BANCO DE ESPAÑA
Eurosistema

USO INTERNO

TÉCNICAS DE *MACHINE LEARNING* APLICADAS A LA IMPUTACIÓN Y CONTROL DE CALIDAD DE LOS MICRODATOS CONTABLES (POC DE LA CENTRAL DE BALANCES DEL BDE)

Seminario sobre Aplicaciones y Desarrollo de Big Data y Data Science en la Banca Central_CEMLA_Junio 2021

DEPARTAMENTOS DE ESTADÍSTICA Y SISTEMAS DE INFORMACIÓN

ÍNDICE

Cuestionarios con información contable de las empresas no financieras españolas:

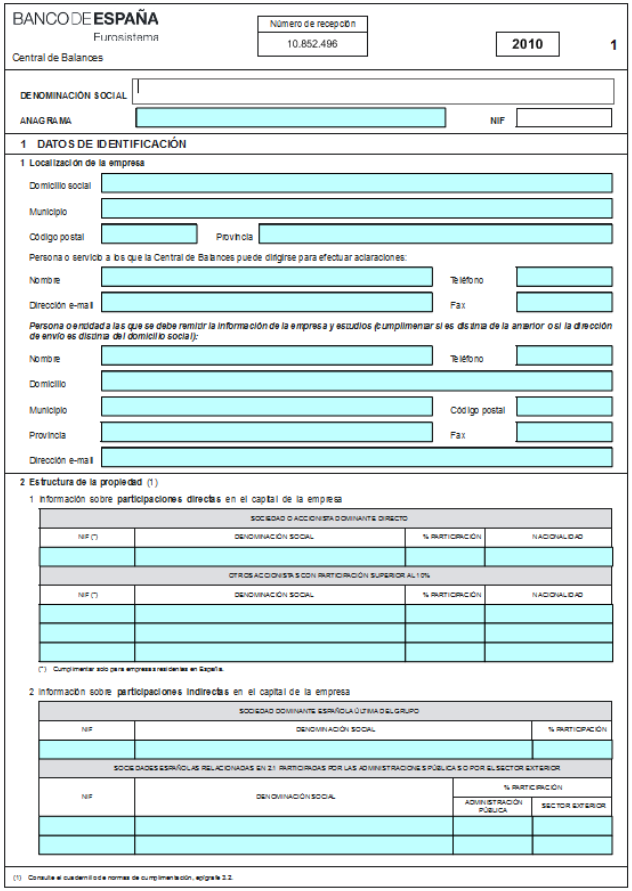
10 ejercicios x 900.000 empresas x 3.000 datos.

Depurados y clasificados mediante procesos automáticos.

Un 20% se clasifican como no aptos para estudio.

¿Puede la IA ayudarnos a mejorar estos procesos?

- Encontrar patrones alternativos para clasificar los cuestionarios: *Caso I. Detección de anomalías.*
- Completar la información omitida: *Caso II. Imputación de valores.*



BANCO DE ESPAÑA
Eurosistema
Central de Balances

Número de recepción
10.852.496

2010 1

DENOMINACIÓN SOCIAL
ANAGRAMA NIF

1 DATOS DE IDENTIFICACIÓN

1 Localización de la empresa

Domicilio social
Municipio
Código postal Provincia

Persona o servicio a los que la Central de Balances puede dirigirse para efectuar aclaraciones:

Nombre Teléfono
Dirección e-mail Fax

Persona o entidad a las que se debe remitir la información de la empresa y estudios (cumplimentar si es distinta de la anterior o si la dirección de envío es distinta del domicilio social):

Nombre Teléfono
Domicilio
Municipio Código postal
Provincia Fax
Dirección e-mail

2 Estructura de la propiedad (1)

1 Información sobre **participaciones directas** en el capital de la empresa

SOCIEDAD O ACCIONISTA DOMINANTE DIRECTO			
NIF (*)	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN	NACIONALIDAD
OTROS ACCIONISTAS CON PARTICIPACIÓN SUPERIOR AL 10%			
NIF (*)	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN	NACIONALIDAD

(*) Cumplimentar sólo para empresas residentes en España.

2 Información sobre **participaciones indirectas** en el capital de la empresa

SOCIEDAD DOMINANTE ESPAÑOLA ÚLTIMA DEL GRUPO		
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN
SOCIEDADES ESPAÑOLAS RELACIONADAS EN 2.1 PARTICIPADAS POR LAS ADMINISTRACIONES PÚBLICAS O POR EL SECTOR EXTERIOR		
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN: ADMINISTRACIÓN PÚBLICA / SECTOR EXTERIOR

(1) Consulte el cuadernillo de normas de cumplimentación, epígrafe 2.2.

