Recuadro 2

PRUEBA DE CONCEPTO REALIZADA EN LA CENTRAL DE BALANCES PARA LA DETECCIÓN DE ANOMALÍAS Y LA IMPUTACIÓN DE VALORES CON TÉCNICAS DE INTELIGENCIA ARTIFICIAL

La Central de Balances (CB) recopila datos anuales de carácter contable de las sociedades no financieras mediante dos fuentes distintas, que determinan la creación de dos bases de datos muy diferenciadas: una de colaboración voluntaria por parte de las entidades (denominada «CBA»), donde el estrato de grandes empresas tiene un mayor peso y donde se depuran manualmente los cerca de 11.000 cuestionarios que la componen; y una segunda base de datos procedente del Registro Mercantil («CBB», según la terminología de la CB), que anualmente recibe las cuentas anuales de cerca de un millón de sociedades y donde las pymes quedan ampliamente representadas.

Cada cuestionario que forma parte de la base de datos CBB, debido al elevado número de empresas que la integran, se somete a un conjunto de contrastes automáticos (que verifican, básicamente, que se cumplen unas reglas lógico-aritméticas), que contribuyen a depurar la información y a determinar su calidad y coherencia interna (para más información, puede consultarse el epígrafe 2.3 del *Suplemento metodológico* de la presente monografía). Aproximadamente, el 20 % de la información es clasificada finalmente como no publicable, por su baja calidad. Uno de los motivos que provoca que los cuestionarios no sean perfectos es la ausencia de datos para determinados desgloses de información, impidiendo así el cuadre de la información contable. Otro grupo de cuestionarios es descartado porque, aunque tiene datos, no supera determinados umbrales de calidad para partidas contables concretas.

Una de las vías que se abren para aprovechar la información de esta muestra de empresas es la utilización de técnicas de inteligencia artificial. En particular, las técnicas de aprendizaje automático *(machine learning)* pueden complementar los procesos de detección de casos anómalos *(outliers)* y permitir procesos de imputación ante ausencia de datos (datos *missing)* mediante la identificación de patrones de comportamiento entre los datos con los que afinar el sistema de filtrado automático. De esta forma, se maximizaría el número de empresas con información coherente y se minimizaría el riesgo de introducir cuestionarios anómalos en la muestra.

Este ha sido el objetivo de una prueba de concepto (PoC, por sus siglas en inglés) llevada a cabo por la Central de Balances con la colaboración del Departamento de Sistemas de Información del Banco de España y del Instituto de Ingeniería del Conocimiento (IIC), asociado a la Universidad Autónoma de Madrid.

Preprocesamiento de los datos

En esta prueba de concepto se contó con una muestra de más de seis millones de cuestionarios de formato abreviado (según la terminología del depósito oficial de cuentas en los Registros Mercantiles), con distintos grados de calidad y correspondientes a los ejercicios 2008 a 2017. Se seleccionaron 94 variables contables, de acuerdo con el criterio de los expertos de negocio, evitando así las partidas que fuesen combinaciones lineales (sumas) de otras.

Además, se normalizaron los datos para evitar que el algoritmo detectase como anómalas a las sociedades solo por su mayor tamaño. Esta normalización se llevó a cabo dividiendo, en cada cuestionario, las partidas del balance por su total activo, y las partidas de la cuenta de pérdidas y ganancias, por su cifra de ventas.

Imputación de valores *missing*

Para este caso, se buscó un algoritmo que realizara *imputaciones no lineales ni predefinidas,* evitando que fuera la decisión humana la que influyera en la definición del modelo al que debería adaptarse el dato (a diferencia de algunos modelos de imputación clásica, que suelen utilizar la mediana, regresiones o medias móviles).

Del conjunto de algoritmos probados[1], el que mejor resultados ofreció fue *Ensembled Regression Chains* (ERC), método que realiza regresiones sucesivas sobre variables explicativas elegidas y ordenadas al azar (esto hace que sea computacionalmente muy costoso para una PoC, lo que ha impedido usar todos los datos disponibles).

Para obtener un adecuado modelo de predicción, el algoritmo trabajó con el 80 % de los cuestionarios que, de acuerdo con los procedimientos aritmético-lógicos, se consideran perfectos (conjunto de entrenamiento *train).* El restante 20 % formó el conjunto del test, en el que se vaciaron aleatoriamente algunos de sus datos (perforación) para poder evaluar en una fase posterior la calidad de la imputación realizada. El modelo de predicción entrenado

1 Se probaron otros métodos; en concreto, *autoencoders* y *multivariable imputation* (MICE), pero en la fase de validación quedaron descartados por su menor precisión.

se aplicó a los formularios incompletos *(missing)* para estimar el valor pertinente que se había de imputar.

