BUSINESS SECTOR CLASSIFICATION AND BEYOND USING MACHINE LEARNING

2024

BANCO DE ESPAÑA
Eurosistema

Notas Estadísticas
N.º 18

Alejandro Morales Fernández

CONTENTS

BUSINESS SECTOR CLASSIFICATION AND BEYOND USING MACHINE LEARNING

ABSTRACT

This statistical note presents the work carried out last year by the Banco de España's Central Balance Sheet Data Office (CBSO) on the sectorisation and classification of holding companies using machine learning. This work has also been presented, in July 2023, at the World Statistics Congress (WSC) in Ottawa, organised by the International Statistics Institute (ISI), and this note is part of a series of talks on central banks organised by the Irving Fisher Committee (IFC) at the same congress.

The work presented can be divided into two parts: first, obtaining an automated procedure to help distinguish companies that, given their economic activity, are either holding companies or head offices. In other words, the aim of this work is to detect companies whose activities may come under codes 6420 or 7010 of the CNAE (Spanish National Classification of Economic Activities, equivalent to NACE, the statistical classification of economic activities in the European Community), by checking whether their data (mainly economic and financial ratios from their annual financial statements) suggest that they are or may be holding companies or head offices (whether or not they report such activities). The second part of the work is the classification of holding companies and head offices into the financial or non-financial sectors (as required by the National Accounts), using the model and information generated by the first part of the project as a starting point.

Artificial intelligence – in particular supervised machine learning classification models – is used to perform both of these tasks. A supervised model requires a prior set of labelled companies, that is to say it needs companies that have already been categorised with complete certainty as holding companies, head offices or other companies in the financial or non-financial sectors. A wide range of companies in the databases of the CBSO Division of the Statistics Department have been categorised manually, so that labelled information – an essential factor for building the model – is available.

Other essential tasks for the creation of the final machine learning model have also been performed, including the integration of various CBSO data sources and their subsequent adaptation to the structure necessary to create the model. Inter alia, variables have been selected, eliminated and transformed using statistical methods, and variables have been selected and/or eliminated for business reasons.

Finally, after the model has been constructed and evaluated, a quality control procedure is proposed. The proposed CNAE codes sometimes differ from those originally recorded. In such cases, two independent actions are proposed as a result of the model's application: the automatic classification of over 8,500 companies, where the model's result is in line with the business rules, and the manual review of approximately 5,300 other companies. As for the institutional sectorisation model, it provides a smaller set of entities for which the sector needs to be reviewed and therefore saves human effort.

The steps taken to build the proposed model, along with other technical details, are described in the annex on the technical details of the model.

Keywords: Machine Learning, Business Classification, Supervised Models, Holdings, Institutional Sectorisation, Head Offices, Data Integration, Variable Selection, Quality Control.

JEL classification: C38, C55, G23.

RESUMEN

El propósito de este documento es presentar el trabajo sobre la sectorización y clasificación de Holdings usando *Machine Learning* (en español, Aprendizaje Automático) que se ha desarrollado en la Central de Balances en el Banco de España durante el último año. Este trabajo también ha sido presentado en el *World Statistics Congress* (WSC) en Ottawa en julio de 2023, organizado por el *International Statistics Institute* (ISI). Este documento es parte de una serie de charlas sobre Bancos Centrales organizadas por el Comité Irving Fisher (IFC) en el mismo congreso.

El trabajo presentado se puede dividir en dos partes diferenciadas: en primer lugar, obtener un procedimiento automatizado que ayude a distinguir compañías como Holding o Sede Central en el contexto de Actividad Económica. En otras palabras, el propósito es detectar entidades con posibles CNAE 6420 o 7010 verificando si aquellas que declaran tales actividades muestran indicadores (ratios económicos y financieros) de serlo, y viceversa, entre aquellas que no declaran esas actividades, sus datos (principalmente sus estados financieros anuales) indican el potencial de serlo. En segundo lugar, el objetivo es realizar una sectorización institucional (es decir, la clasificación necesaria para los sistemas de Cuentas Nacionales, diferente a la mera actividad económica) de compañías Holding/Sede Central, es decir, clasificarlas en sectores Financiero/No Financiero. Para lograr esto, se utiliza como punto de partida el modelo y la información generada en la primera parte del proyecto.

Para cumplir con ambas tareas, se utiliza Inteligencia Artificial, en particular modelos de aprendizaje automático supervisado para clasificación. Un modelo supervisado requiere un conjunto previo de compañías etiquetadas, lo que significa que necesita compañías categorizadas de antemano y con total certeza como Holding/Sede/otras o Financiera/No Financiera. En las bases de datos disponibles en la Central de Balances (de ahora en adelante, CB) del Departamento de Estadística, hay una amplia gama de compañías previamente procesadas por el personal de negocio, y esto ha resultado en tener información etiquetada, un factor esencial para construir el modelo.

Además, se han realizado otras tareas imprescindibles para la creación del modelo final de aprendizaje automático. Entre ellas, está la integración de varias fuentes de datos del CB y la posterior adaptación a la estructura necesaria para la creación del modelo. Esto incluye la selección, eliminación y transformación de variables utilizando métodos estadísticos, así como la selección y/o eliminación de variables por razones de negocio.

Finalmente, después de construir y evaluar el modelo, se propone un control de calidad. Los CNAE propuestos a veces difieren de los CNAE originalmente registrados. En tales casos, se proponen dos acciones independientes como resultado de la aplicación del modelo: la asignación automática de más de 8.500 compañías donde el resultado del modelo se alinea con las reglas de negocio, y la revisión sugerida, manualmente, de aproximadamente 5.300 compañías. En cuanto al modelo de sectorización institucional, proporciona un conjunto más pequeño de entidades para revisar su sector y, por lo tanto, ahorra esfuerzo humano.

En el Apéndice: Detalles Técnicos del Modelo, se describen los pasos seguidos para llegar al modelo propuesto, junto con otros detalles técnicos.

Palabras clave: Aprendizaje Automático, Clasificación Empresarial, Modelos Supervisados, Holdings, Sectorización Institucional, Sedes Centrales, Integración de Datos, Selección de Variables, Control de Calidad.

Códigos JEL: C38, C55, G23.