1. INTRODUCCIÓN

Metodología en la POC de 2019

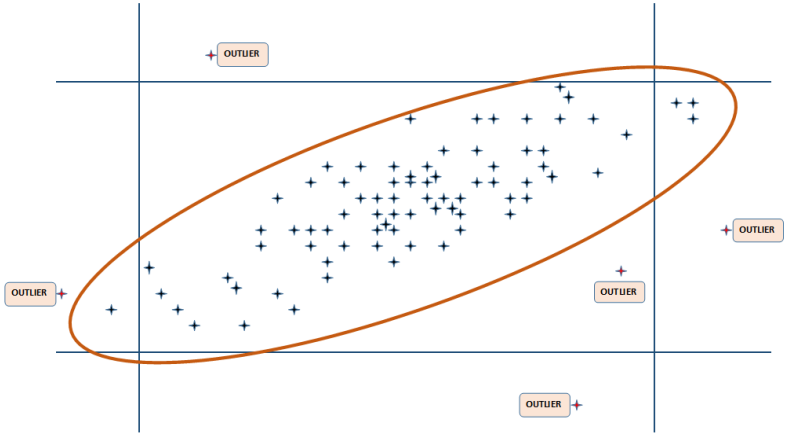
BANCO DE ESPAÑA
Eurosistema

iic
instituto de ingeniería del conocimiento

RECUPERAR CUESTIONARIOS PARA ESTUDIO

SCORE DE ANOMALÍAS

Índice de anomalía valorando n dimensiones



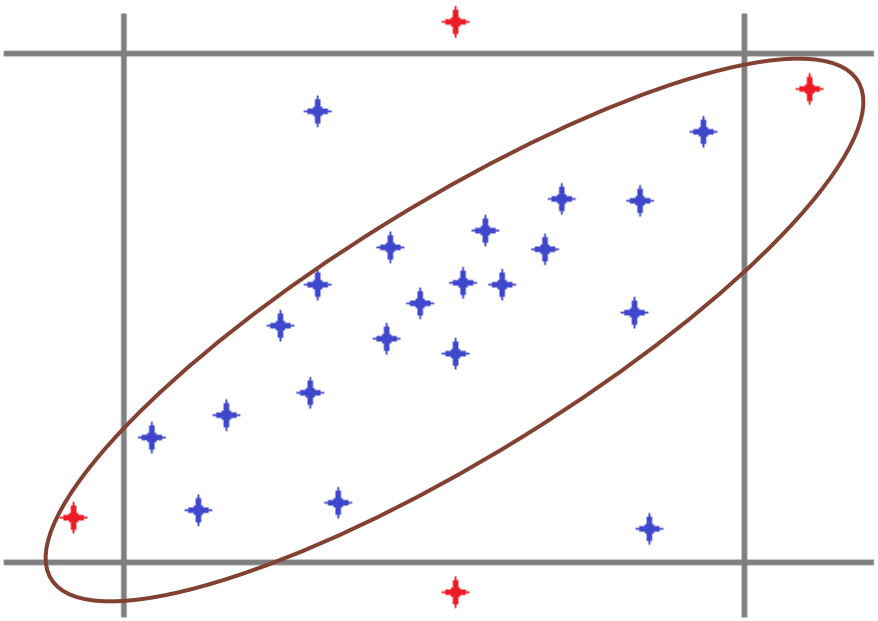
IMPUTACIÓN DE VALORES en:

(i) Descuadres más comunes y (ii) empleo

PATRIMONIO NETO Y PASIVO			NOTAS DE LA MEMORIA	EJERCICIO 2017 (1)
C)	PASIVO CORRIENTE	32000		95.200,00
I.	Pasivos vinculados con activos no corrientes mantenidos para la venta	32100		4.000,00
II.	Provisiones a corto plazo	32200		1.200,00
III.	Deudas a corto plazo	32300		5.000,00
1.	Deudas con entidades de crédito	32320		
2.	Acreedores por arrendamiento financiero	32330		
3.	Otras deudas a corto plazo	32390		
IV.	Deudas con empresas del grupo y asociadas a corto plazo	32400		2.000,00
V.	Acreedores comerciales y otras cuentas a pagar	32500		
1.	Proveedores	32580		
a)	*Proveedores a largo plazo*	32581		
b)	*Proveedores a corto plazo*	32582		

- **Selección de variables: 94 claves contables + clave empleo + 2 campos de sector de actividad (Sector y Gran Sector).**
- **Normalización contable: Dividir los campos de *Perdidas y Ganancias* entre el *Importe neto de la cifra de negocios*. El *Balance* entre el *Total Activo*.**
- **Filtro de cuestionarios: Subtipo reducido de 2008 a 2017, descartando los *No normalizables* (*Importe neto de la cifra de negocios = 0; ~2 millones instancias).***
- **Generar nuevas variables: Medias de cada valor en los últimos 2-5 años, número de sectores declarados, edad de la empresa…**
- **Separar cuestionarios según su calidad:**
  - Perfectos (5,323,000)
  - Baja calidad (476,000)
  - *Missing* (469,000)

