IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

An artificial intelligence application for accounting data cleansing[1]

Pablo Jiménez and Tello Serrano,
Bank of Spain

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Banco de España

AN ARTIFICIAL INTELLIGENCE APPLICATION FOR ACCOUNTING DATA CLEANSING

Central Balance Sheet Data Office (CBSO) Division. SMEs Unit.
Management Systems Division II. Economics and Statistics Systems Unit.
Collaboration: Instituto de Ingeniería del Conocimiento (IIC).

Abstract

The CBSO maintains, among others, a large annual accounting database of the Spanish non-financial corporations sector. By application of automatic rules, the CBSO considers the data of approximately 20% of companies as invalid for studies. Faced with this situation, the dual objective of this proof of concept (PoC) was to explore the application of machine learning techniques, to complement the detection of anomalous cases (outliers) and allow imputation in the absence of data (missing data). In this way, with the help of these new techniques, the CBSO seeks to maximise the number of companies with coherent information and minimise the risk of introducing anomalous questionnaires in the sample. This seems possible in view of the encouraging results of this study.

Keywords: isolation forest, outliers, anomaly detection, missing data, ERC, regression chains, accounting data, imputations, Shapley values, artificial intelligence (AI), machine learning.

JEL classification: C61 y C81

1. Introduction

The CBSO collects annual accounting data from non-financial corporations (NFCs) through two different sources that determine the creation of two very different databases: the Central Balance Sheet Data Office Annual Survey (CBA), which relies on voluntary contributions by reporting firms, where large corporations account for a larger share and where the nearly 11,000 questionnaires that make up the survey are manually refined (standardised questionnaire adapted to the Spanish General Chart of Accounts); and the CBB database which, based on a collaboration agreement between the Banco de España, the Ministry of Justice and the Spanish Association of Property and Mercantile Registrars (CORPME), receives the annual accounts

deposited in the Mercantile Registers each year by Spanish companies in standardised models and where small and medium-sized enterprises (SMEs) are widely represented.

The data obtained from the Mercantile Registers are used to check the information held and provide information on a large sample of NFCs. They enable population totals to be inferred and make it possible to monitor NFCs which are underrepresented in the database built on the voluntary contributions of the CBSO reporting firms. The CBSO thus holds data on approximately one million firms for each financial year, of which more than 800,000 may ultimately be used for the preparation of studies, once the various automated data quality and consistency processes have been carried out.

There are two main reasons for the low quality of around 20% of the questionnaires determined by the automated validation system applied in the CBB sample:

- Data mismatches, due either to errors in the recording of values or to missing data.

- Data inconsistencies from a logical-accounting standpoint, such as, for example, high data variance between two consecutive years that cast doubt on their comparability.

Given that the data of approximately 20% of companies are being rejected as low quality, and to gain a more accurate picture of the population of NFCs, the PoC aims to meet two different objectives through the use of machine learning techniques applied to the CBB sample: (1) imputation of missing data; and (2) detection of anomalous questionnaires (outliers).

2. PoC objectives

2.1. Imputation of missing values

In each questionnaire there are certain information headings that are broken down in turn into detailed information; if these sections are not completed or are completed incorrectly the questionnaires are considered invalid. Appropriate imputation of the missing information would enhance the final quality.

For this test, four accounting items were selected whose amount is generally reported but whose breakdown is often incomplete. One of these items is "short-term debts", which consists of three **addends**: debts with credit institutions, finance lease creditors and other debts. The algorithm must fill in the missing addends, **subject to the restriction that their sum matches the total**.

There are numerous ways to impute values: imputing the mean, the median, regressions, moving averages, etc. The use of machine learning techniques seeks to ensure that **the imputation is neither linear nor pre-defined** or, in other words, that the imputed data are not biased by the aggregate chosen to obtain the supposedly analogous values, but that Artificial Intelligence evaluates the complete set of companies and learns from their characteristics to determine which imputation is correct. In this way, **no human decisions affect the data,**

introducing biases, pre-defined views of how each company should look or pre-defined functional forms to which the data must adapt. The main idea is not to make a priori assumptions.

The **ensemble of regressor chains** method (explained below) was chosen for the imputation of missing values.

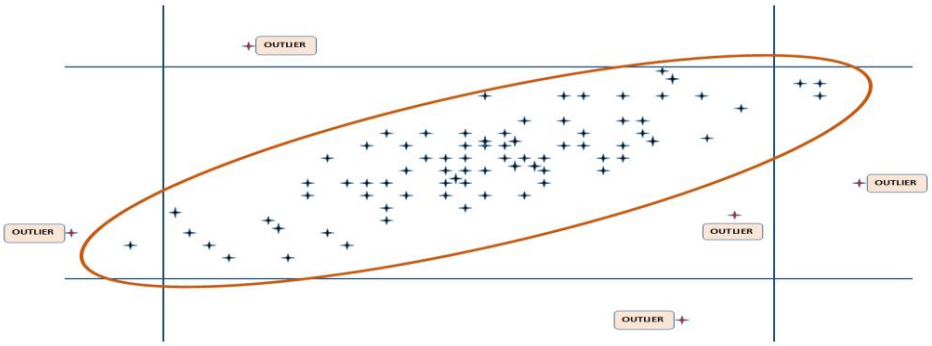
2.2. Detection of outliers

Value anomalies are commonly analysed independently of other variables, that is, from a one-dimensional standpoint. However, in this study an algorithm was used that captures the structure of the data. A value might be anomalous if considered individually, but it might not be when the **specific structure** of the data set is considered.

In Figure 1, the red points correspond to corporations that appear anomalous from a one-dimensional perspective because they lie outside the rectangle that marks the limits of that perspective. But they would not be outliers if observed **multidimensionally** because they remain within the normal structure of the point cloud. Similarly, there could be instances where, despite being within the limits set for the detection of outliers, the points are distant from the densely populated areas of the problem space and are, therefore, relatively anomalous.

The points that are outside the ellipse would be considered anomalous.

Figure 1: Multidimensional outliers



For the sake of clarity, the figure shows two dimensions, although the real problem addressed in the PoC has 97 dimensions (corresponding to the 97 variables used).

The method chosen for detection of outliers was an **Isolation Forest** (explained below).

As a by-product of the study, hidden patterns were discovered that are not visible using other more common statistical techniques. **Shapley values** or ratios**, the usual basis of the XAI (eXplainable AI)**, stand out for these purposes.

3. Characteristics of the microdata, selection and preparation of the sample

3.1. Data pre-processing

To enable the algorithms used to reach a solution, it is essential to work with as **many instances as possible**, use **comparable formats** and reduce the number of variables or **dimensional space** of the problem, without losing crucial information. This pre-processing of the data was carried out taking advantage of the CBSO's **expertise** in this field.

The CBB questionnaires from the Spanish Mercantile Registers may be one of two types, depending on the level of information they contain: normal or abbreviated. For this study only the abbreviated format was considered, since this is the one generally used by SMEs when they file their accounts. Accordingly, large corporations were excluded from the scope of the PoC. The abbreviated questionnaire format consists of four parts: some identification data, balance sheet data, profit and loss account data and model financial statements.

In order for the data to be comparable, the sample had to be **standardised**. This was done by dividing the Balance sheet variables by Total assets and the Profit and loss account variables by Net turnover. Likewise, the Employment variable was divided by Net turnover, in order to normalise it, although this makes no sense from an accounting standpoint.