BUSINESS SECTOR CLASSIFICATION AND BEYOND USING MACHINE LEARNING[1]

1 Introduction

1.1 Initial motivations

The Central Balance Sheet Data Office (CBSO) Division[2] of the Banco de España's Statistics Department collects economic and financial information, as well as other types of non-financial company data, mainly through two channels: questionnaires voluntarily sent by companies to the CBSO (CBA: Central Balance Sheet Data Office Annual Survey) and annual accounts obtained from the financial statements compulsorily filed by companies in the Mercantile Registries (CBB). The information available in the CBA is more detailed as it includes information additional to that in the annual accounts filed, but there is a much smaller number of companies available (10,000 compared with approximately 1,000,000 in the CBB database). The non-financial information obtained from both sources - with varying levels of detail - is essential for categorising companies.This information includes, for example, the number of employees, geographic location and economic activity in which the company is engaged. This document focuses on information about the economic activity carried out by companies. This information is collected in both data sources and is standardised by requesting companies to declare their activity according to the National Classification of Economic Activities (CNAE). The CNAE is the standardized classification for Spain and NACE is its equivalent in the European Community, both being fully coincident at the 3-digit level. The information that companies include in the CBA questionnaires is individually and manually reviewed and refined, unlike the information obtained from the filed accounts, which is unfeasible given the number of companies and is therefore treated and filtered applying automated methods that eliminate 20% of the annual financial statements filed.

The objective of the first work summarised in this document was to obtain an automated procedure that assists in detecting companies of two specific branches of activity, namely holding companies and head offices (HCs + HOs), which have certain specific characteristics. These branches of activity correspond to business sectors 6420 and 7010. If an entity is not classified into the two previous types, which is the majority of cases, the "Other" label is assigned, thereby providing an initial classification of the companies in the CBSO. To achieve this, Machine Learning has served as an additional component in the classification process of these types of companies, aiding in the initial classification as holding companies or head offices. The algorithm used to carry out this classification, due to its good performance, is Xgboost (extreme gradient boosting).

[1] The author is grateful to Rafael Arroyo Juan and Encarnación Colodrás Lozano for their contributions to this project.

[2] For more information, refer to the website https://www.bde.es/wbe/en/areas-actuacion/central-balances/.

The second project originates in the Directories and Publications Unit of the Banco de España due to the need to label a group of primarily small-sized companies (total assets less than 50 million euro) for which there is no information available about their shareholders or other crucial information in the current database (but there could be in their corporate documents), as explained in section 3.1. Therefore, the usual business rules used in the unit to classify these types of companies cannot be applied. The approach for this work has been similar to the business sector project. In particular, the data integration code is the same, with minor adaptations, as it is based on a subset of the same sources. For the selection of variables, previous procedures have been applied, but with improvements that are explained throughout the document. The final model is also an xgboost model, but with different variables. Additionally, techniques for interactive visualisations have been used for model interpretation and validation.

1.2 Previous work carried out by other central banks

Researching previous papers prepared by other central banks has been key for us to conduct a state-of-the-art analysis.

In 2018 the Bank of England (Noyvirt, 2018) published a paper on the classification of financial institutions. They achieved good results in some 3 and 5-digit Standard Industrial Classification (SIC) codes: 6491-0, “Financial leasing” and 6420-2, “Holding companies in production sector”.

In 2019 the central banks of Austria and Germany published two articles relating to the classification of production branches of holding companies using Machine Learning (ML) with accounting variables.

In the presentation made by the Austrian central bank (Oesterreichische Nationalbank, 2019), they first conducted data exploration and an unsupervised analysis of data from companies in their CBSO equivalent. They concluded that the holding company and real estate branches of activity have distinctive characteristics that can be distinguished from the rest using data science techniques. Therefore, they performed a supervised analysis of the same population, using various machine learning models to discriminate between these branches of activity.

The German central bank (Raulf & Schürg, 2019), however, focused directly on a supervised analysis to discriminate between Holding and Non-Holding branches of activity. They were able to successfully discriminate a large portion of these entities after applying a sequential ML model.

In 2019 the CBSO Division conducted a study based on exploratory data analysis and other visualisation techniques, including dimensionality reduction. They concluded that using appropriate machine learning techniques could lead to the automatic categorisation of holding companies. The difficulty pointed out was the proper selection of variables for the model and the fact that sometimes the classification of the economic activity is not straightforward, as there are entities that are truly on the boundary between two or more groups, and even humans have difficulty classifying them.

1.3 Preprocessing and variable selection

Before creating the machine learning model, it is necessary to have a population that meets at least two conditions:

1. It is a representative sample of the total population (in this case, the population of entities from the CBSO databases) with which an ML model will be trained for the corresponding extrapolation of Holding company / Head office / Other. This sample must contain a Holding company, Head office, or Other label based on human reviews. The labels have been encoded as 1 for Holding company, 2 for Head office, and 0 for other companies. In the institutional sector project, non-financial is encoded as 0 and financial as 1.

2. It contains a set of explanatory variables (also known as features) that meet certain characteristics, primarily: they have a certain relationship with the target or objective (Holding company/ Head office / Other, Financial / non-financial), they are a reduced set without duplicates or high correlations, and they are numerical (and if they are not, they are transformed using feature engineering methods) and are, preferably, interpretable.

With the sample mentioned in point 1 and the variables detailed in point 2, a datamart is built, that is, a reduced and high-quality dataset is achieved.

1.3.1 Data Engineering

Data engineering and transformations are essential techniques in any data science work, as they allow for the incorporation and merging of different sources. In this case, merging the sources is a straightforward task, as the Spanish ecosystem has a universal identifier for entities: The Tax Identification Number (NIF). Additionally, the CBSO has a unique identifier for each entity, which in most cases corresponds one-to-one with the NIF. This identifier has been used for the merging process. The three data sources used are as follows:

- CBA (Annual Central Balance Sheet Data Office dataset): This database provides an annual compilation of detailed financial statements (balance sheets, profit and loss accounts, etc.), along with additional complementary information. It contains approximately 12,000 entities per year that report directly to the Banco de España.

- CBB (Annual Financial Statements obtained from the Mercantile Registers): This dataset consists of individual questionnaires that are compulsorily filled out by Spanish entities and integrated as part of the Central Balance Sheet Data Office datasets, sourced from Mercantile Registers. It contains between 800,000 and 1,000,000 entities, depending on the required data quality.