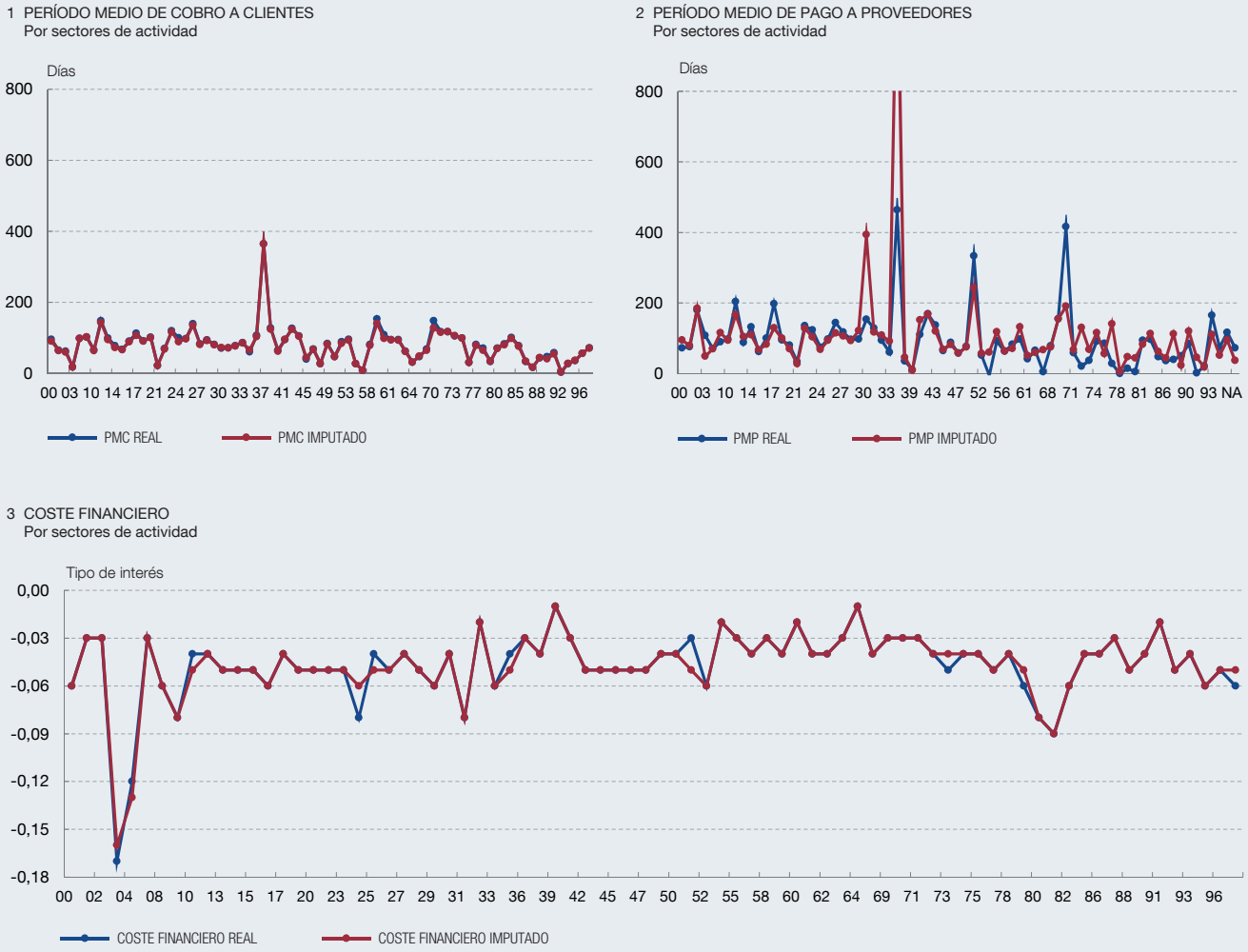
Los resultados del empleo de este algoritmo resultan aceptables para las variables contables en las que se han podido utilizar muchos datos para «entrenarlo». Con el fin de evaluar la calidad de las imputaciones, se realizó un ejercicio que consistió en recalcular tres indicadores contables utilizando los datos imputados y se compararon con los valores reales del conjunto del test. Los indicadores analizados fueron el período medio de cobro (PMC), el período medio de pago (PMP) y el coste financiero, que se representan en el gráfico 1, por sector de actividad, y donde se observa bastante similitud entre las ratios calculadas con datos reales y las obtenidas con datos imputados.