The **number of dimensions** also had to be reduced: starting from more than 3,000 variables, linear relationships were eliminated (some being the result of the sum of others). Finally, the 97 most significant variables for studying a company, according to the accounting experts' criteria, were selected.

In addition, various **auxiliary variables** were generated to represent certain information that is useful when studying a company's accounts:

a. New variables containing the average values of each field in the last one, two, three, four and five years. These variables will provide the fundamental historical information when predicting a non-imputed value.

b. Age of the company, calculated from the date of the first questionnaire completed.

c. Number of different sectors reported by the company in its history.

d. Number of different large sectors reported.

Finally, in order to **classify the instances** according to their quality, null values owing to lack of information had to be distinguished from those that denote zero.

This process consisted of checking the values of the variables with the sum of their breakdowns: null values that participate in a correct sum were considered zero values; all others were considered missing. Instances were thus divided into three groups:

1. **Perfect** questionnaires: instances with no missing values, considered suitable for study according to the CBSO's automated debugging rules.
2. **Low quality** questionnaires: instances with no missing values, considered unsuitable for study.
3. **Missing data** questionnaires: instances with empty values.

3.2. Data selection criteria

In summary, the criteria adopted for selection of the questionnaires were:

a. Abbreviated subtype CBB questionnaires from 2008 to 2017 (the last complete year when the PoC started) from which 97 variables were selected:
   - 94 accounting items on the balance sheet and profit and loss account corresponding to the current year (the variables from the previous year were discarded).
   - Total average employment of each company.
   - Two sector variables: large sector of activity and 2-digit NACE code.

b. Both perfect and low quality questionnaires were included.

c. Non-standardisable instances were excluded, that is, instances where net turnover or total assets were equal to zero.

d. Other instances that the CBSO discards for different reasons (their main variables are blank, they have high negative values in positive variables or financial sector instances) were also eliminated.

Altogether, **more than 6.2 million questionnaires were included in the PoC**, of which, according to the groups explained above, 5.3 million were 'perfect', 0.5 million were 'low quality' and 0.5 million were 'missing data'.

4. Methods applied

4.1. Methods applied to detect anomalous observations

To begin with, different methods were proposed to undertake the project. Several common techniques start by considering that the data come from normal distributions, others do not address the problem of the joint structure of the data, others require some type of initial assumption, others overlook the

problem of data imbalance (by definition anomalies represent only a tiny percentage of the total data), etc.

Some of the techniques initially considered were: PCA (Principal Component Analysis), Mahalanobis distances, KNNs (K Nearest Neighbors), K-means, etc.

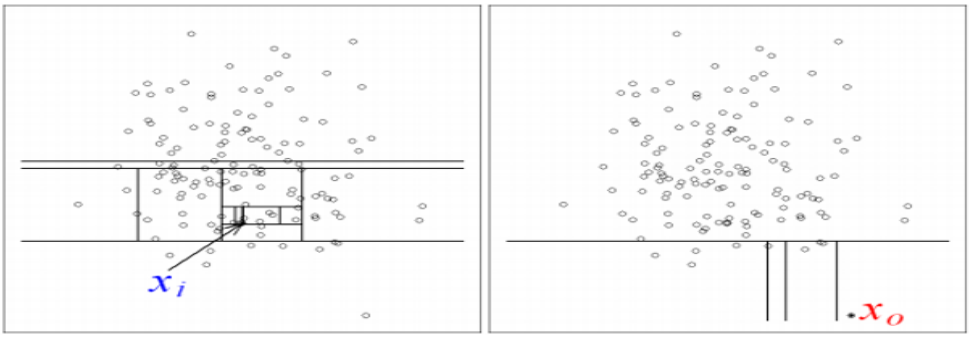
After applying some of these techniques to the data, analysing them and discussing the results, the **Isolation Forest** technique was finally chosen. This is an unsupervised algorithm that consists of "cutting" the space that houses the n-dimensional points by means of random secant planes of dimension n-1 (see Figure 2).

The main idea is that the more cuts it takes to isolate a point, the less anomalous that point is. From the opposite perspective, if a single cut is able to isolate a point, that point is far from the rest, therefore it is anomalous.

Figure 2 shows, in 2 dimensions, that isolating point x_i requires more cuts than isolating point x_0.

Image 2: Isolation Forest

Source: Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou (2008), "Isolation-based Anomaly Detection" (page 4).



Since the cuts are random, many are executed and the average number of cuts needed to isolate each point in successive attempts is calculated. This mean, normalised between 0 and 1, is the anomaly score, with 1 being the maximum anomaly indicator and 0 the minimum anomaly indicator.

The formula for each point x is: **score = 2 ^(** $-\boldsymbol{E}(\boldsymbol{cuts}(\boldsymbol{x})) * \boldsymbol{standarisation}$ **)**

The Isolation Forest technique has a problem when there are unreported variables. In those cases, the instances with unknown values in the variable by which the space is being subdivided cannot be classified in either of the two resulting subspaces. To deal with these cases, IIC created and implemented the **Missolation Forest** algorithm ("missing" + "isolation"). This algorithm takes into account the anomaly score that would be assigned to the instance if it were in each of the two subspaces, giving as a result the average of those values weighted by the probability of it being in one or the other subspace, according to the number of instances in each subspace.

4.2 Methods applied for imputation of values: ensemble of regressor chains (ERC)

After trying other methods such as self-similar neural networks (variational autoencoders) and MICE (multiple imputations) and obtaining a lower level of accuracy, an ERC was chosen.

Training this algorithm consists of randomly permuting the set of target variables (Y3, Y4, Y1, Y2) and constructing a regression for the first variable (Y3). In the next step, another regression is built for the next variable (Y4), but including as another regressor the result of the first regression (the estimate of Y3) as an added regressor for Y4. Thus, successively, **the regressions are "chained".**

In this particular case, the target variables are the fields to be completed: short-term debts with credit institutions, finance lease creditors and other debts.

Naturally, the results depend on the order obtained in the permutation. In our example, since Y3 is estimated before Y4, the effect that Y4 could have on Y3 does not appear. To avoid this problem, k groups of observations (bootstrap) and different permutations of the target variables are used, and the final prediction is computed as the **mean of the k individual predictions for each target variable**.

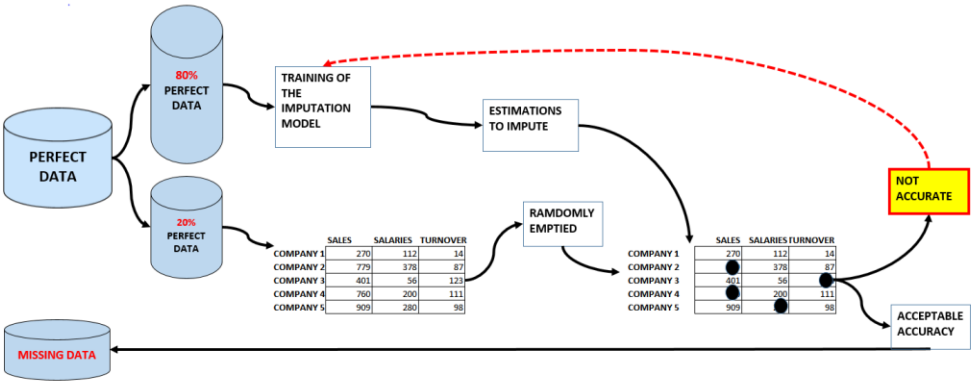
At the initiative of IIC, each regression was run through a **random forest** algorithm comprising a thousand trees and with randomly selected explanatory variables.