- CBH (Holding Companies' Central Balance Sheet Data Office dataset): This source contains detailed annual financial statements of financial holding companies, thoroughly reviewed by business personnel due to the methodological importance of distinguishing these companies from non-financial companies in the Spanish National Accounts. On average, it contains 1,000 records per year.

In addition to these primary sources, company data has been significantly enriched with MCB concepts (Microdata integrated by the Spanish CBSO from the aforementioned datasets CBA and CBB). This auxiliary source is instrumental in providing a deeper understanding as it contains ratios and values calculated from the microdata found in the above-mentioned databases.

An important aspect of our data engineering effort was the temporal scope of our analysis. We focused on data spanning two years, specifically 2019 and 2020, which were the most recent years available when this project began. Additionally, the model is being applied and used for subsequent years, with continuous reviews on the impact and interpretation of variables.

To ensure a cohesive and comprehensive analysis, the CBA, CBH and CBB Questionnaire keys were meticulously matched using their unique key identifiers. This alignment was crucial to ensure that the model had access only to the keys common to all three sources. In total, we identified 982 common keys across these sources. Regarding the MCB concepts, we found 397 common concepts consistently present in all the databases. Figure 1 provides a schematic summary of this alignment process.

It is important to keep in mind that the data engineering process involves not only the merging of different datasets, but also other types of transformations. In our case, once the joint table is obtained, sampling is done to train the model, since the data quality of the companies in CBH and CBA is better, having been reviewed by business personnel. Other transformation and sampling tasks required for a good achievement of the supervised model have been carried out and are detailed later on in the document.

1.3.2 Feature Engineering

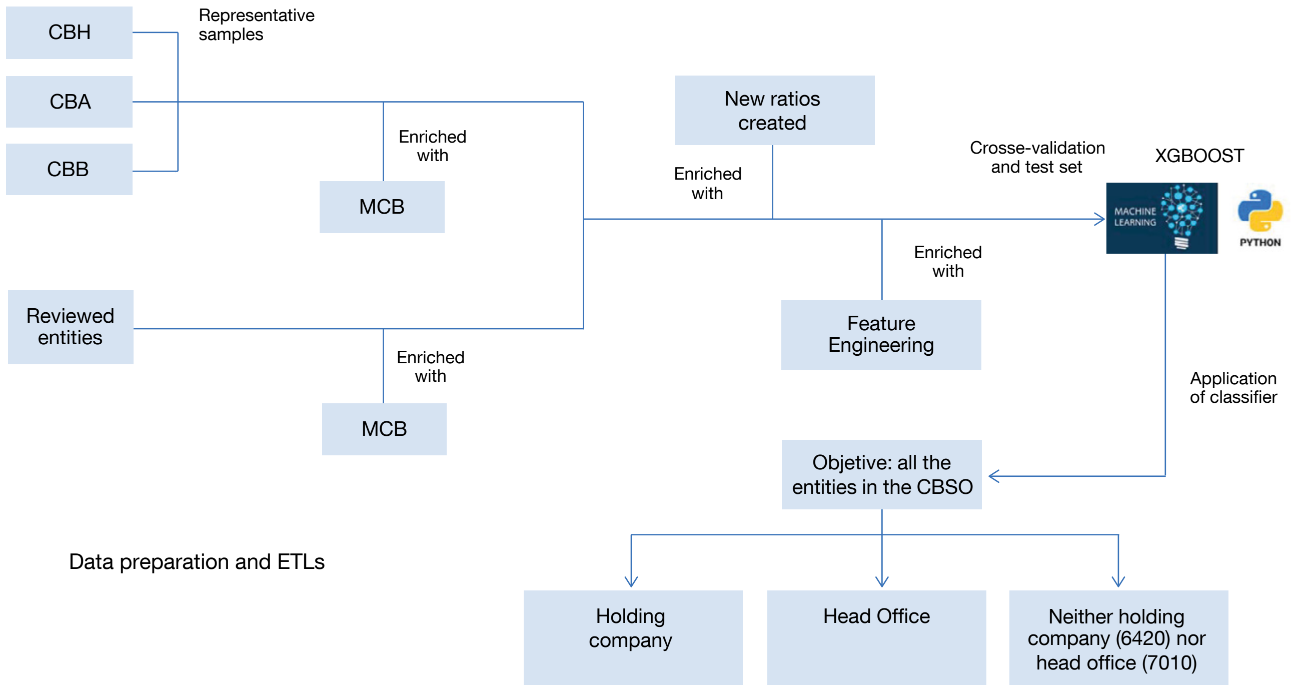
For the selection and construction of variables, the following criteria have been used:

- Elimination of variables: constant variables and those with a large number of missing values are removed.

- Variable selection:

  - Discarding variables that have a high correlation between them (70% Pearson correlation).

Figure 1
Summary of the pre-data transformation tasks required to build the ML model



SOURCE: Central Balance Sheet Data Office.

- Variables that are related to the target using Random Forest models.

- Pruning of variables using SHAP values (Shapley additive explanations). In this case, a subset of variables selected by Random Forest is evaluated for their Shapley value. This value, standardised for each variable, is used to rank and select the best variables. This method provides a much better result for selecting an optimal subset of variables from a high-quality previous subset. Therefore, Random Forest is used as a massive feature filtering technique, and Shapley values are used for fine and final selection.

— Construction of new variables: For categorical variables (such as postal code), binary variables associated with each class are created. Finally, variable selection models and human expertise determine that these variables do not contribute to the classification value for this branch of activity prediction project.

— Prioritisation of current year variables over the previous year. In case of doubt, the variable from the previous year is always prioritised for elimination instead of the current year.

For a more detailed description of the previous processes, refer to section 3.1.

2 Construction of the supervised business sector classification model

A supervised classification model is one which aims to predict a particular feature of the population called the target or objective (in the case of this project, whether a company is a holding company, head office, or other or whether a holding company is financial or not) based on previous learning from data with expert knowledge. In other words, in this case, we start with a series of companies for which it is known in advance, with certainty, whether they are a holding company, head office, or not (either because the declared CNAE code by the company has been accepted as valid, or because after a review by CBSO personnel, the most appropriate one has been chosen).

2.1 Business rules associated with holding companies and head offices

The standard business criterion for classifying companies as a holding company or head office is as follows:

— *The percentage of equity instruments in group and long-term associated companies over total assets is equal to or greater than 50%.*

In the CBSO, the above percentage or ratio is calculated by dividing equity instruments in group and long-term associated companies per year by the total assets. The numerator of the previous variable is only available in normal CBSO questionnaires. If it is not that type of questionnaire, the equivalent definition is applied:

— *The percentage of long-term investments in group companies is equal to or greater than 50%.*

The above percentage or ratio is calculated by dividing the long-term investments in group and associated companies by total assets.

