

4.3. Resultados

Los resultados – aun no siendo totalmente concluyentes - muestran que el emparejamiento aleatorio *e0* es el que presenta mejores resultados en el conjunto de los ochos modelos analizados. En una breve síntesis, *e0* es la técnica con mejores resultados en 3 de los 8 modelos analizados, siendo la segunda mejor técnica en otros 2 modelos y la tercera en otro de los 8 (resultados obtenidos como la ponderación del número de soluciones en la denominada zona de interés - aproximadamente el 5% de las soluciones - y mejor rendimiento de las 10 mejores soluciones obtenido por medio del coeficiente normalizado de Gini sobre los valores estimados). Estas cifras no se alcanzan con otra de las técnicas (*e2*, *e3* o *e4*). Los resultados resumidos (con ponderación 1 para el número de soluciones en la zona de interés del experimento y 2 para los resultados de las 10 mejores soluciones en dicha zona) son los expuestos en la siguiente tabla.

Tabla 5: Selección observaciones conjunto de entrenamiento - Puntuaciones ponderadas 1:2 por modelos (soluciones en zona de interés del experimento)

modelo	%zona de interés	e0	e2	e3	e4	1	2	3
m1_1	5,00%	0,4654	0,0000	0,1171	1,0000	e4	e0	e3
m2_2	5,06%	1,0000	0,2327	0,2331	0,0000	e0	e3	e2
m3_3	5,04%	0,0000	0,0897	0,5479	1,0000	e4	e3	e2
m1_3	5,05%	0,2023	0,6720	0,7731	0,2197	e3	e2	e4
m4_4	4,99%	0,4844	0,2649	0,8664	0,1343	e3	e0	e2
m4_6	4,95%	1,0000	0,2416	0,0000	0,1426	e0	e2	e4
m8_8	5,00%	0,2692	0,4897	0,0012	1,0000	e4	e2	e0
m7_9	4,94%	1,0000	0,5934	0,8195	0,0000	e0	e3	e2

Los resultados anteriores no se alteran si se igualan las antedichas ponderaciones.

En consecuencia, para conformar el conjunto de entrenamiento, tanto las observaciones de empresas fracasadas, como las correspondientes a empresas no fracasadas, se elegirán de forma aleatoria.

5. OTRAS PARAMETRIZACIONES RELEVANTES

Hay al menos dos parametrizaciones relevantes en esta fase de la modelización que no se abordan en este documento por falta de espacio:

5.1. Tamaño y partición del conjunto de entrenamiento.

Referido a cómo se fija el tamaño del conjunto de entrenamiento y cómo se conforman los subconjuntos del conjunto de entrenamiento (subconjunto de aptitud -*fitness* – y validación – *validation*).

5.2. Reducción de la dimensión del conjunto de variables de entrada.

Aplicando PG, sin recurrir a algoritmos externos.

6. COMPARACION CON REFERENCIA EXTERNA

En este capítulo se trata de cotejar los resultados obtenidos (de los modelos anuales y con las únicas parametrizaciones vistas hasta el momento: Conjunto de primitivas,

emparejamiento y selección del conjunto de entrenamiento, tamaño y partición del conjunto de entrenamiento y reducción de dimensionalidad), con una referencia externa a fin de tener una primera aproximación a la bondad de dichos resultados.

6.1. Situación de partida de los modelos anuales

Hasta el momento, se han abordado las siguientes decisiones:

- Selección del conjunto de primitivas (r97_log_gred)
- Emparejamiento de observaciones de empresas fracasadas y no fracasadas del conjunto de entrenamiento (50% fracasadas y 50% no fracasadas).
- Selección de las observaciones de empresas no fracasadas y fracasadas dentro de dicho conjunto de entrenamiento (e0: ambos conjuntos se seleccionan de forma aleatoria).
- Tamaño del conjunto de entrenamiento, que se ha fijado igual al número de observaciones de empresas fracasadas en el modelo. Dentro de dicho conjunto, habrá igual número de observaciones de empresas fracasadas y no fracasadas.
- Partición del conjunto de entrenamiento entre los subconjuntos de aptitud y de validación (fitVal_100-30) es decir, el subconjunto de aptitud (*fitness*) será el 100% del conjunto de entrenamiento y el de evaluación (*validation*) el 30% de dicho conjunto.

Se dispone de un experimento con 1.000 ejecuciones (cada una de las cuales evalúa el mismo número de soluciones: 100 generaciones y un tamaño de la población de 1.500), para cada uno de los nueve modelos anuales analizados (m1_1, m2_2, …, m8_8, m9_9). Las características comunes de dichos experimentos son las que se han visto previamente, referidas a los puntos antedichos sobre los que se ha ido tomando decisiones. Recuérdese que los datos utilizados corresponden a los ejercicios 2005-2007 ambos inclusive.