According to certain interpretations of this algorithm in the literature, it could be compared to a deep learning neural network in which the layers of the network are replaced by random regressions.

Since this algorithm is based on permutations, it is **computationally expensive**, especially in this case owing to the complexity added by the random forests used for each regression. Accordingly, it was considered appropriate to cut the number of observations to be used to 240,000 instances. Even so, it is a slow process to train.

To obtain an adequate prediction model, the algorithm was applied to 80% of the questionnaires which, according to the arithmetic logic procedures, were considered perfect ('training set'). The remaining 20% formed the test set, in which some of the data (perforation) were randomly emptied in order to evaluate the quality of the imputation carried out at a later stage. The trained prediction model was applied to the missing forms to estimate the relevant value to be imputed.

Figure 3: Training cycle and imputation
Source: Own elaboration.



5. Analysis of results

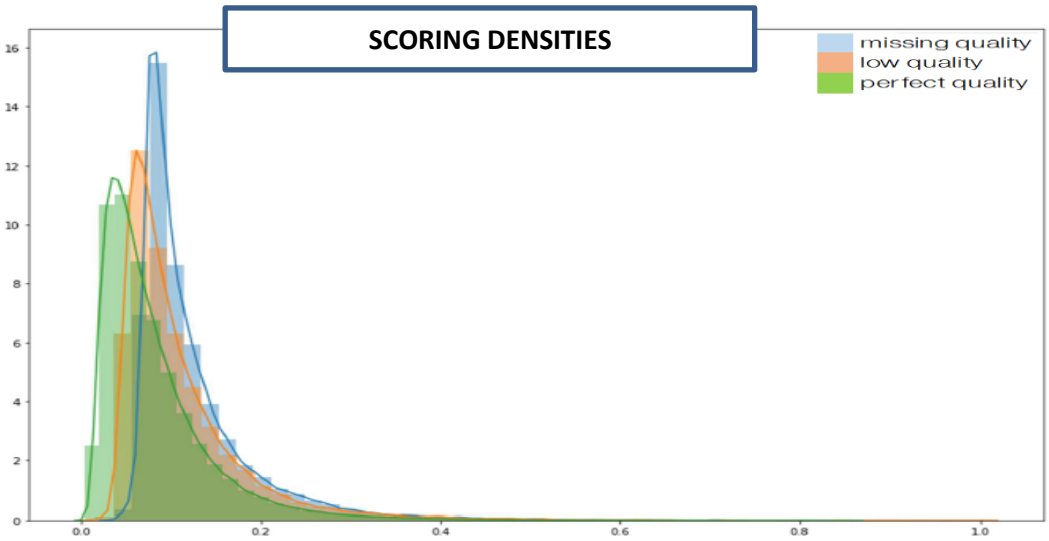
5.1. Results obtained in the detection of anomalous questionnaires

The **degree of anomaly** is indicated by a score between 0 (not anomalous) and 1 (highly anomalous). Since the concept of "anomaly" is not dichotomous, to put the algorithm into production a limit beyond which an observation is considered anomalous must be accepted. According to this distribution, a limit of 0.25 would render 5% of the observations anomalous.

The degree of anomaly estimated by the algorithm is similar, on average, to what the CBSO had been classing as "low quality". The companies which according to the CBSO had perfect quality show a low degree of anomaly (green histogram in Chart 1), those which according to the CBSO had low quality show slightly more anomaly (orange histogram) and those which lacked data show the highest degree of anomaly (blue histogram). This suggests that the algorithm captures the meaning of the task to be performed, at least on average, and that the path followed in the PoC may be the correct one.

Chart 1: Distribution of anomaly scores

Source: IIC.

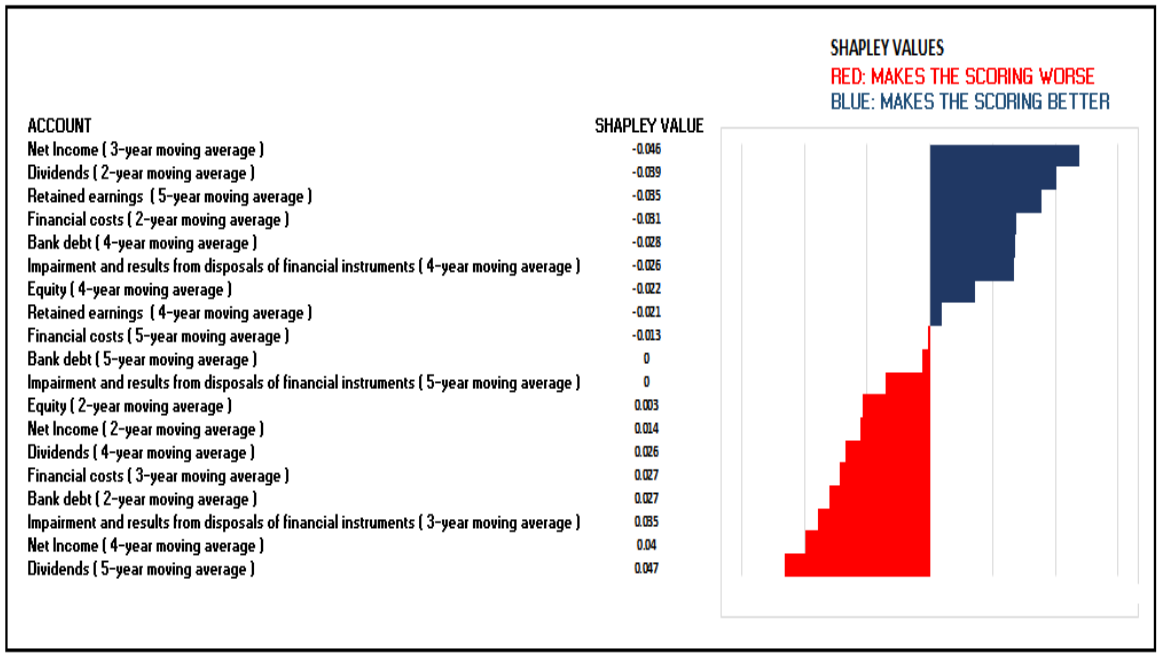


One way to dive into the origin of the anomalies is by using the **Shapley values**. These provide information on the extent to which each accounting item affects the anomaly score of each company. As Chart 2 shows, the "3-year moving average of dividens" is the item that most affects the degree of anomaly.

At the beginning of this PoC, the power of these values was not taken into consideration. Accordingly, from the start, the PoC was not designed to capture their explanatory power in an ideal way. For example, it was decided to use moving averages for each variable; this makes interpretability difficult because it disguises these effects within the moving averages themselves. To be more explicit, it could be the case that, for the same variable, the 3-year moving average was significantly positive but the 2-year moving average was significantly negative. They would thus be contradicting each other, and although they do not actually contradict each other, this complicates the interpretability.

Chart 2: Shapley values for the group of companies with the highest anomaly scores

Source: IIC and own elaboration.



In practice, using this automatic automated algorithm to decide if a questionnaire is anomalous or not means deciding in accordance with Table 1, which states, for example, that 41,626 company questionnaires are considered "not perfect" by the CBSO and yet the algorithm considers them statistically normal.

Table 1: Quality of the questionnaires from two perspectives

Source: Banco de España and IIC. Own elaboration.