In order to avoid handling two variables simultaneously and generating collinearity, the following variable is created:

— *Percentage of group investments over total assets = percentage of equity instruments in group and long-term associated companies, if available; otherwise, it imputes the value of the percentage of long-term investments in group companies.*

This variable, which combines both keys into one, summarises the information better and is preferred to be used by machine learning models, as will be seen later on.

2.2 Final distinction between holding companies and head offices based on the employment business rule

The business rule used to distinguish a holding company from a head office is based on the average employment data of the entity:

- If the Number of employees ≤ 5, then the company is classified as a holding company (CNAE 6420). Otherwise it is classified as a head office (CNAE 7010).

  At the beginning of the project, it was assumed that the balance structures of holding companies and head offices are similar, and therefore it would not be appropriate to create two separate models to distinguish them. The statistical difficulty for classification would be significant, and the gain would not be significant either since the previous business rule is sufficient to distinguish them.

  However, in later phases of the project, when examining the companies that are on the border between holding company and head office more closely, it was observed that certain companies that slightly exceeded the threshold of 5 employees in a given year still correctly retained the CNAE 6420 classification. Conversely, there are a small number of companies with 5 or fewer employees that perform head office functions.

  Through a deeper analysis, it was found that there are other variables that help distinguish holding companies from head offices, although to a lesser extent than employment. Head offices tend to have slightly lower percentages of investments within the group and higher average personnel expenses compared with holding companies, among other factors.

  Therefore, the distinction between holding company and head office is incorporated into the model itself. The supervised classification model is of a multi-class type, with the possible classes being 0 (Other), 1 (Holding company), and 2 (Head office)."

The distinction between financial holding companies and non-financial holding companies is still binary, as head offices are not considered in the analysis, as shown later on.

2.3 Final results for the business sector model

In total, data from 1,682 entities for 2019 and 2020 have been used to develop the model (see Table 1).

The reason for training with a relatively small dataset compared to the whole population is to use the highly quality data reviewed by business staff. The reason for choosing an approximate ratio of 2 to 1 in 2019 compared to 2020 is to mitigate the possible atypical effects that the year of the COVID-19 pandemic may have had.

Table 1
Number of entities by year

Year	Entities used for model training	Total entities
2019	1,083	850,984
2020	599	827,014

SOURCE: Central Balance Sheet Data Office.

Table 2
Number of entities by source

Source	Used for model training	Total available records (including two years)
CBA	597	23,972
CBB	250	1,651,357
CBH	564	2,699
Sampled reviewed by business staff	271	306

SOURCE: Central Balance Sheet Data Office.

This data come from different sources and different extractions throughout the last quarter of 2022. The origin of these companies and their volumes can be seen in Table 2.

The companies included in the "'Sample reviewed by business" group are a small set of 271 companies that have been thoroughly reviewed by CBSO staff. Therefore, they have a higher reliability. The reason why 271 companies are analysed but there are 306 records is because some entities have been analysed by both the Small Business Unit and the Large Business Unit. In all cases, the same conclusion was reached regarding the reported CNAE code, assigning 6420, 7010, or Other in different cases.

Most of the 1,411 entities from the standard sources have been analysed, but none for this specific purpose. Nevertheless, they include entities from different sources from which the model should learn and to which quality control business rules have been applied in order that the algorithm does not learn incorrectly.

Out of the total number of companies, 1,429 (85%) have been selected as the training sample, and 253 (15%) have been assigned to the test sample. The breakdown by source is shown in Table 3.

Cross-validation has been utilised in the so-called training set. Therefore, a single model has been trained for each sub-dataset in that set. The test set is designed to show unbiased metrics. In terms of the model's performance, the results in both the training and test samples can be seen in Table 4.

Table 3

Number of entities by set

Training	Test
1,429	253

SOURCE: Central Balance Sheet Data Office.

Table 4

Accuracy accomplished for each set

Sample	Accuracy %
Training	98
Test	95.7

SOURCE: Central Balance Sheet Data Office

Table 5

Confusion matrix for the training dataset

Real / Predicted	Other	Holding company	Head office
Other	727	3	4
Holding company	0	448	5
Head office	5	12	225

SOURCE: Central Balance Sheet Data Office

Table 6

Confusion matrix for the test dataset

Real / Predicted	Other	Holding company	Head office
Other	125	2	4
Holding company	0	79	1
Head office	1	3	38

SOURCE: Central Balance Sheet Data Office

It is important to note that, unlike other models with binary class, we cannot talk about false positives or false negatives here, since there are three classes. The definitions of precision, recall, F_1 Score, sensitivity, and specificity are not as well known for the case of multiclass supervised models (although generalisations do exist). Instead, it seems more intuitive to show the confusion matrices for the training and test samples (see Tables 5 and 6).

Table 7

Variables selected for the final model and their description

Variable	Description	Type
Ratio of equity instruments and investments to total assets	The numerator of the ratio is the long-term equity instruments in group and associated companies, if available. In the case of a reduced questionnaire, it is imputed as long-term investments in group companies: shares, loans to companies, securities, derivatives, or other financial assets. In both cases, it is divided by total assets	Calculated ratio
Average number of employees	Average number of employees per year	Questionnaire key
Ratio of provisions to total assets	Ratio of company's inventories in the current year divided by total assets	Calculated ratio
Ratio of fixed assets to total assets	Ratio of tangible fixed assets of the company in the current year divided by total assets	Calculated ratio
Average personnel expense	Ratio of personnel expenses during the year to total average employment	Calculated ratio
Financial income from holding companies	Financial income from holding companies	Questionnaire key
Ratio of stock to total assets	Inventory ratio of the company in the current year to total assets	Calculated ratio

SOURCE: Central Balance Sheet Data Office.

The interpretation of the above-mentioned confusion matrices is as follows: in the test sample, for example, there would be a company with a reported CNAE code of 7010, but the model predicts that it does not have either that CNAE code or the 6420 code. It can be observed that the model is fairly balanced in terms of errors, with no type of error predominating over the other. One of the main objectives pursued in this phase has been to have a high-quality sample to train the model, and this has been achieved thanks to the sample of companies analysed by the business staff, which are carefully analysed by several CBSO units. As a result of this work, a high-quality model is obtained whose main goal will be extrapolation to companies not reviewed. As shown in Section 2.7, there are a number of companies that meet certain conditions and consistently assign themselves an incorrect CNAE code.