6.2. Algunas consideraciones sobre los modelos anuales

Es preciso recordar algunos aspectos que deben tenerse en cuenta a la hora de analizar la comparación:

- Los modelos realizados por en este trabajo están concebidos para poder analizarse a lo largo de la totalidad de la curva ROC. Ello, como se ha visto previamente (apartados de Metodología y métricas y Conjunto de primitivas -

Funciones), conlleva ciertas restricciones respecto al evaluador y a las funciones a utilizar.

- Se intentan evitar soluciones con buenos datos AUC-ROC obtenidas con umbrales inaceptables por sus porcentajes de TP y TN, con lo que se impone que las soluciones adoptadas se encuentren en la denominada zona de interés de un experimento (ver apartado de Metodología y métricas), lo que conlleva que cada solución ha de pasar a la derecha de un umbral de mínimos.
- La elección de un perfil común de parametrización para la totalidad de modelos del trabajo, sacrificando parametrizaciones que podrían ser más adecuadas para un modelo anual en concreto.
- Las soluciones de los modelos están sin optimizar (no se ha ahondado en múltiples aspectos de parametrización - p.ej.: número de soluciones evaluadas, tamaño de la solución, … - ni en la simplificación de las soluciones finales).

Los antedichos aspectos suponen de facto una minoración en las AUC de las soluciones de los modelos obtenidos por PG.

6.3. Selección de la referencia externa

La referencia utilizada para evaluar si los resultados obtenidos hasta el momento son aceptables es el trabajo de Altman y col. (Altman et al. 2015) en el que se contemplan modelos con predicciones hasta 10 años previos al fracaso, con variables financieras únicamente y con variables financieras y no financieras. La métrica de comparación será AUC calculada sobre el conjunto de test.

La elección de esta referencia se basa en el hecho de que es uno de los pocos estudios que permite una comparación homogénea en diferentes horizontes temporales (1 a 10 años previos al fracaso). En el mencionado estudio, los autores utilizan el análisis de regresión logística para, a partir de los datos del ejercicio 2003, realizar predicciones de 1 a 10 años previos del fracaso empresarial. Por otra parte, tanto el tamaño de los conjuntos de test como de entrenamiento son similares a los utilizados en este estudio (en el trabajo de Altman y col. (2015), el tamaño del conjunto de test es de 23.469 y el del conjunto de entrenamiento oscila entre 50 y 174, mientras en este trabajo, el tamaño del conjunto de test es de 22.330 y el de entrenamiento oscila entre 82 y 340 en lo que se refiere a los modelos anuales objeto de comparación).

6.4. Resultados

En la tabla siguiente (tabla 6) se muestra un resumen de los resultados obtenidos. Por una parte, para los modelos anuales de este trabajo, obtenidos por PG, se muestran las AUC correspondientes a los límites inferior y superior de las 5 mejores soluciones contenidas en la zona de interés del experimento (5% de las soluciones), a fin de contextualizar la bondad de las mismas. Por otra parte, se muestran los resultados del trabajo de Altman y col. (2015). En ambos casos, los modelos utilizan únicamente variables financieras y la evaluación de los mismos se realiza sobre el conjunto de test.

Tabla 6: Comparación con referencia externa (test): Primeros resultados

	AUC de max.Gini en selección 5 mejores soluciones	AUC de min.Gini en selección 5 mejores soluciones	AUC (Altman) Financial
m1_1	94,29%	93,62%	88,13%
m2_2	91,53%	90,88%	81,69%
m3_3	85,60%	85,49%	78,30%
m4_4	85,78%	85,51%	79,39%
m5_5	80,84%	80,31%	76,47%
m6_6	79,02%	78,60%	66,83%
m7_7	76,08%	74,26%	72,94%
m8_8	67,39%	66,77%	73,29%
m9_9	70,21%	68,27%	68,95%

Como puede observarse, los resultados de los modelos obtenidos en este estudio con PG – comparados en términos de AUC-ROC sobre el conjunto de test – mejoran claramente, modelo a modelo, a los obtenidos en el antedicho estudio, con excepción del modelo correspondiente al horizonte temporal de 8 años previos al fracaso. Nótese que, en los modelos con horizonte temporal de 1 a 7 años, incluso las AUC mínima de las 5 mejores soluciones en la zona de interés seleccionada de los modelos de este estudio, superan a las AUC de la referencia externa.