QUALITY OF QUESTIONNAIRES

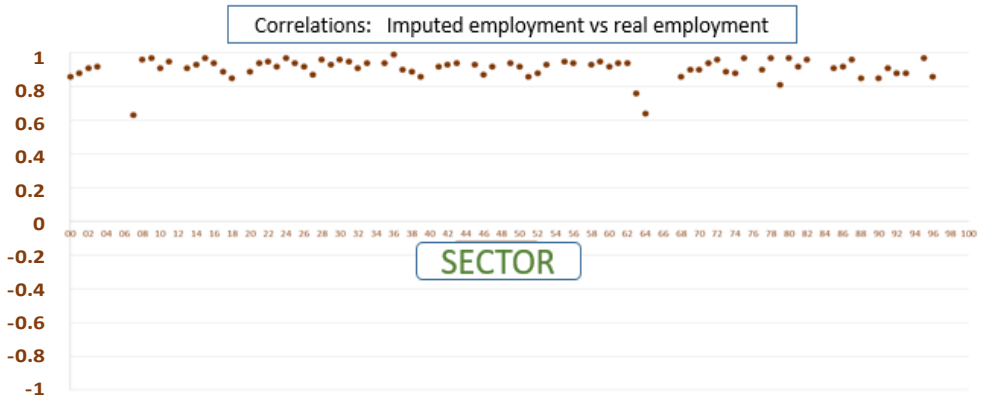
Scoring (0=right, 1=wrong)	PERFECT	NOT PERFECT	TOTAL	% TOTAL ACCUMULATED
0 - 0.1	411,973	41,626	453,599	71%
0.1 - 0.2	118,439	28,942	147,381	94%
0.2 - 0.3	20,380	5,404	25,784	98%
0.3 - 0.4	5,154	1,377	6,531	99%
> 0.4	2,299	853	3,152	100%
TOTAL	558,245	78,202	636,447	

5.2 Results obtained in the imputation of missing values

As a first approximation to the similarity between the real and the imputed values, the correlation between the two was calculated (grouped by each 2-digit NACE code) for each of the imputed variables.

A correlation close to 1 indicates a high degree of similarity. Chart 3 shows the employment correlation.

Chart 3: Correlations between real and imputed data
Source: Banco de España and IIC. Own elaboration.

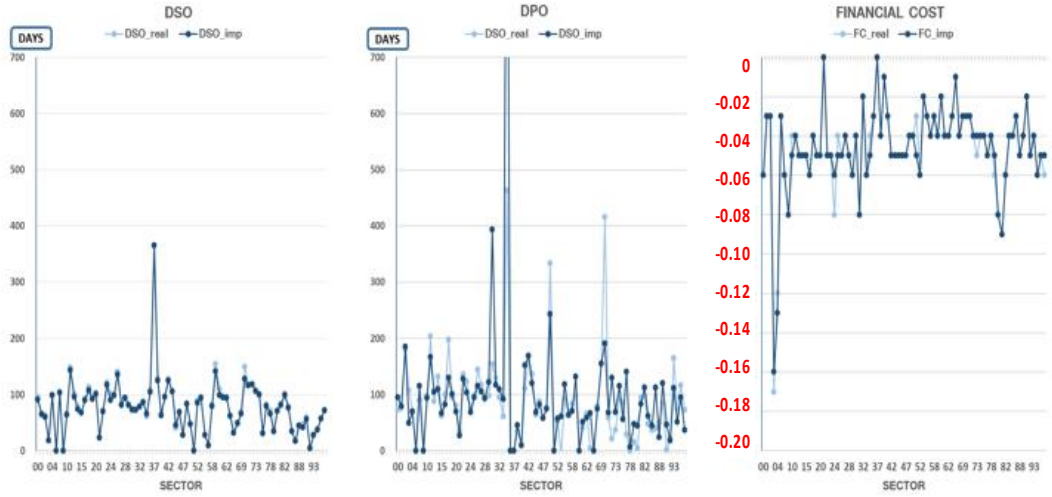


The degree of success is high when the variable considered is "Clients", but it is somewhat lower for "Suppliers" (an item on which there are fewer data).

In general, the worst ¿least successful? imputations correspond to those variables on which there are fewer data. The extreme case is "Called-up share capital", which is usually non-existent in the accounts; the correlation obtained between the real and the imputed data is zero, that is, practically at random. **Our interpretation is that when there are few data, the algorithm cannot learn**. On the contrary, it follows that if more data had been used, the imputation would have been more accurate (it should be remembered at this point that the number of observations had to be limited for the ERC because of the high computational cost).

A simulation of how some accounting ratios would stand if these imputations were used can be seen in Chart 4. For the DSO (days sales outstanding = 365*clients / sales) the adjustment is very good. It is slightly less good for the DPO (days payable outstanding = 365*suppliers / purchases) on account of the problem indicated above with suppliers. For financial costs it is reasonably acceptable.

Chart 4: Comparison of ratios calculated with real vs imputed data
Source: Banco de España and IIC. Own elaboration.



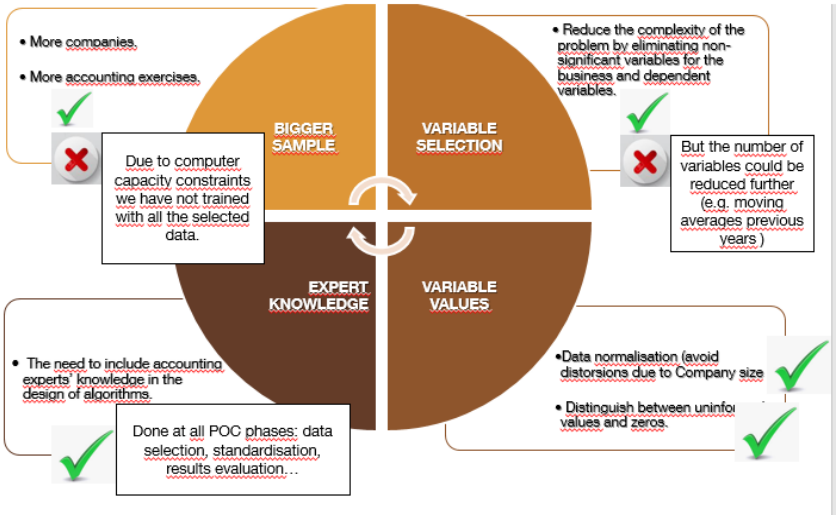
6. Conclusions and lessons learned

- Emphasis should be placed on the **feature selection**. This is normally a primary issue in the entire field of artificial intelligence and there are no magic bullets. In addition, accounting has its own difficulties, including, notably, the high degree of interrelation between the items. After all, the accounts reflect the large balance sheet items on which data have been built up over the years. That particular aspect of the accounting data affects the algorithms, but it can also provide clues that can contribute to their correct formation.
- The problem of the **distinction between empty** (non-existent) **values and values saved as zero** in the database was relevant here because the aim was precisely to impute missing values.
- The failure to normalise the data in the first tests resulted in all large corporations appearing as anomalous. A statistical normalisation would not have solved this problem because the relative scale does not vary. For subsequent analysis usual **"accounting standardisation"** was used: dividing balance sheet items by total assets and income statement items by sales. In consequence, some two million companies were dispensed with because they lacked sales figures, but the results improved substantially.
- The **computational cost** is a major problem. The cost varies significantly **depending on the algorithm chosen**, although this effect is generally exclusive to the training part of the model because, once trained, the application is usually fast. In general, the more data available the greater the accuracy, but that in turn requires more computation. IIC's access to computational resources was essential.