2.4 Final variables and model interpretation of the business sector model with Shapley Values

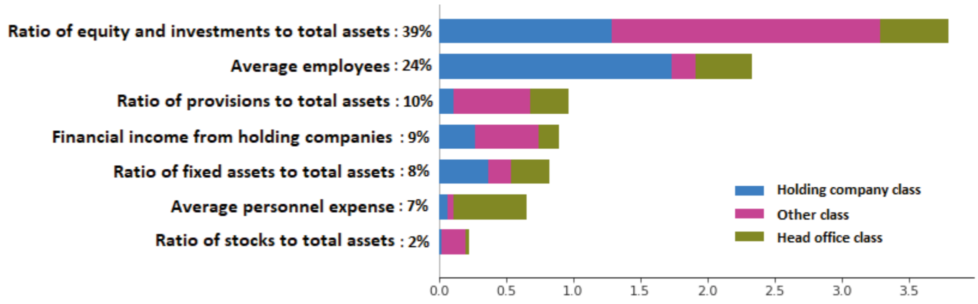
After all the variable selection processes explained in detail in Appendix 5.1, the final model incorporates the following 7 variables (see Table 7).

The influence of the variables in the model has been interpreted using the Shapley values (See Chart 1) as follows: the greater the absolute SHAP value on the x-axis of the graph, the greater the influence of the variable in the final model. If the value is positive, it will have a positive influence, while negative values indicate a negative influence on the model. The colours indicate the value of the target variable. Blue indicates low values and red indicates high values. The explanation is as follows:

1 *Ratio of investments to total assets:* It is the variable that contributes the most to the determination of the holding company and head office branches of activity, with greater emphasis on holding companies.

Chart 1

Variables of the final model and their Shapley contributions



SOURCE: Central Balance Sheet Data Office.

2. *Average number of employees:* higher in head offices and other companies.
3. *Ratio of provisions to total assets:* higher in head offices and especially in holding companies
4. *Ratio of fixed assets to total assets:* very low in holding companies
5. *Average personnel expenses:* very high in head offices.
6. *Financial income from holding companies:* almost definitive for classification as a holding company, but with numerous missing values.
7. Ratio of stock to total assets: values are very low for holding companies

The previous interpretations have been validated from a business perspective and are coherent with the accounting knowledge at the Central Balance Sheet Data Office.

2.5 Review tasks performed by business staff

Based on the model results and discrepancies with respect to CNAE codes, various review actions have been taken on the selected companies. This quality control is additional to the controls usually performed on the CBSO database.

For the integration, familiarisation, and validation of the model, it has been decided that the Treatment Units will review it in two phases: CBA (reviewed large and medium-sized companies) and CBB (non-reviewed small and medium-sized companies). By the time this paper was drawn up, the first stage had been carried out.

Table 8

Confusion matrix of the revision by source in 2020

Real / Predicted	Class	Other	Holding company	Head office
CBA	Other	11	80	24
CBA	Holding company	3	266	2
CBA	Head office	19	9	267
CBB	Other	796	17	5
CBB	Holding company	8	3	10
CBB	Head office	91	3	890
CBH	Other	98	95	3
CBH	Holding company	28	1	12
CBH	Head office	2	1	32

SOURCE: Central Balance Sheet Data Office.

The model was applied to the entire set of companies in the CBSO for 2019 and 2020 (a total of 1,677,998 entities counting the duplicates for both years), resulting in a series of actions that are explained in this subsection.

2.5.1 First Review (CBA)

Firstly, a list of companies whose questionnaires complied with the CBA model for 2019 and 2020, but did not have CNAE code 6420 or 7010 assigned, was sent to the treatment units. The SME Unit analysed a total of 172 entities, with the model achieving a 25% accuracy rate, while the Large Firms Unit analysed 146 companies, achieving 0% accuracy for companies with more than 100 employees and 53% accuracy for entities with fewer than 100 employees.

This revision led to important changes in the model and can be seen in section 5.3.

2.5.2 Second Review (CBB)

The second phase of the review –on CBB- is of vital importance, for several reasons. CBA companies already had a previously revised CNAE code, which means there is less propensity for a CNAE code change. For this very reason, CBB entities should have a bit more propensity to change. In addition, CBB companies are smaller in size, which means that they will not fall largely within entities with high average employment and turnover of more than 50 million (these thresholds have been further detailed below). Finally, by only evaluating in the model the keys and concepts present in the reduced questionnaire, the human validation will be somewhat more similar to the result of the machine, mainly because the percentage of investments in equity instruments is not available. This validation is still pending, although a quick review has been performed with good success.

The information of the pending review summarised can be seen in Table 8. The columns indicate what the conclusions of the model are and the rows indicate the official classification.

The previous decision to change a CNAE code has been thought through in detail, as we have to give some value to what an entity has reported as its CNAE code. A 90% quality control threshold of probability is chosen to change the CNAE code of an entity.

3 Construction of the supervised institutional sector classification model

The correct institutional classification of each entity is crucial in the preparation of the statistics compiled by the Banco de España, as it will impact the creation of different data aggregates produced by the CBSO and other divisions of the Statistics Department.

To determine the institutional sector, it is important to establish whether the entity has decision-making autonomy, that is, whether its main activity is carried out by the entity itself or it is subordinate to the decisions made by the direct or indirect parent company of that entity.

3.1 Business rules associated with financial holding companies and financial head offices

The criteria defined for the institutional categorisation of financial holding companies and financial head offices by the Task Force (TF) on head offices, holding companies and special purpose entities (SPEs) of the OECD, Eurostat, and ECB, in June 2013, are translated into the following business rules:

— The entity must be considered an “Institutional Unit” (IU), meaning it possesses decision-making autonomy:

- If employment > 5, it is considered an IU; therefore, it would be a head office.
- If employment <= 5, it is not considered an IU, unless its parent is non-resident. If none of its shareholders holds a stake of more than 50%, autonomy of decision is assumed by agreement and, therefore, it is an IU, it would be a holding company.
- If employment <= 5, it is not considered an IU and consolidates with its parent, excluding the cases indicated in the previous point. Only if the parent is financial will an analysis be conducted of whether it should be categorised in the sector of its parent or in the financial holding sector if the former is not possible (e.g. banks, savings banks...) and it meets the following criteria:

— The percentage of equity instruments in group companies must be more than 50% of the total assets.