SCORE DE ANOMALÍA

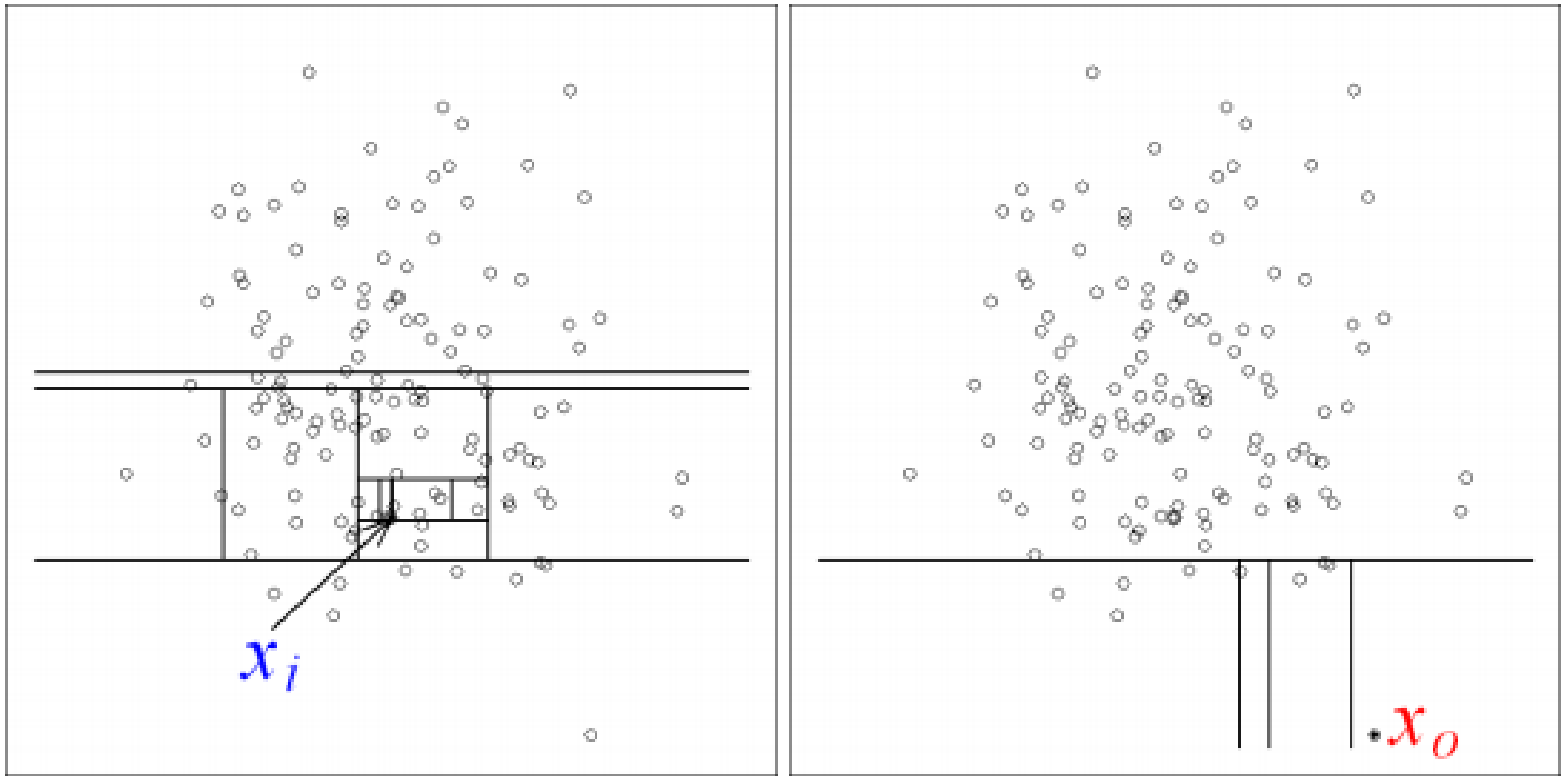


Cálculo de *score* de anomalía *[0,1]* vs. detección de *outliers (Sí/No).*

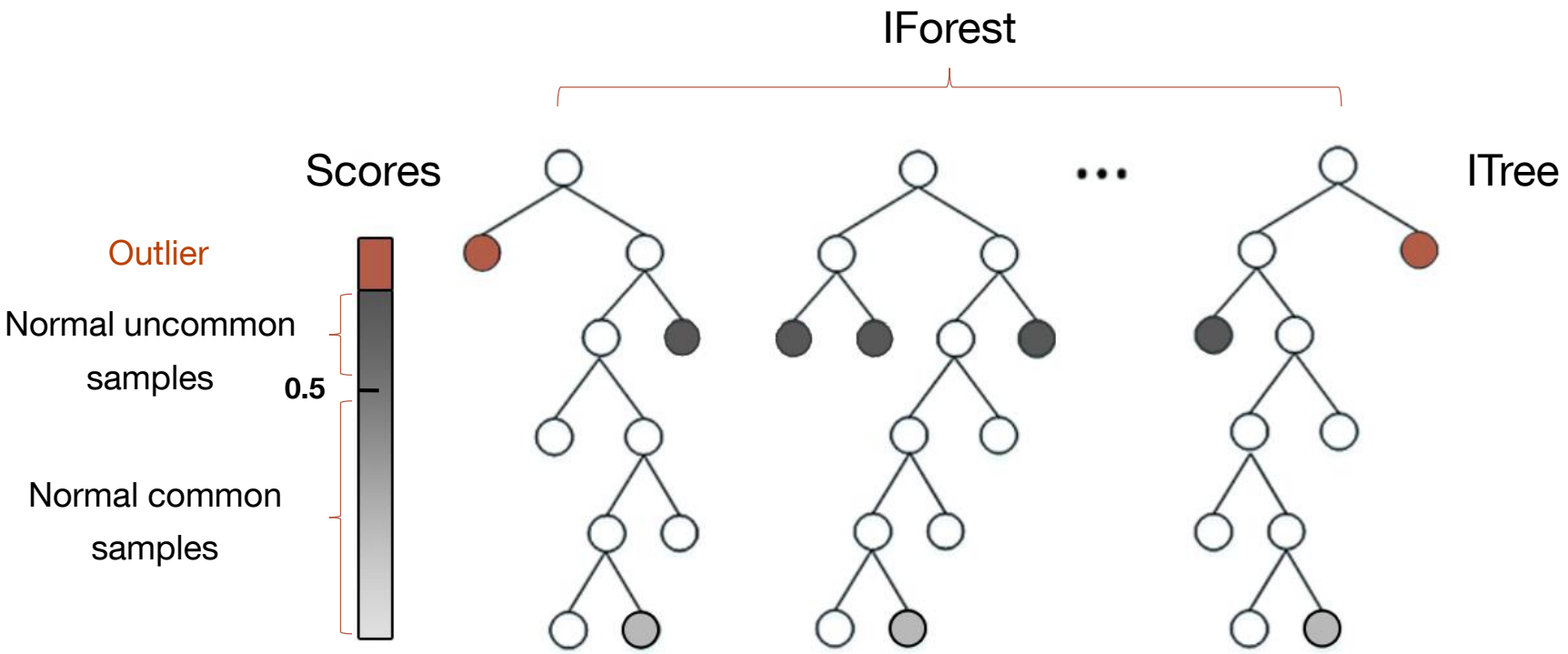
Algoritmo empleado: **Isolation Forest**.

ISOLATION FOREST

Las instancias anómalas se aíslan fácilmente mediante divisiones aleatorias del espacio.



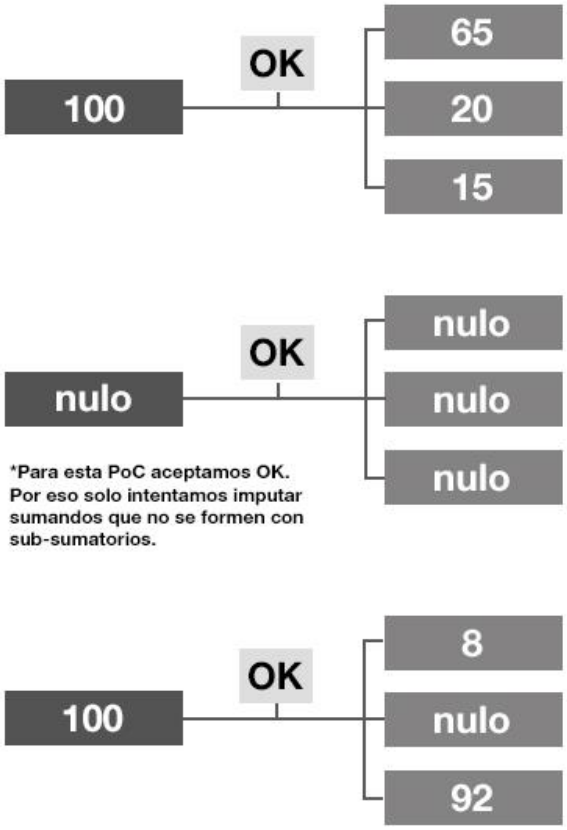
(a) Isolating x_i (b) Isolating x_o



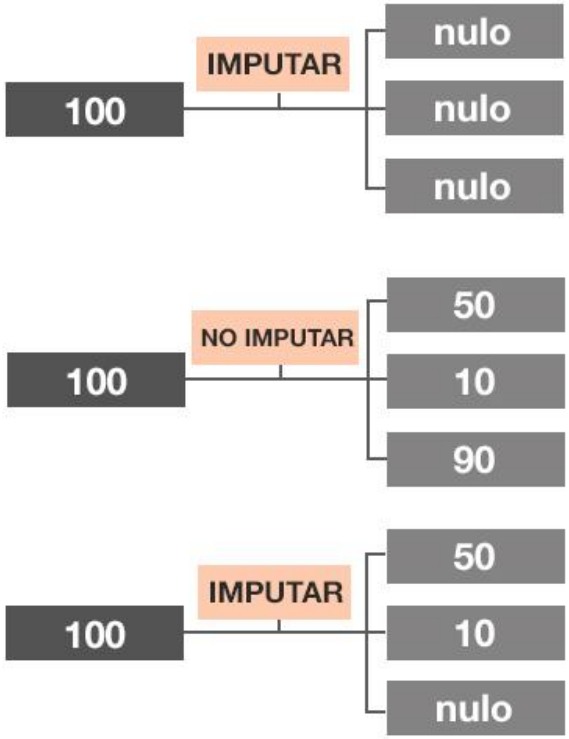
Entrenamiento con 5.000 árboles y todos los cuestionarios perfectos: 5.300.000

Se evalúa sobre los cuestionarios perfectos, de baja calidad y missing de 2017.