- **Expert accounting knowledge** was key at certain times during the process. Applying mathematically correct solutions that do not take accounting into consideration can produce strange effects that do not go unnoticed by an accountant. For example, in some instances, after imputing the addends of a summation, the summation did not match, meaning that the imbalance had to be distributed among the addends. Initially a linear distribution was made, which gave data that were clearly illogical for accounting purposes but which the system accepted with no problem. Finally the issue was corrected by distributing the mismatch proportionally to the addends. Closer collaboration can save a lot of time in such cases.

Chart 5: Lessons learned

Source: Own elaboration.



BANCO DE ESPAÑA
Eurosistema

INTERNAL USE

AI TOOLS IN OUTLIER DETECTION AND MISSING DATA IMPUTATION

POC DEVELOPED BY BANCO DE ESPAÑA´S CBSO

IFC WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING"

October 19th-22th, 2021

Pablo Jiménez
Tello Serrano

STATISTICS AND INFORMATION SYSTEMS DEPARTMENTS

INDEX

1. **Introduction. Scope of the 2019 initiative**
2. **ML models**
   - I. Anomalies detection
   - II. Missing value imputation
3. **Analysis of results**
   - I. Anomalies
   - II. Missing value imputation
4. **Lessons learned and next steps**

Questionnaires with accounting information of Spanish non-financial corporations:

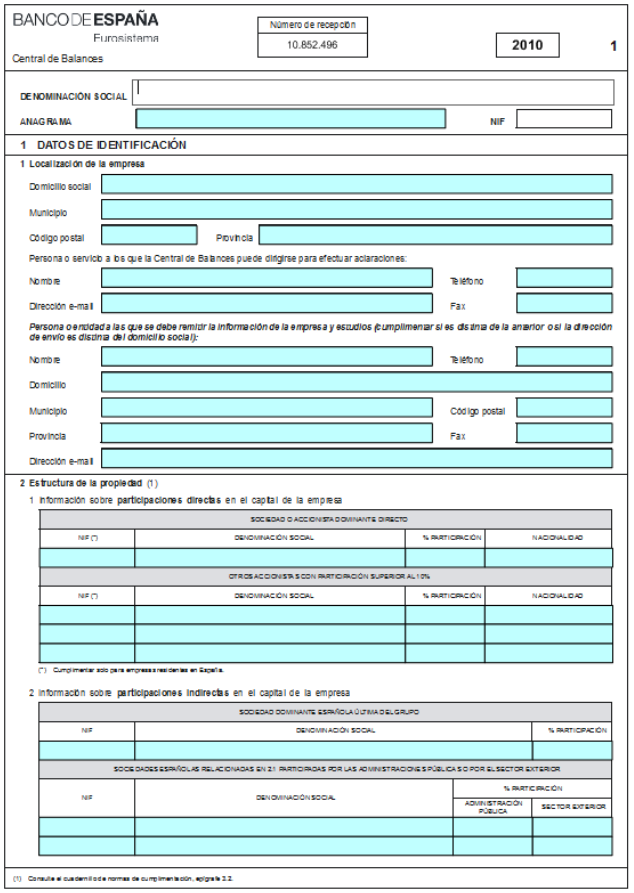
10 exercises x 900,000 companies x 3,000 data

Treated and classified by automatic processes

20% are classified as unsuitable for study

Can AI help us to improve these processes?

- Find alternative patterns to classify the questionnaires: *Case I. Anomaly detection*
- Complete the omitted information: *Case II. Value imputation*



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 solo para empresas residentes en España.

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

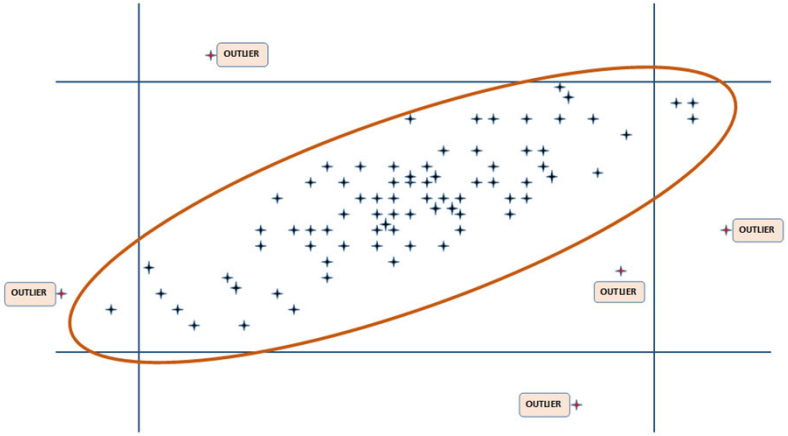
SOCIEDAD DOMINANTE ESPAÑOLA ÚLTIMA DEL GRUPO		
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN
SOCIEDADES ESPAÑOLAS RELACIONADAS EN 2.1 PARTICIPADAS POR LAS ADMINISTRACIONES PÚBLICAS O POR EL SECTOR EXTERIOR		
NIF	DENOMINACIÓN SOCIAL	% PARTICIPACIÓN: ADMINISTRACIÓN PÚBLICA / SECTOR EXTERIOR

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

RECOVER QUESTIONNAIRES FOR STUDY

ANOMALY SCORE

Anomaly index valuing n dimensions



VALUE IMPUTATION

(i) Most common imbalances and (ii) employment

ACCOUNTS TO IMPUTE		
Short term debt	=	5000
...from banks	=	?
...leasing	=	?
...others	=	?

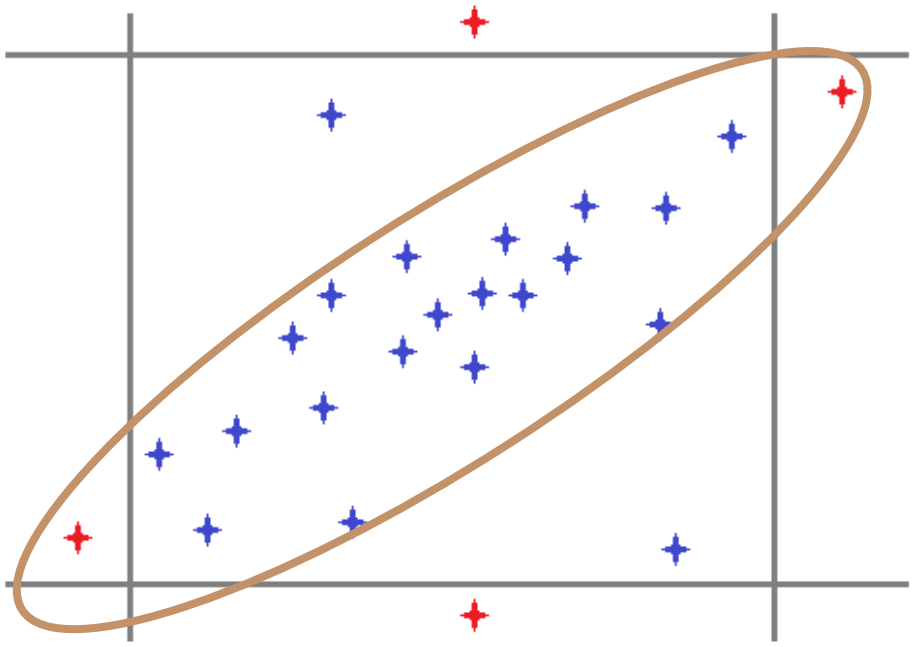
- **Variable selection: 94 accounting keys + employment + activity sector**
- **Accounting standardisation:**