Detección de anomalías

El segundo objetivo de la PoC era incorporar la multidimensionalidad en el análisis de detección de anomalías o *outliers,* esto es, usar una técnica que calcule valores extremos teniendo en cuenta las posibles relaciones entre las variables.

Como se puede observar en el gráfico 2, si se usa una técnica con perspectiva unidimensional, las empresas representadas mediante puntos rojos fuera del rectángulo quedarían marcadas como *outliers,* porque son casos extremos vistos desde las perspectivas vertical y horizontal. Sin embargo, dejarían de ser considerados anómalos si se tuviera en cuenta la estructura de la nube de puntos.

Gráfico 1
COMPARACIÓN DEL COSTE FINANCIERO Y DEL PERÍODO MEDIO DE COBRO Y DE PAGO UTILIZANDO DATOS IMPUTADOS FRENTE A DATOS REALES

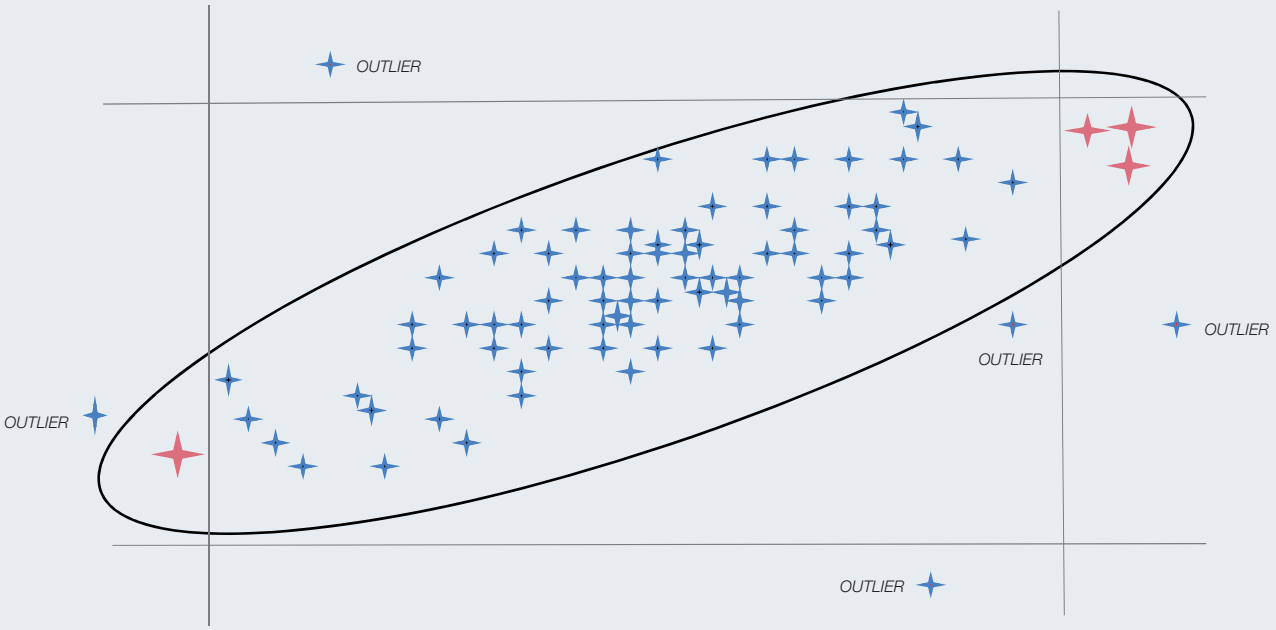


FUENTE: Banco de España.