Imputación de valores



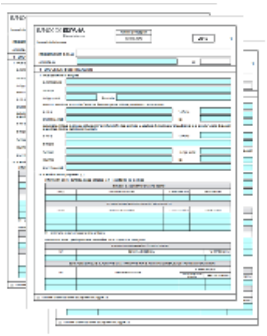
*Para esta PoC aceptamos OK. Por eso solo intentamos imputar sumandos que no se formen con sub-sumatorios.



* Si el sumatorio no es igual a la suma de sumandos: IMPUTAR sumandos.
Restricción: Tras la imputación deben cuadrar.
Holgura del cuadre: +/- 2%

Selección de datos:

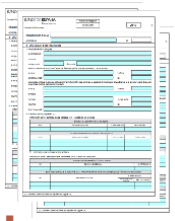
- Descuadres más comunes (4 ejercicios)
- Empleo (1 ejercicio)

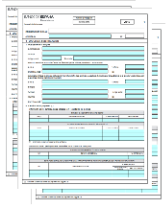


PERFECTOS

80%
TRAIN

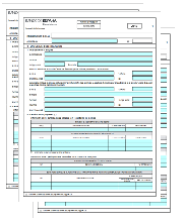
20%
TEST

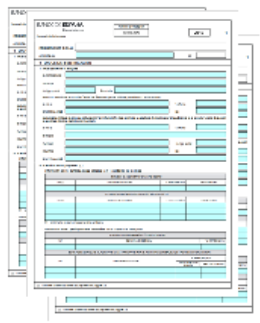




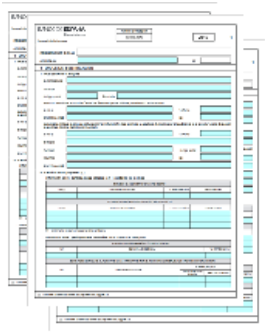
PERFORADO

MODELO
ENTRENADO





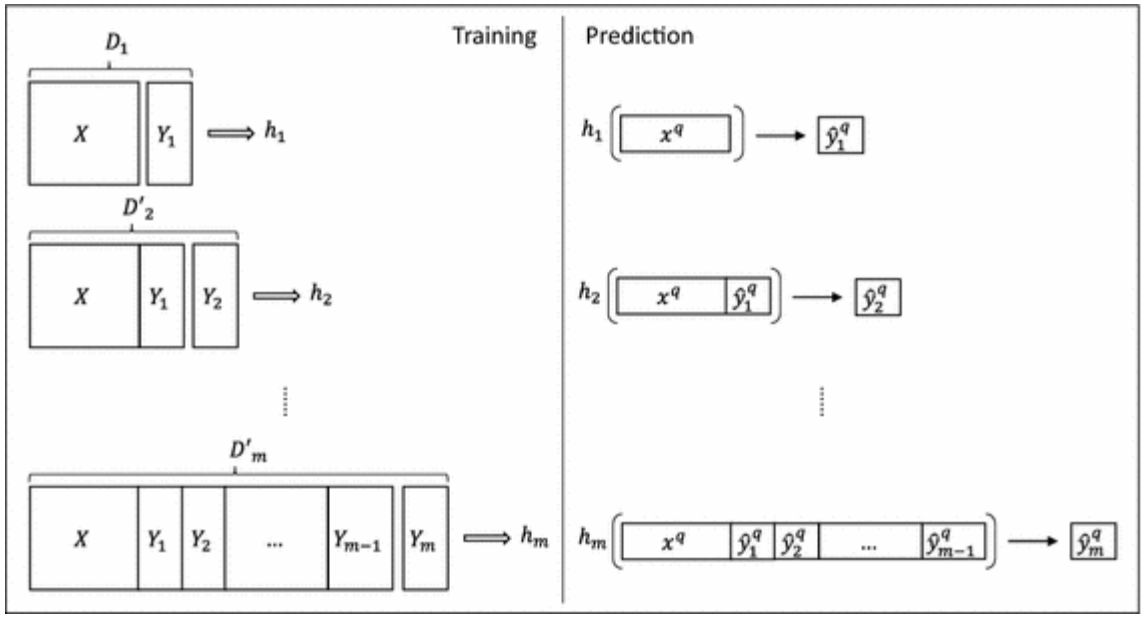
PREDICCIÓN



MISSING

ID_INST_CUEST	12300	12380	12370	12390	31200	31220	31230	31290	32300
13623376763	NaN	NaN	NaN	NaN	76.00000	NaN	NaN	NaN	-0.04868
13623376769	NaN	NaN	NaN	NaN	23.47207	NaN	NaN	9.63073	0.57878
13623376773	2.67538	2.02641	NaN	0.64897	3.11943	NaN	NaN	NaN	-6.45118
13623376777	91.91667	91.91667	NaN	NaN	9.31272	NaN	NaN	NaN	16.13428

Ensamble of Regressor Chains (ERC): Construir varios modelos de regresión de forma incremental. Cada modelo predice una variable que luego es empleada para entrenar el siguiente modelo.



Train: 240.000

Test: 60.000

Modelo de regresión:

Random forests 1.000 árboles

Computacionalmente muy costoso.

El orden de predicción de las variables (cadena) teóricamente afecta al resultado, dando mayor peso a las primeras variables elegidas. Se prueban 5 cadenas aleatorias.

¿Falsos positivos? Analizar para detectar posibles mejoras en nuestros sistemas de filtrado.

¿Falsos negativos? Analizar para en su caso 'relajar' nuestros sistemas de filtrado