  Divide the P&L fields by net revenues

  Divide the *Balance fields* between *Total Assets*

- **Generate new variables: Averages of each value in the last 2-5 years, number of declared sectors, company age…**
- **Separate questionnaires according to their quality:**
  - Perfect (5.323.000)
  - Low quality (476.000)
  - *Missing* (469.000)

1) ANOMALIES DETECTION



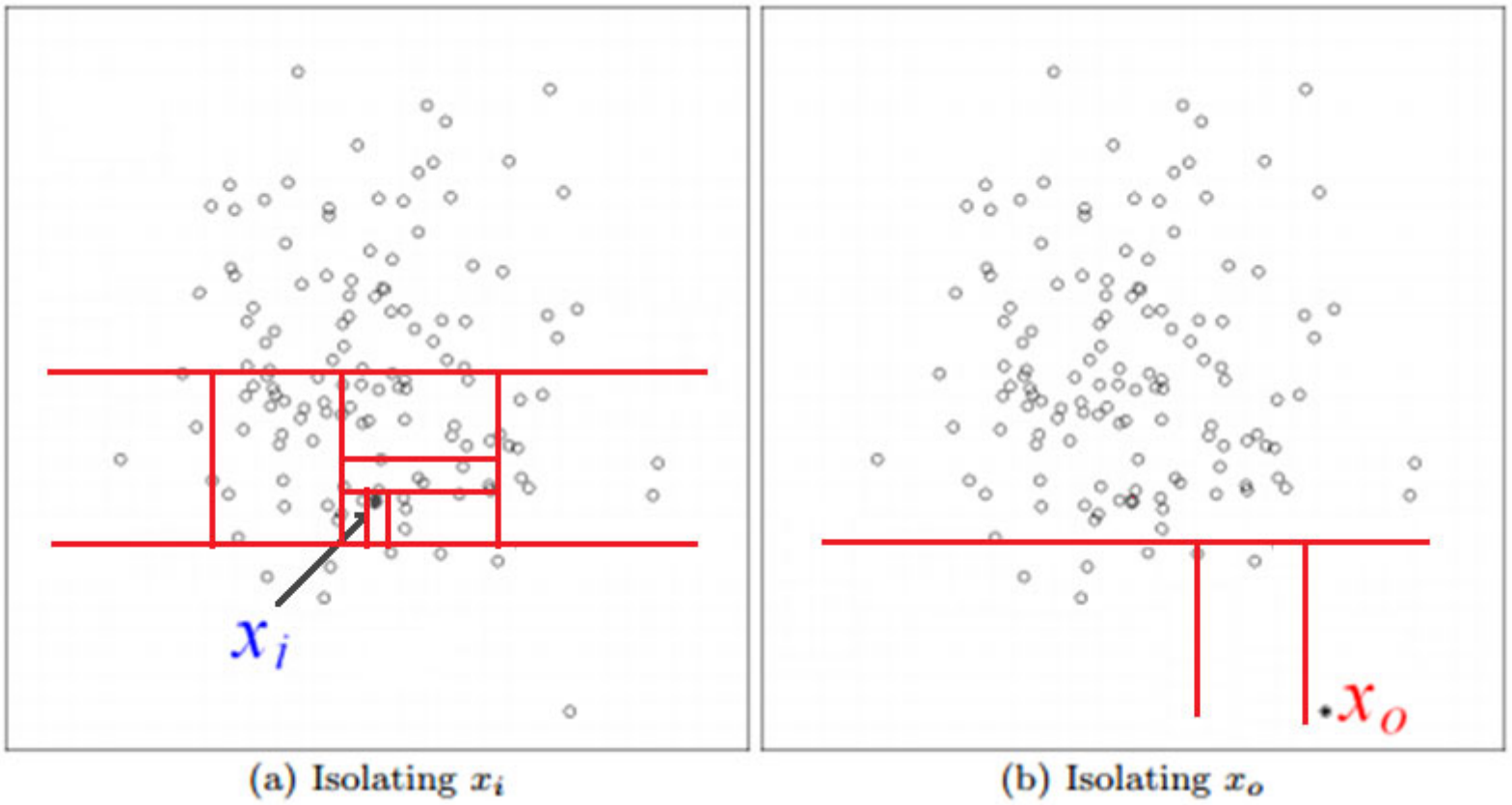
Anomaly score calculation *[0,1]* vs.
Outlier detection *(Yes/No)*

Unsupervised learning

Algorithm: **Isolation Forest**

ISOLATION FOREST

Anomalous instances are easily isolated by random divisions of space



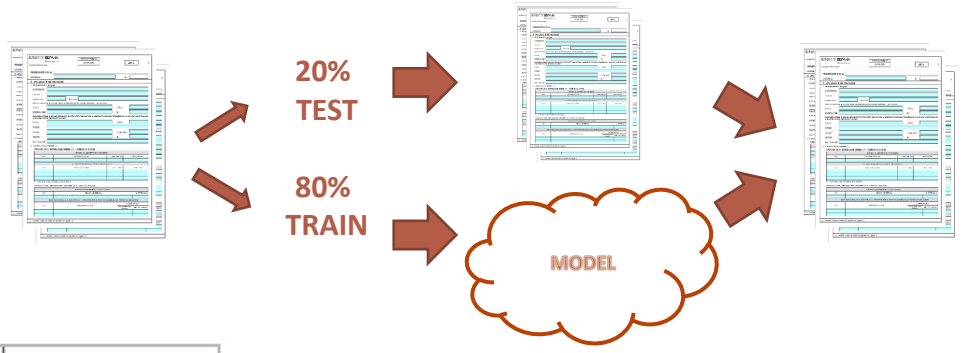
MissolationForest: custom modification of Isolation Forest algorithm to allow estimation of anomaly score when missing values are present in the data

Liu et al – Isolation Forest

2) VALUE IMPUTATION

Supervised learning

Most common imbalances and employment variables



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		

Tested algorithms:

- ❑ Variational AutoEncoder (VAE)
- ❑ Multivariate Imputation by Chained Equations (MICE)
- ❑ **Ensamble of Regressor Chains (ERC)**

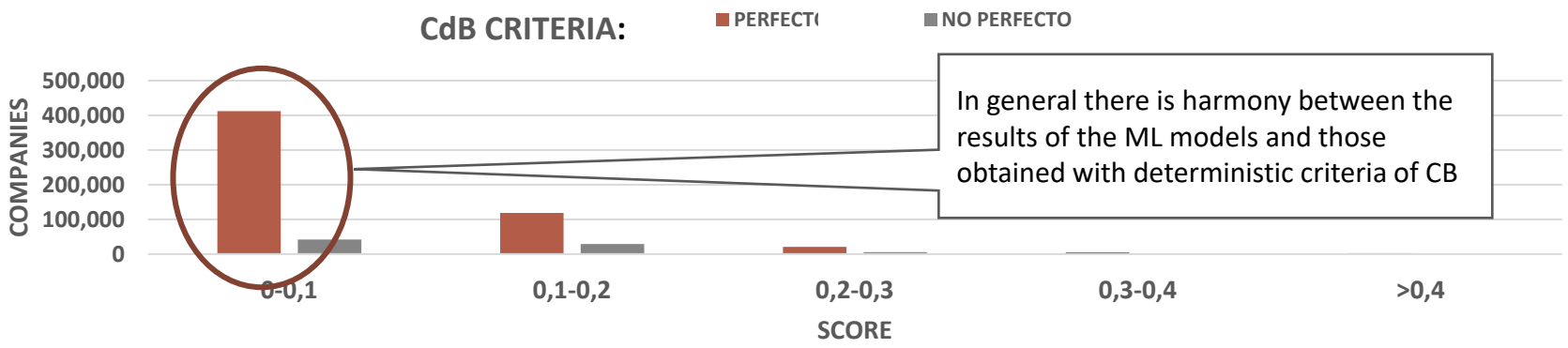
False positives? Analyze to detect possible improvements in our filtering systems