— Additionally, in the case of financial head offices, the majority of their subsidiaries must be financial corporations.

As can be appreciated, the criteria go beyond using accounting variables (employment and equity). Information about their parent and subsidiaries is also used, making the definitions of financial holding companies and financial head offices more abstract than the definitions of holding companies and head offices. Therefore, throughout the project, we have doubted and learnt quite considerably about what type of variables should be influential.

3.2 Challenges and focus

Owing to its good performance in the business sector project, only the xgboost model has been used for this project. Moreover, given the few financial head offices that exist in the Spanish environment, the head offices were discarded from any analysis as very little benefit could be yielded from this.

Distinguishing between the financial and non-financial sectors in companies from different industries is statistically easier than distinguishing financial holding companies from non-financial holding companies. That is why having an excellent quality sample has been a key requirement, i.e. a sample where the financial-non-financial labels are 100% sure. A second problem arose owing to the small size of the holding companies under analysis, with some variables not having been filled out or having been filled out with a low reliability. To solve this problem, we only used well-informed variables (assets, net amount, investments, fixed assets, etc.) and no other calculated items. The variables must be well-informed in both financial, non-financial, and target population terms. Also, we try to make the model not very sensitive to entity size, which is why ratios have been widely used.

The initial models for institutional sectorisation yielded good results with very few explanatory variables. This was unusual and raised some concerns because it was known in advance that differentiating between financial and non-financial entities is not a trivial problem.

To increase the consistency of the variables included in the model, certain variables were manually eliminated. It was observed that other similar keys and concepts entered the model, leading to the conclusion that a more automated procedure would be appropriate. The conditions explained in section 3.4 were then applied.

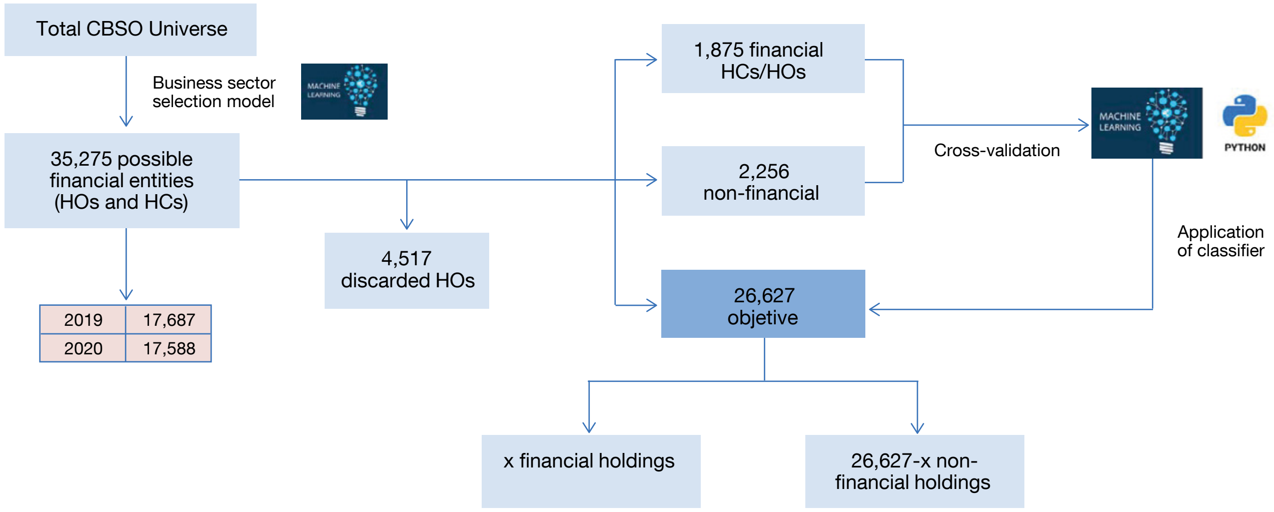
3.3 Data Engineering

We have started with the same three data sources used in the business sector model:

- CBA variables (Central Balance Sheet Data Office Annual Survey)
- CBH variables (Holding companies’ Central Balance Sheet Data Office dataset)
- CBB variables (individual questionnaire from the Mercantile Registers’ Deposit in the Central Balance Sheet Data Office)

Figure 2

Schematic summary of the data sources and subsequent actions for the present project



SOURCE: Central Balance Sheet Data Office.

Table 9

Classification of entities in the institutional sector project

Sectorised – Objective	Financial – Non-financial	Volume
Labelled	Non-financial	1,875
Labelled	Financial	2,256
Objective	Objective	26,627
Head offices	Not considered	4,517

SOURCE: Central Balance Sheet Data Office.

Additionally, the entities have been enriched with ratios and calculated variables from MCB concepts. As in the previous model, there are initially a total of 1,351 common questionnaire keys. As for the MCB concepts, there are 397 common concepts too. A schematic summary can be seen in Figure 2.

The total volume of companies included in the population comes from two different years (2019 and 2020). This subset of companies has been selected from the total number of entities in the Central Balance Sheet Data Office, imposing the condition that the business sector model resulted in the entity being a holding company. The head offices have been discarded as there are very few of them and the trade cost-benefit is very high. Out of these 35,275 records, the Directory Unit, taking into account the companies that have been manually analysed and also running an automatic institutional sectorisation software using R, has provided an initial dataset. After some quality checks, the volumes can be seen in Table 9.

Table 10
Performance of the model

%

Sample	Accuracy	F_1 score
Training	87	88
Test	83	85

SOURCE: Central Balance Sheet Data Office.

As with the business sector project, an 85% training sample and a 15% test sample have been used. Also, cross–validation is utilised.

3.4 Feature Engineering

The ideas used are similar to those mentioned in section 1.3.2, with two additional modifications:

— *Variable elimination:*

- Elimination of variables with a high proportion of constant or null values, stratified by target subset (sample vs. target sample to which the algorithm is to be applied). This improvement was necessary because the companies to which the algorithm was to be applied (target set) showed different values in the variables that the variable selection model chose as optimal. Generally, these variables were not extensively reported in the sample or had a value equal to zero.