¿El 94% de los cuestionarios se concentran en un rango de anomalía entre 0 y 0,2

CALIDAD CUESTIONARIOS CBB 2017

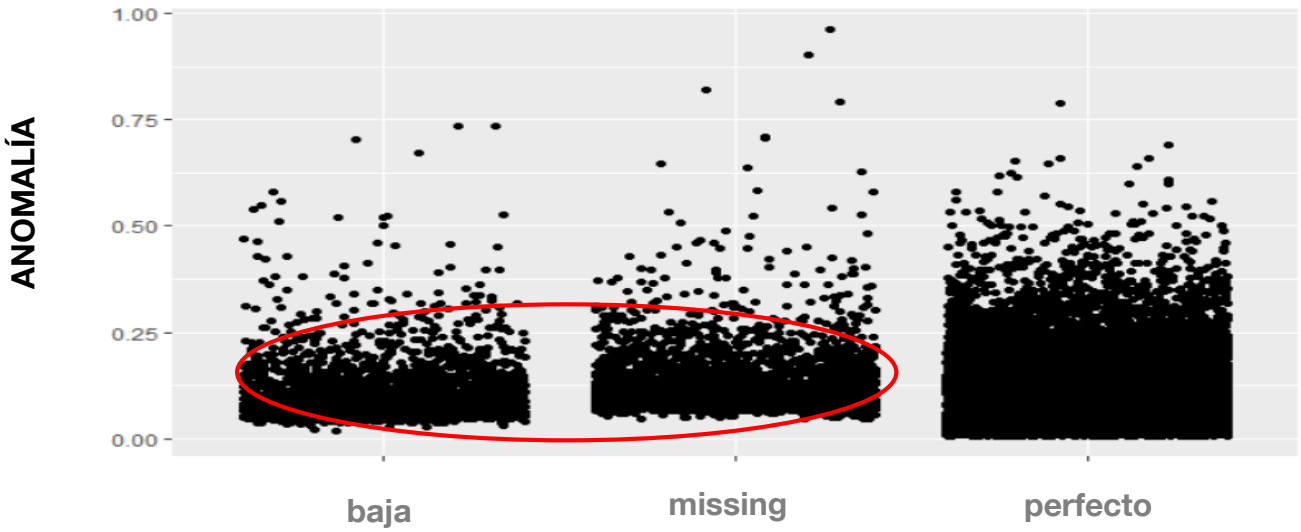
Scoring IIC (0=Bueno; 1=Malo)	PERFECTO	NO PERFECTO	TOTAL	% Total acumulado
0-0,1	411.973	41.626	453.599	71,3%
0,1-0,2	118.439	28.942	147.381	94,4%
0,2-0,3	20.380	5.404	25.784	98,5%
0,3-0,4	5.154	1.377	6.531	99,5%
>0,4	2.299	853	3.152	100,0%
TOTAL	**558.245**	**78.202**	**636.447**	

En general existe sintonía entre los resultados de los modelos de ML y los obtenidos con criterios deterministas de CB

En resumen:

Aceptando este score…	…en neto ganamos o perdemos sociedades	…renunciando a estas…	…e incorporando estas …
0,1	-104,646	-104,272	41,626
0,2	42.735	-27.833	70.568
0,3	68.519	-7.453	75.972
0,4	75.050	-2.299	77.349

SCORING DE ANOMALÍA SEGÚN IIC VS NIVEL DE CALIDAD SEGÚN CENTRAL DE BALANCES

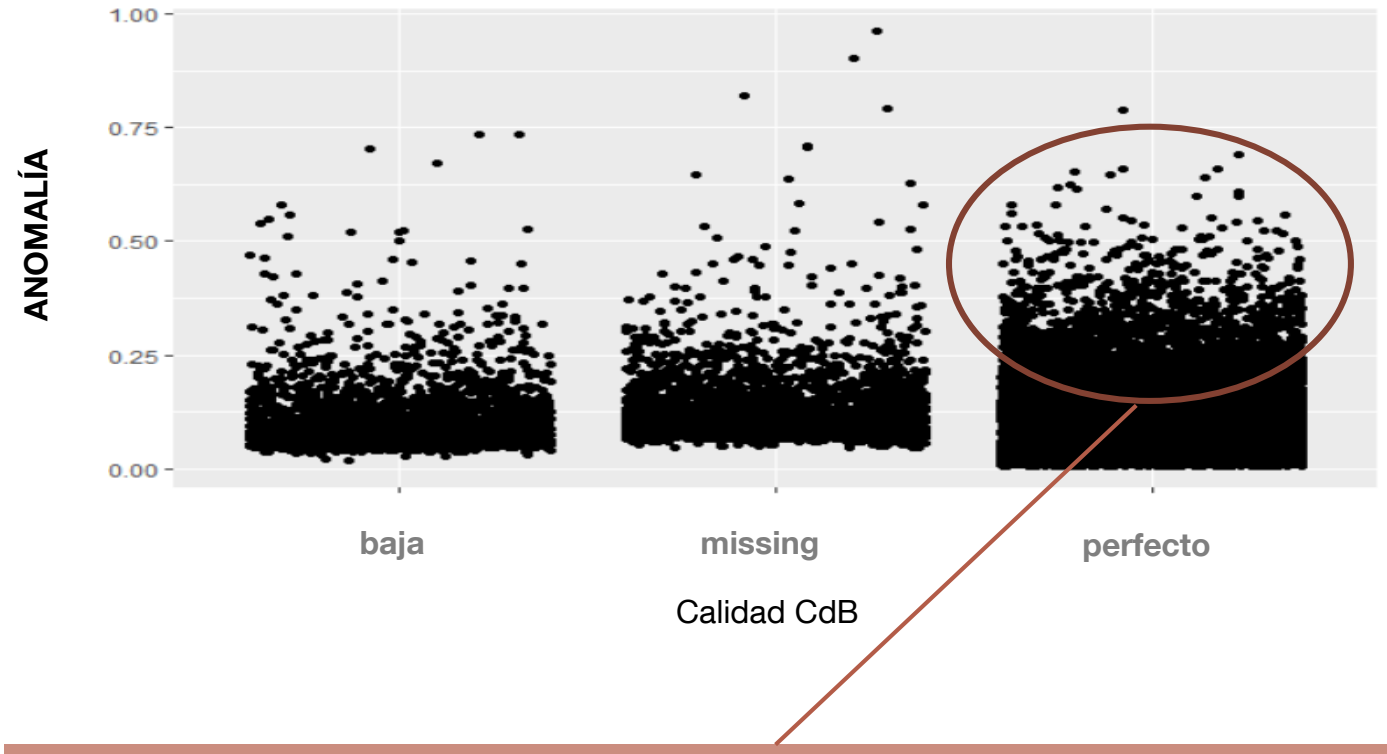


¿Complementar con el score de anomalías para recuperar empresas?

¿Recuperables mediante imputaciones?

SCORE	TOTAL CUESTIONARIOS	NO CUADRADA		UNIDADES NO FIABLES		INCOMPLETA		NO CUMPLE CONDICION VAREXCESIVA		PERSONAL NO COHERENTE	
hasta 0,05	2.242	438	20%	96	4%	193	9%	1.462	65%	67	3%
hasta 0,1	41.626	6.092	15%	2.733	7%	2.503	6%	13.268	32%	20.833	50%
hasta 0,15	62.482	10.426	17%	4.128	7%	3.420	5%	18.506	30%	32.915	53%
hasta 0,2	70.568	12.821	18%	4.620	7%	3.724	5%	20.720	29%	37.077	53%