Recuadro 2

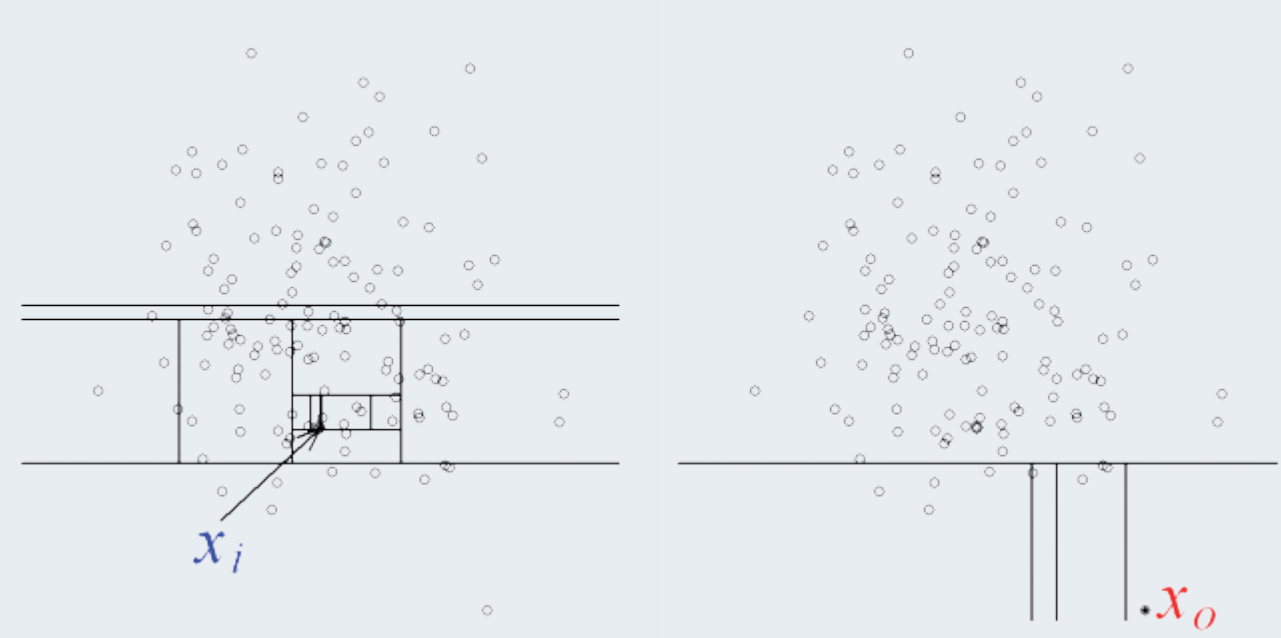
PRUEBA DE CONCEPTO REALIZADA EN LA CENTRAL DE BALANCES PARA LA DETECCIÓN DE ANOMALÍAS Y LA IMPUTACIÓN DE VALORES CON TÉCNICAS DE INTELIGENCIA ARTIFICIAL (cont.)

Gráfico 2
SISTEMA MULTIDIMENSIONAL DE DETECCIÓN DE OBSERVACIONES ANÓMALAS O *OUTLIERS*



FUENTE: Banco de España.

Gráfico 3
ISOLATION FOREST COMO TÉCNICA EN LA DETECCIÓN DE *OUTLIERS* MEDIANTE CORTES ALEATORIOS DEL ESPACIO MUESTRAL



FUENTES: Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (diciembre de 2008).

Recuadro 2

PRUEBA DE CONCEPTO REALIZADA EN LA CENTRAL DE BALANCES PARA LA DETECCIÓN DE ANOMALÍAS Y LA IMPUTACIÓN DE VALORES CON TÉCNICAS DE INTELIGENCIA ARTIFICIAL (cont.)