False negatives? Analyse whether or not it is necessary to relax our filtering systems

94% of the questionnaires are concentrated in a range of anomaly between 0 and 0.2

QUALITY OF CBB QUESTIONNAIRES 2017

Scoring IIC (0=Right; 1=Wrong)	PERFECT	NOT PERFECT	TOTAL	% Total accumulated
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**	

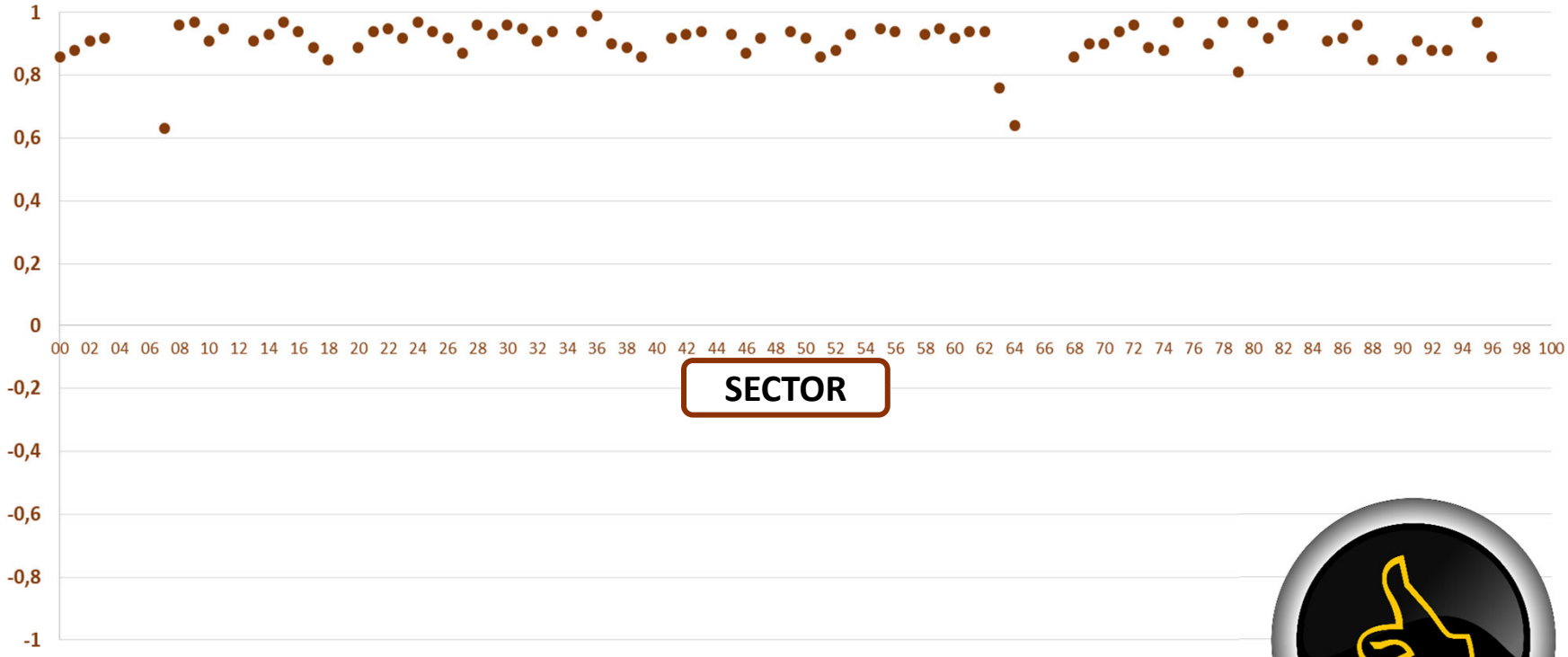


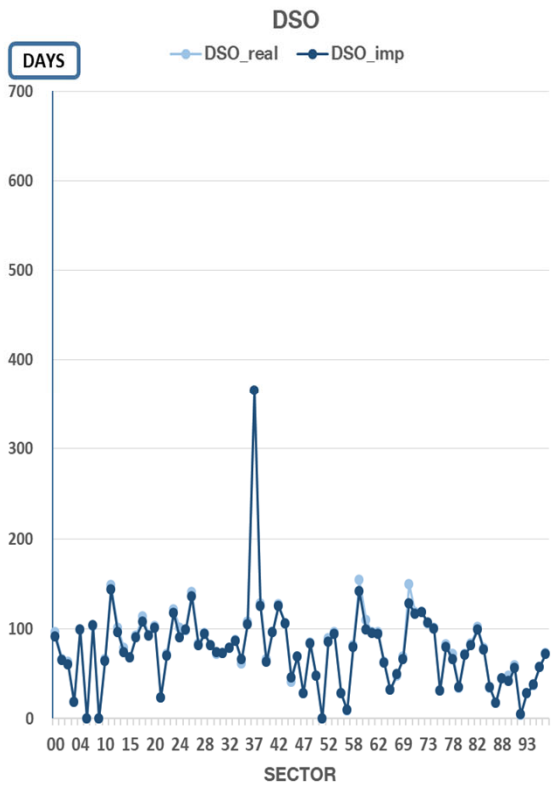
In general there is harmony between the results of the ML models and those obtained with deterministic criteria of CB

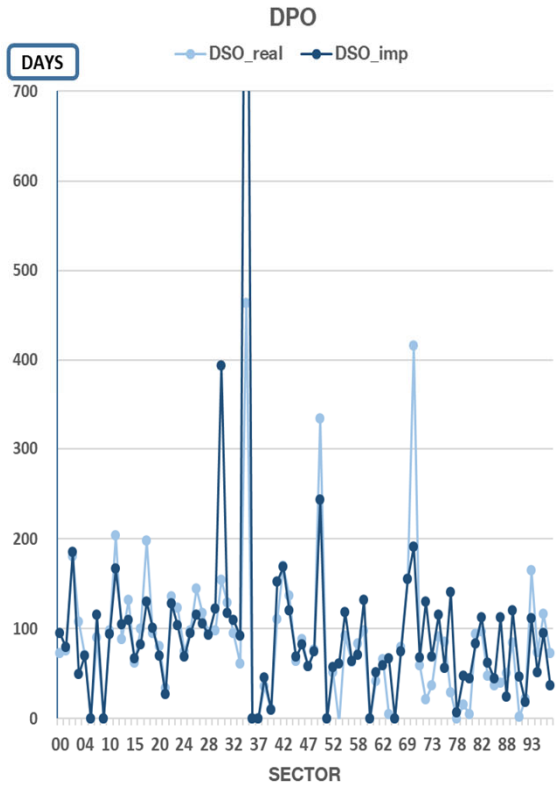
In summary:

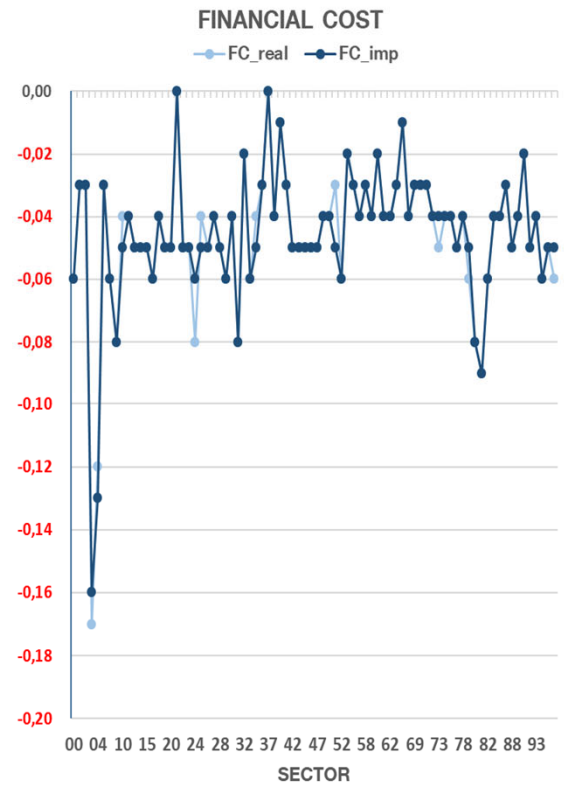
Accepting this score…	…we add or lose these companies	…giving up on these…	…and including these…
0.1	-104,646	-146,272	41,626
0.2	42,735	-27,833	70,568
0.3	68,519	-7,453	75,972
0.4	75,050	-2,299	77,349

IMPUTATION: Correlations: Imputed employment vs real employment





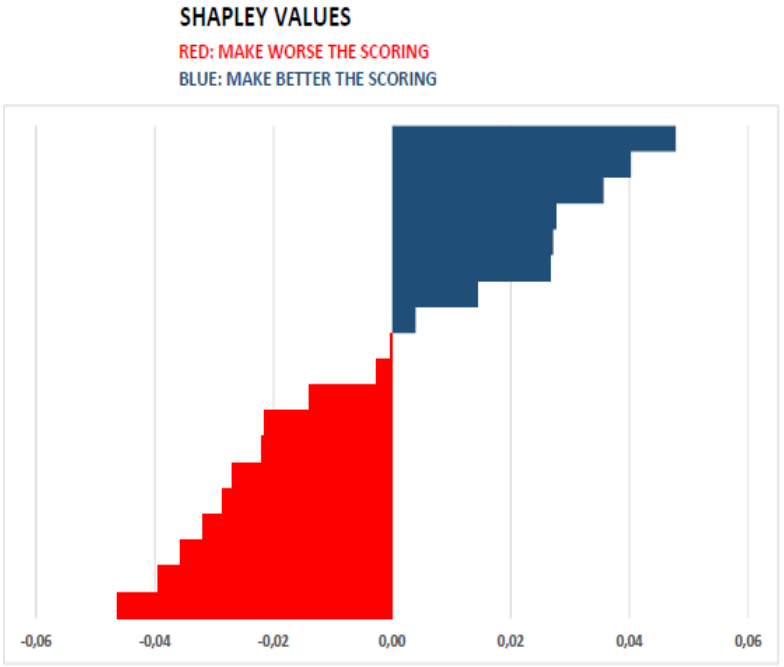




The correlations are acceptable for DSO and financial cost, but are lower for DPO, perhaps because fewer imputations have been made in the supplier key

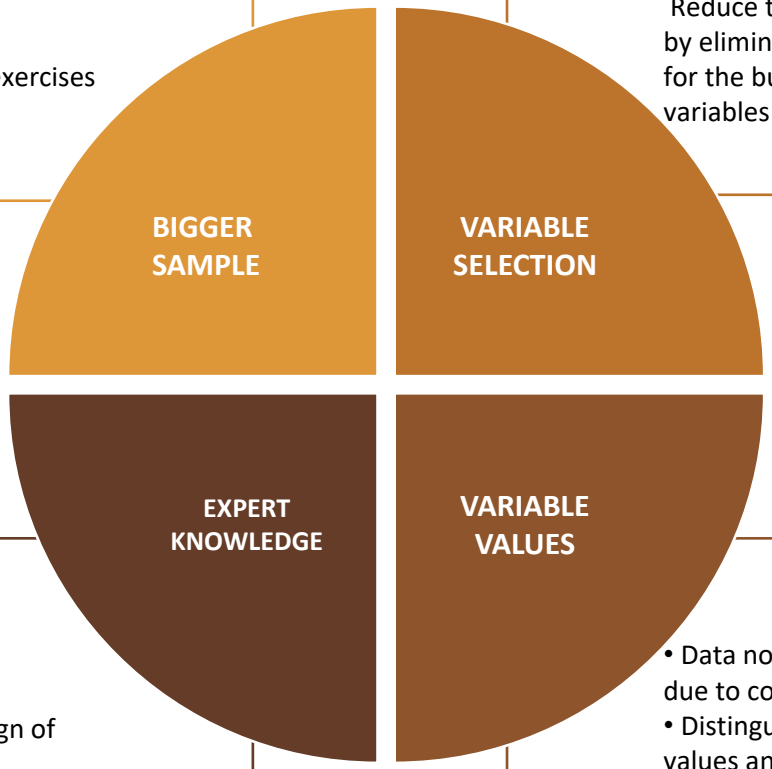
Shapley values are additive, which allow us to compute the global influence of a variable for the whole dataset or for a subset of the data

ACCOUNT	SHAPLEY VALUE
Retained earnings at the end of period(1 years moving average)	-0,05
Impairment and results from disposals of financial instruments(3 years moving average)	-0,04
Dividends(1 years moving average)	-0,04
Net wealth(3 years moving average)	-0,03
Net Income(1 years moving average)	-0,03
Financial expenditures(1 years moving average)	-0,03
Net Income(2 years moving average)	-0,02
Net Income(2 years moving average)	-0,02
Dividends(2 years moving average)	-0,01
Impairment and results from disposals of financial instruments(2 years moving average)	0,00
Bank debt(1 years moving average)	0,00
Net wealth(3 years moving average)	0,00
Impairment and results from disposals of financial instruments(1 years moving average)	0,01
Fiannacial expenditures(5 years moving average)	0,03
Bank debt(5 years moving average)	0,03
Dividends(3 years moving average)	0,03
Fiannacial expenditures(1 years moving average)	0,04
Retained earnings at the end of period(3 years moving average)	0,04
Bank debt(4 years moving average)	0,05



4. LESSONS LEARNED AND NEXT STEPS

BANCO DE ESPAÑA
Eurosistema



- More companies.
- More accounting exercises

Due to computer capacity constraints we have not trained with all the selected data

Reduce the complexity of the problem by eliminating non-significant variables for the business and dependent variables

But the number of variables could be reduced further (e.g. moving averages previous years)

The need to include accounting experts' knowledge in the design of algorithms

Done at all POC phases: data selection, standardisation, results evaluation…

- Data normalisation (avoid distorsions due to companies size)
- Distinguish between uninformed values and zeros

BANCO DE ESPAÑA
Eurosistema

THANK YOU FOR YOUR ATTENTION

STATISTICS AND INFORMATION SYSTEMS