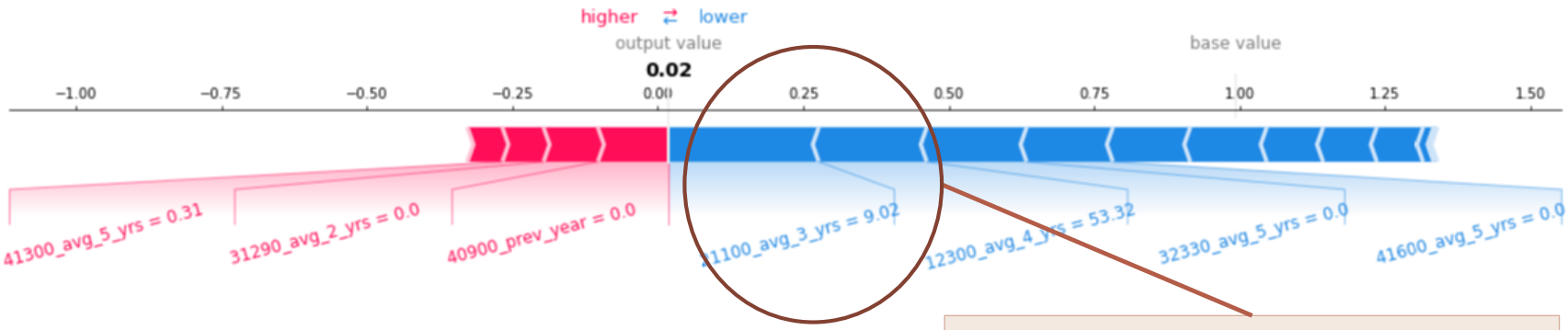
SCORING DE ANOMALÍA SEGÚN IIC VS NIVEL DE CALIDAD SEGÚN CENTRAL DE BALANCES



Intentando entender el algoritmo… lo primero, bajar al microdato

CONTRIBUCIONES AL ÍNDICE DE ANOMALÍA: RATIOS DE SHAPLEY

Indican qué claves han contribuido más al índice de anomalía

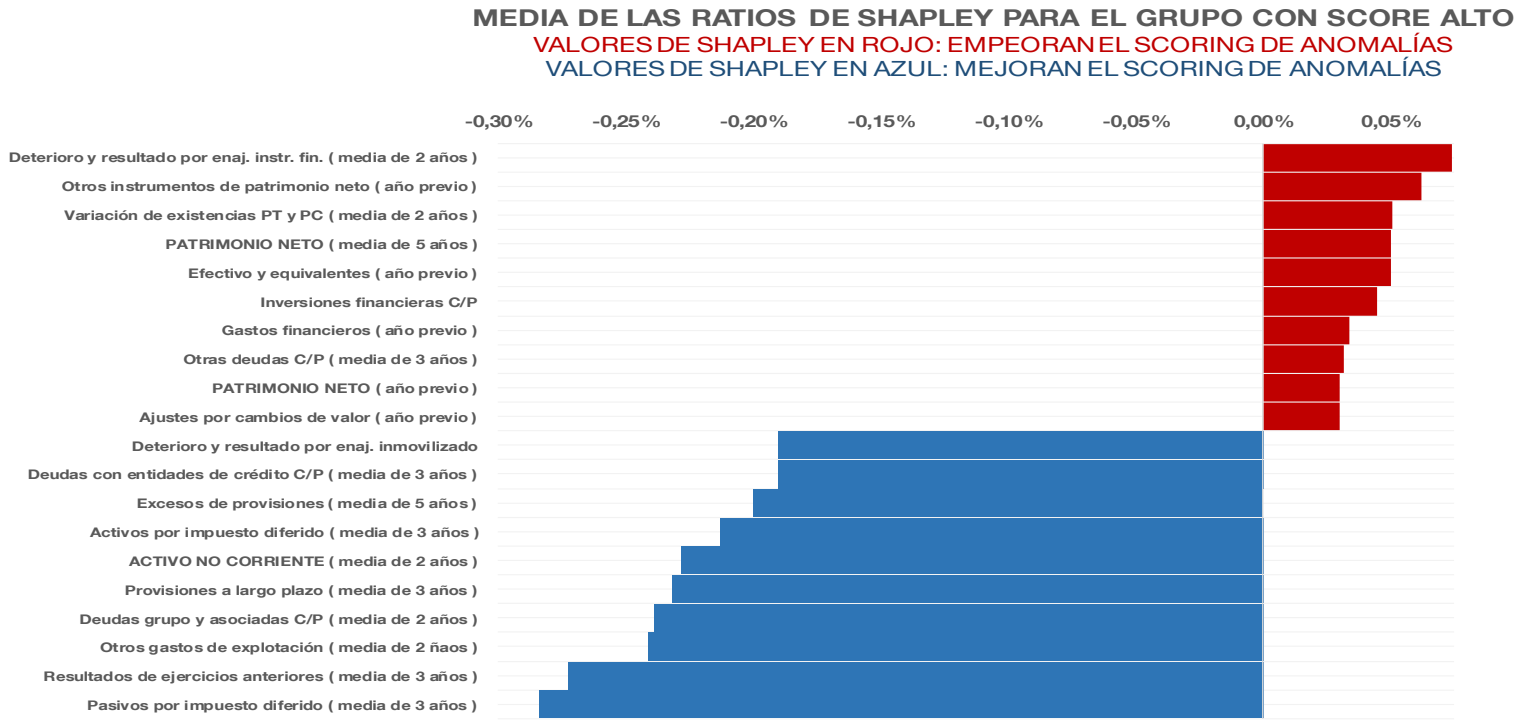


Significa que la variable *"media de los 3 últimos años de la clave 21100"*, al tener un valor de 9,2, reduce el escore 0,26 puntos aprox.

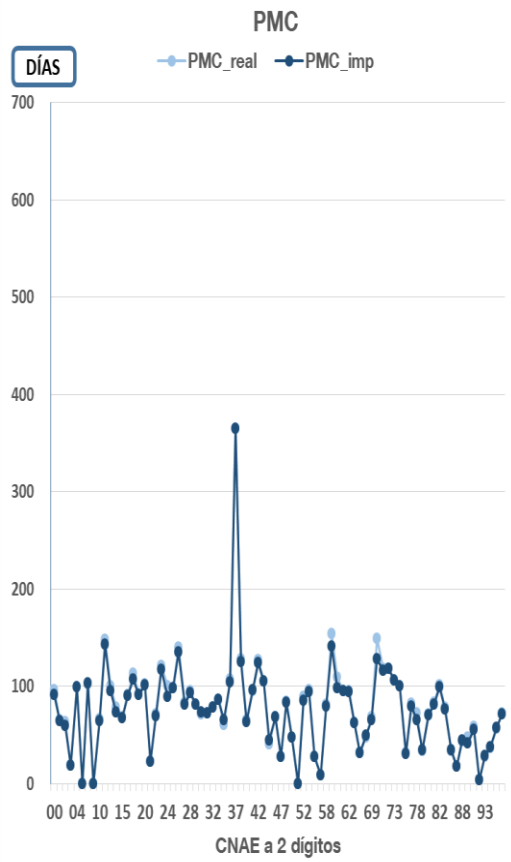
Interpretación individual: poco útil en los casos revisados