- Elimination of variables with a high proportion of constant or null values, stratified by each source (CBA, CBB, and CBH): in this case, it is necessary to do this because certain variables take different values in the case of the CBH source, which is the source with the highest proportion of holding companies and head offices by a large margin.

3.5 Final model

The model achieves 83% of accuracy in the test sample and the main metrics can be seen in Table 10.

The final model trained on a grid described in Section 5.4 led to a model having the 10 variables in Table 11.

3.6 Variable interpretation

A SHAP value analysis has been performed as in section 2.4. The influence of the variables in the model can be seen in Chart 2.

Table 11

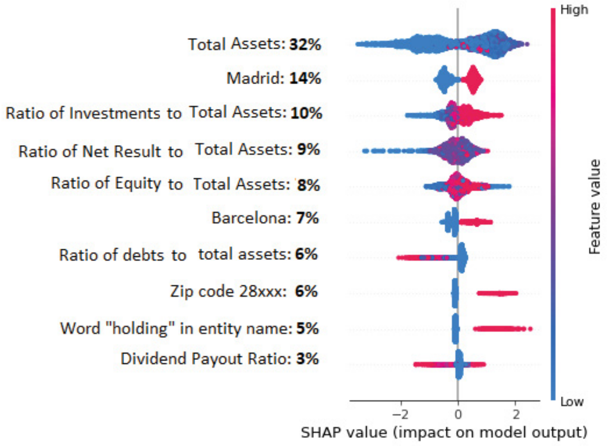
Variables selected for the final model and their description

Variable	Description	Type
Total Assets	Total assets of the company in the current year	Questionnaire Keys
Madrid Associated Postal Code (14%)	Binary Variable. Filled with 1 if the Postal Code is from Madrid, 0 otherwise	Calculated variable by one-hot encoding
Investment to total assets ratio	The numerator of the ratio is the long-term equity instruments in group and associated companies, if available. In the case of a reduced questionnaire, it is imputed as long-term investments in group companies: shares, loans to companies, securities, derivatives, or other financial assets. In both cases, it is divided by total assets	Calculated ratio
ROA	Ratio of net result of the company in the current year to total assets	Calculated ratio
Equity to total assets ratio	Ratio of equity of the company in the current year to total assets	Calculated ratio
Barcelona Postal Code	Binary Variable. Filled with 1 if the Postal Code is from Barcelona, 0 otherwise	Calculated variable by one-hot encoding
Debt to total assets ratio	Debt (Long and Short-term) divided by Total Assets in the current year	Calculated ratio
Postal code 28xxx	ZIP code associated with a particular district in Madrid	Calculated variable by one-hot encoding
Word "Holding" in entity name	1 in the entity name contains "holding", otherwise 0	Calculated variable
Dividend to net income ratio	Dividends divided by net result during the current year. The reason for using net income as the denominator instead of the distribution base is that the former key is available in both questionnaires	Calculated ratio

SOURCE: Central Balance Sheet Data Office.

Chart 2

Variables of the institutional sector model



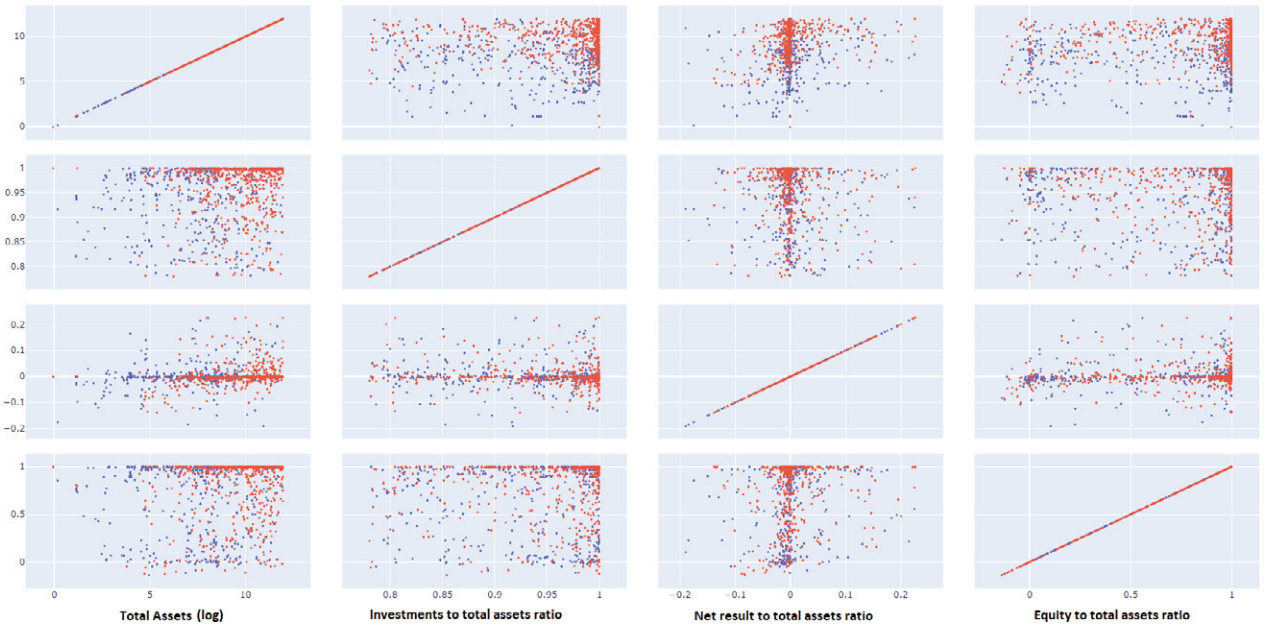
SOURCE: Central Balance Sheet Data Office.