7. CONCLUSIONES

Los principales aspectos a destacar de este trabajo son básicamente los siguientes:

- La parametrización es un aspecto clave en la aplicación de la técnica de PG de cara a la obtención de las mejores soluciones según los objetivos fijados. Tómese, p.ej.: el conjunto de variables de entrada para el modelo 2 años previos al fracaso, permaneciendo constantes el resto de parámetros. La diferencia del

AUC de las 10 mejores soluciones para la mejor alternativa (r97_log) supera en un 1,27% a la penúltima (r97_nyd), dado que la última alternativa utilizada no tiene soluciones en la zona de interés del experimento. Son variaciones pequeñas, pero que al ir acumulando la parametrización hace que las opciones de parametrización sí sean relevantes en el resultado final.

- La estrategia de parametrización propuesta de los modelos de predicción del fracaso empresarial realizados con Programación Genética (PG) y basada en el establecimiento de un área de interés, es adecuada, a la vista de los resultados obtenidos, no únicamente comparando alternativas de parametrización, sino comparando resultados obtenidos con referencias externas.

De cualquier forma, es preciso indicar que en este estudio: 1) la parametrización se hace de forma global (p.ej.: se escoge un conjunto de primitivas y un emparejamiento para la totalidad de los modelos), cuando podría hacerse modelo a modelo escogiendo la mejor combinación no en términos medios sino en términos individuales, lo que podría mejorar más los resultados y 2) la parametrización es incompleta, puesto que en el trabajo no se abordan aspectos tan relevantes como el número de soluciones evaluadas (función del número de generaciones y del tamaño de la población), el tamaño de las soluciones (profundidad y longitud máximas del árbol que representa la solución), etc.

BIBLIOGRAFIA

Alaka, Hafiz A., Lukumon O. Oyedele, Hakeem A. Owolabi, Vikas Kumar, Saheed O. Ajayi, Olugbenga O. Akinade, y Muhammad Bilal. 2018. «Systematic review of bankruptcy prediction models: Towards a framework for tool selection». *Expert Systems with Applications* 94:164-84. doi: 10.1016/j.eswa.2017.10.040.

Altman, Edward I., Małgorzata Iwanicz-Drozdowska, Erkki Laitinen, y Arto Suvas. 2015. «Financial and non-financial variables as long-horizon predictors of bankruptcy». *SSRN Electronic Journal*. doi: 10.2139/ssrn.2669668.

Beade, A., J. Santos, y M. Rodríguez López. 2022. «Predicción del fracaso empresarial por medio de Programación Genética: Estrategias de parametrización». en *IX Jornada Internacional AECA de Valoración, Financiación y Gestión de Riesgos*.

Dodd, Lori E., y Margaret S. Pepe. 2003. «Partial AUC Estimation and Regression». *Biometrics* 59(3):614-23.

Hand, David J. 2009. «Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve». *Machine Learning* 77(1):103-23. doi: 10.1007/s10994-009-5119-5.

du Jardin, Philippe, y Eric Séverin. 2012. «Forecasting Financial Failure Using a Kohonen Map: A Comparative Study to Improve Model Stability over Time». *European Journal of Operational Research* 221(2):378-96. doi: 10.1016/j.ejor.2012.04.006.

Koza, John R. 1992. *Genetic programming: on the programming of computers by means of natural selection*. Cambridge [etc.]: The MIT Press.

McClish, D. K. 1989. «Analyzing a Portion of the ROC Curve». *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 9(3):190-95. doi: 10.1177/0272989X8900900307.

Palepu, Kg. 1986. «Predicting Takeover Targets - a Methodological and Empirical-Analysis». *Journal of Accounting & Economics* 8(1):3-35. doi: 10.1016/0165-4101(86)90008-X.

Poli, Riccardo, W. B. (William B.). Langdon, Nicholas F. McPhee, y John R. Koza. 2008. *A Field Guide to Genetic Programming*. [S.l.] : [Lulu Press], lulu.com.

Wagner, Stefan, Gabriel Kronberger, Andreas Beham, Michael Kommenda, Andreas Scheibenpflug, Erik Pitzer, Stefan Vonolfen, Monika Kofler, Stephan Winkler, Viktoria Dorfer, y Michael Affenzeller. 2014. «Architecture and design of the HeuristicLab optimization environment». Pp. 197-261 en *Advanced Methods and Applications in Computational Intelligence*. Vol. 6, *Topics in Intelligent Engineering and Informatics*, editado por R. Klempous, J. Nikodem, W. Jacak, y Z. Chaczko. Springer.