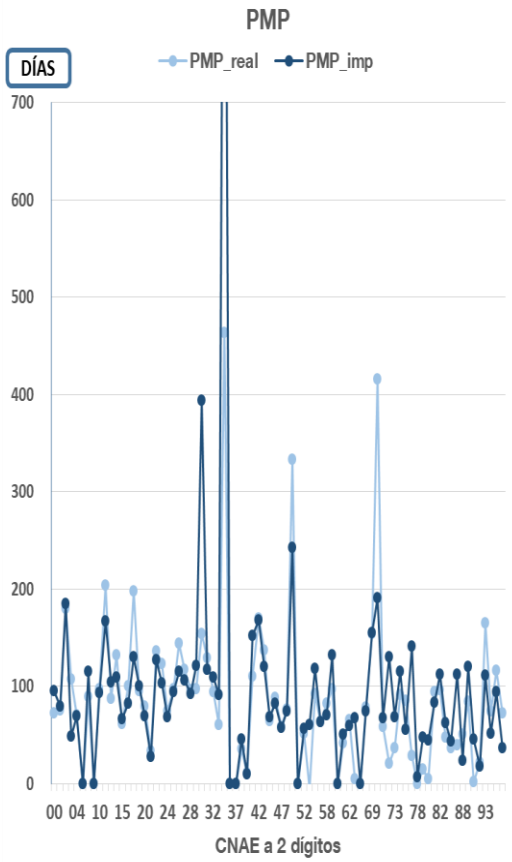
Propiedad aditiva para analizar un nodo concreto

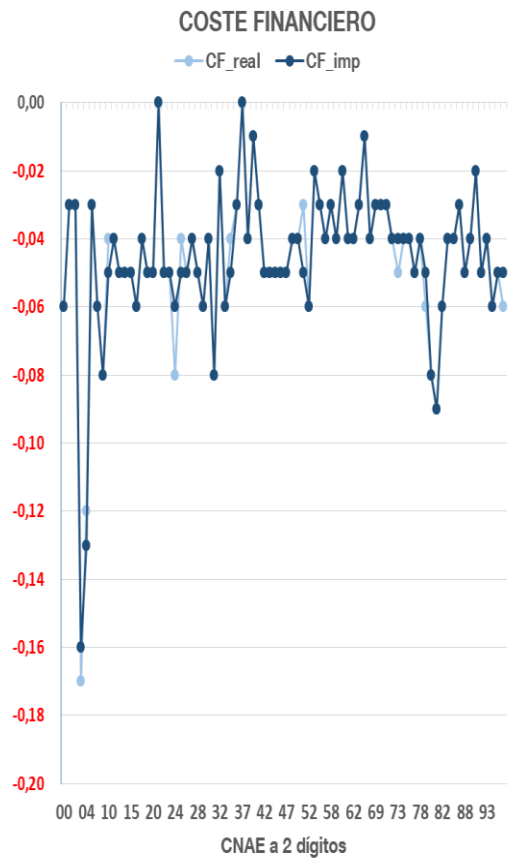
Las ratios de Shapley ***AGREGADAS*** permiten interpretar cuál es el efecto de cada variable en el **conjunto de sociedades** que se elija (en este caso son aquellas con scoring alto y provenientes de empresas perfectas; pero a efectos de la CdB se podría elegir un nodo, un tamaño de empresa concreto, una comunidad autónoma, o cualquier otro conjunto).



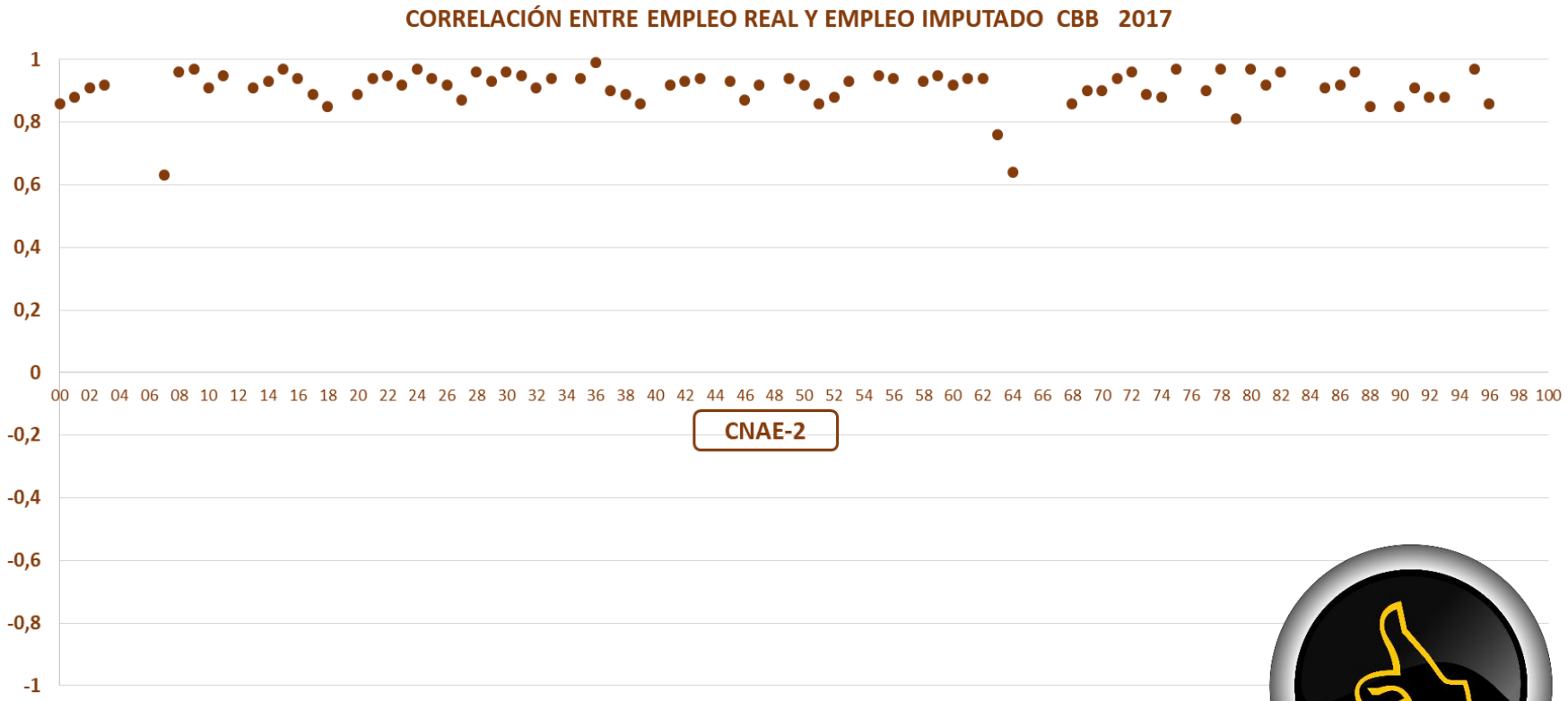
Herramienta potente para investigar pero los datos disponibles no lo permiten actualmente