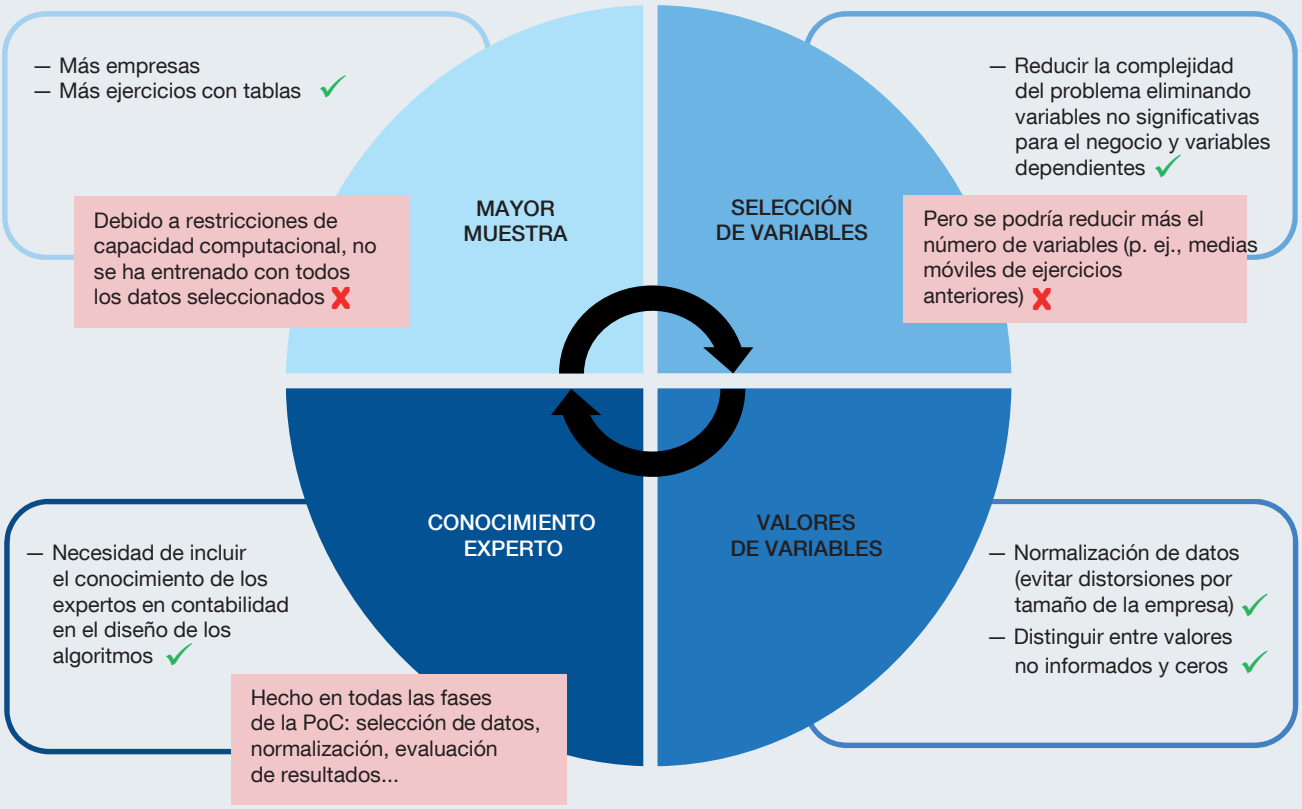
La técnica empleada ha sido *Isolation Forest.* Este tipo de algoritmos se caracteriza por cortar aleatoriamente muchas veces el espacio de puntos, asignando la característica de anómalo a aquel punto que necesita muy pocos cortes para aislarse respecto del resto de las observaciones del espacio muestral. En nuestro caso, cada sociedad es un punto y sus variables contables son las coordenadas. Aunque, en aras de la claridad, el gráfico expresa solo dos dimensiones, el problema que se planteaba tenía 94 dimensiones (el número de variables analizadas).

Como se puede apreciar en el gráfico 3, aislar el punto *outlier* de la imagen derecha requiere menos cortes que aislar el punto de la imagen izquierda.

Usando este procedimiento, se obtiene un *scoring* de anomalía de entre 0 (poco anómalo) y 1 (muy anómalo), que permite la construcción de una matriz de confusión al enfrentar estos niveles con los tradicionales de coherencia utilizados por la Central de Balances.

Los resultados de esta prueba fueron bastante favorables, ya que pusieron de manifiesto que el 94 % de las empresas clasificadas como coherentes por la Central de Balances se sitúa, para el algoritmo, en un intervalo de anomalía de entre 0 y 0,2. El conocimiento de los casos que resultan ser posibles falsos positivos y/o negativos permitirá mejorar el sistema de filtrado automático aplicado en la Central de Balances para el control de calidad de la información. Para un mejor entendimiento del resultado del algoritmo sobre el nivel

Esquema 1
LECCIONES APRENDIDAS DE LA PoC DE CENTRAL DE BALANCES



FUENTE: Banco de España.

Recuadro 2

PRUEBA DE CONCEPTO REALIZADA EN LA CENTRAL DE BALANCES PARA LA DETECCIÓN DE ANOMALÍAS Y LA IMPUTACIÓN DE VALORES CON TÉCNICAS DE INTELIGENCIA ARTIFICIAL (cont.)

de *scoring,* los valores de Shapley[2] se han revelado en esta PoC como una herramienta potencialmente útil.

Algunas lecciones aprendidas

Durante el desarrollo de esta PoC hubo un aprendizaje continuo sobre aspectos esenciales que condicionan de manera significativa el mayor o menor éxito de los resultados. La normalización de datos y la integración del conocimiento experto en todas las fases de la PoC han sido elementos muy relevantes para alcanzar unos resultados satisfactorios. Otros aspectos, como la correcta dimensión de la muestra o una precisa selección de las variables relevantes, fueron progresivamente mejorados, pero aún persisten algunas ineficiencias que podrían superarse en el futuro, de cara a una próxima puesta en producción de este proyecto.

2 Los valores o ratios de *Shapley,* herramienta habitual de la XAI *(eXplanaible Artificial Intelligence),* permiten aplicar una línea de investigación innovadora en el campo de la inteligencia artificial (y prácticamente inédita en el ámbito de la contabilidad) para conocer la importancia que tiene cada partida contable en el *scoring* final de anomalía y para descubrir patrones contables difícilmente observables de otro modo.