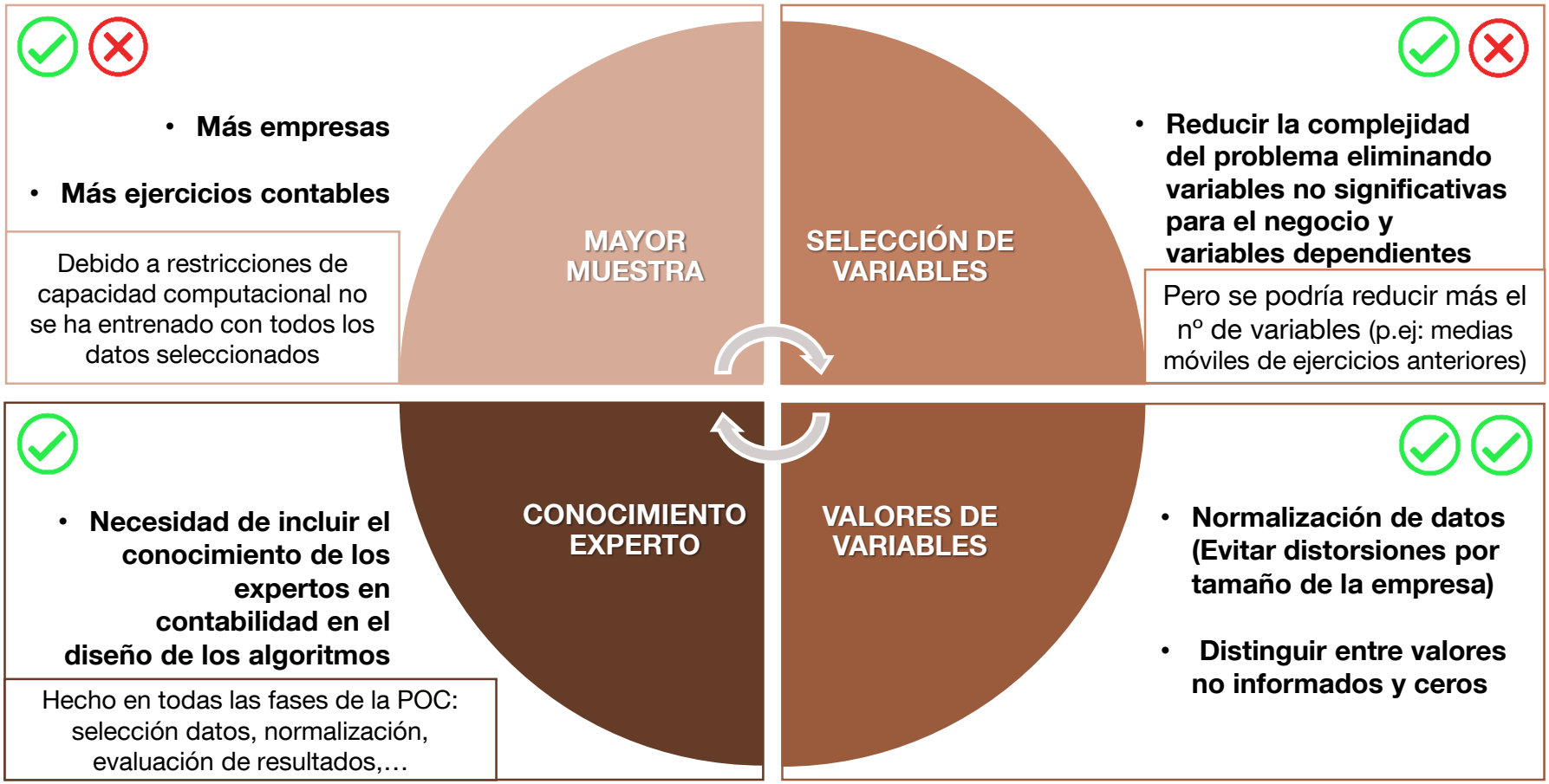
Las correlaciones son aceptables para PMC y Coste financiero, pero bajan para PMP, quizá porque se han realizado menos imputaciones en la clave de proveedores.


CORRELACIÓN ENTRE EMPLEO REAL Y EMPLEO IMPUTADO CBB 2017
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
00 02 04 06 08 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100
CNAE-2

4. LECCIONES APRENDIDAS Y SIGUIENTES PASOS

Qué mejora la POC de 2019 respecto del Piloto de 2018

BANCO DE ESPAÑA
Eurosistema



- **Más empresas**
- **Más ejercicios contables**

Debido a restricciones de capacidad computacional no se ha entrenado con todos los datos seleccionados

- **Reducir la complejidad del problema eliminando variables no significativas para el negocio y variables dependientes**

Pero se podría reducir más el nº de variables (p.ej: medias móviles de ejercicios anteriores)

- **Necesidad de incluir el conocimiento de los expertos en contabilidad en el diseño de los algoritmos**

Hecho en todas las fases de la POC: selección datos, normalización, evaluación de resultados,…

- **Normalización de datos (Evitar distorsiones por tamaño de la empresa)**
- **Distinguir entre valores no informados y ceros**

Para dar validez al score de anomalías es necesario:

- Disponer de ratios de **Shapley agregadas *customizadas*** a las necesidades de negocio (determinado sector, tamaño,…)

Para dar validez a las imputaciones es necesario:

- Disponer de ratios de **Shapley para imputaciones** y no sólo para anomalías
- Revisar el **patrón de perforado** en el conjunto de test (p.ej: clave de proveedores con pocos datos imputados)
- Probar a repetir las **imputaciones después de la eliminación de cuestionarios anómalos**

En definitiva, más análisis…

BANCO DE ESPAÑA
Eurosistema

GRACIAS POR SU ATENCIÓN

ESTADÍSTICA Y SISTEMAS DE INFORMACIÓN

Análisis de las 5 sociedades del CNAE-2 como ejemplo de caso que presenta una diferencia elevada entre ratio PMP con valores reales y PMP con valores imputados.

EMPRESA	EJERCICIO	PMP_real	PMP_imp	Aprovisionamientos	Proveedores real	Proveedores imputado
1558271	2012	239,6	198,5	-15,5	10,1	8,4
1964308	2013	86,6	345,4	-83,3	19,8	78,8
2007712	2014	126,9	1378,3	-95,2	33,1	359,4
2809470	2015	54,9	93,9	-292,2	44	75,2
995966	2017	414,6	398,6	-136,9	155,5	149,5
TOTAL CNAE_2 = 30		153,8	491,5	-623,1	262,5	671,3