The interpretations of the values, the impact on the model (expressed as a percentage) and additional details of the most interpretable variables are as follows:

- *Total assets (32%):* This is the only non-binary variable in the model in absolute value (not a ratio), and it represents the entity's size. It is also the denominator for most of the ratios. In a very general sense, low values of this variable indicate a positive impact on the non-financial holding company allocation.
    - Additional detail: Its average value is lower in non-financial holding companies (€13 million) compared with financial holding companies (€20 million).
- *Madrid postal code associated with Madrid(14%):* High values of this variable (i.e., residence in Madrid) indicate a positive impact on the financial holding allocation.
    - Additional Detail: 67% of financial holding companies are located in Madrid, compared with 32% of non-financial holding companies with Madrid residence.
- *Investment to total assets ratio (10%):* Low values of this variable indicate a positive impact on the non-financial holding company allocation.
    - Additional detail: Its average value is higher in financial holding companies (~92%) than in non-financial holding companies (~84%).
- *ROA, net result / income to total assets ratio (9%):* Low values of this variable indicate a positive impact on the non-financial holding company allocation.
    - Additional detail: Its average value is higher in financial holding companies (~8%) than in non-financial holding companies (~0%).
- *Equity to total assets ratio (8%):* High values of this variable indicate a positive impact on the financial holding company allocation.
    - Additional detail: Its average value is higher in financial holding companies (~68%) than in non-financial holding companies (~46%).
- *Barcelona postal code (7%):* High values of this variable (i.e., residence in Barcelona) indicate a positive impact on the Financial Holding company allocation.
    - Additional detail: Only 27% of holding companies not residing in Madrid or Barcelona are non-financial.
- *Debt (long and short-term) to total assets ratio (6%):* High values of this variable indicate a positive impact on the non-financial holding company allocation.
    - Additional detail: Its average value is higher in non-financial holding companies (~6%) than in financial holding companies (~2%).

Chart 3

Dispersion matrix plot of the top 5 interpretable variables. Financial entities in blue, non-Financial in red



SOURCE: Central Balance Sheet Data Office.

— *Postal code 28xxx (6%):* High values of this variable (i.e., residence in a particular neighbourhood of Madrid) indicate a positive impact on the financial holding company allocation.

- Additional detail: 16% of all financial holding companies are located in this district.

— *Presence of the word “holding” in the company name (5%):* High values of this variable (i.e., the name contains “holding”) indicate a positive impact on the financial holding company allocation.

- Additional detail: Higher presence in financial holding companies (~11%) compared with non-financial holding companies (~2%).

— *Dividend to net income ratio (3%):* High values of this variable in large companies (total assets greater than €100 million) indicate a positive impact on the non-financial holding company allocation.

- Additional detail: In large entities, the dividend ratio is 14.8% for financial holding companies, compared with 26.2% for non-financial ones.

In order to check the previous statements on the real data, dispersion plots have been depicted. One of them can be seen in Chart 3.

Table 12
Business revision of the institutional model

Model / Business	Financial holding companies	Non-financial holding companies
Financial holding companies	11	12
Non-financial holding companies	0	8

SOURCE: Central Balance Sheet Data Office.

3.7 Review tasks performed by business staff

For the revision, 10,938 companies common to both 2019 and 2020 were selected following the business sector model (excluding headquarters), with the condition of being a holding company in both 2019 and 2020, and with a probability threshold of 90%. Based on these questionnaires, the model concludes the following:

- 414 are deemed financially secure according to the model.
- 3.367 are deemed not financially secure according to the model.
- 7,157 do not meet the chosen quality threshold in this analysis to determine their categorization. Bear in mind that the 90% probability requirement is quite restrictive.

The business personnel have reviewed both non-financial and financial companies that the model has raised (out of the previously mentioned 414 and 3.367). Among the entities checked, only 31 have enough information in our sources (questionnaires, documents and other internal and external sources) to determine the financial sector, as explained in section 3.1. (See Table 12 for more information).

The 8 companies are businesses that are not located in Madrid and have a low asset value (generally less than €100,000). In contrast, the 11+12 companies that the model identifies as financial are mostly located in Madrid and their asset value is slightly higher. Most of them also meet a ratio of investments in group companies to assets and equity to assets that is very close to 100%. It is important to highlight that this sample is biased, as small companies typically have less information available to determine their sector. Therefore, the performance metrics of the model cannot be fully measured with this analysis. Nevertheless, we can refer to the 83% of accuracy in the test sample explained in section 3.5.

4 Conclusions and lessons learnt

4.1 Conclusions

A machine learning model has been created to automatically detect holding companies and head offices, which helps better identify the CNAE codes, providing additional and robust quality control. It also constitutes the baseline population for the institutional sectorisation AI model.

The institutional sectorisation ML model serves as a powerful tool in the institutional sectorisation to validate, select and filter financial holding companies.

4.2 Lessons learnt

To achieve a good performance and interpretable model, the use of a high quality sample to train was key. In this case, a set of entities reviewed by business staff was essential for the model to learn correctly.

It is important to give some value to the NACE code declared by the company itself, and even greater value to the one recorded by business staff. Therefore, it is advisable to be conservative and only consider entities as prone to NACE changes if they meet a wide confidence threshold.

4.3 Next steps

Next steps include performing subsequent revisions of the business sector model and concluding the revision of the institutional sector project.

Additionally, this work is covered under a project of sectorisation with machine learning within the Statistics Department of the Banco de España. The next project to be covered involves early sectorisation of entities using balance of payments data. In this forthcoming project, as the amount of accounting information is scarcer, other approaches related to text mining and contextual variables will be researched, utilising NLP, semantic embeddings, and/or large language models.

5 Annex: technical details of the models

This chapter aims to explain the technical details of the algorithms and techniques used throughout the project, as well as how and why they have been chosen. It also aims to detail some of the procedures or paths that have been discarded. The details and conclusions presented in this chapter apply to both models, although the experimental part has been mostly conducted on the business sector model.

5.1 Variable selection and feature engineering

In this section, a more detailed description is provided of the various processes used to select the best variables.

5.1.1 Elimination of variables due to high correlations

The Python library “collinearity” (Malato, 2021) is used. In an iterative process, it removes variables that have a correlation higher than a certain threshold, which is requested as input. To choose which correlated variables remain in the model and which ones do not, priority is given to features that have a strong statistical relationship with the target variable (which has also been introduced in the function), ordered based on the Snedecor’s F-test (ANOVA).

After several tests together with the business staff, working with variables that are known in advance to be related, the threshold correlation coefficient was set at 70%.

5.1.2 Categorical variable treatment

A common task before creating a machine learning model is handling categorical variables, as many models do not accept such variables as input. To address this issue, in this case, the Python Library Feature-engine (Galli, 2021) has been used. This library allows for the automatic selection of the most frequent values of the categorical variables provided as input and generates the corresponding binary variables. For this project, the top 5 most frequent values of each variable have been selected, and the less important variables from each of those 5 (as well as the remaining numerical variables) have been subsequently eliminated using the other selection methods employed.

5.1.3 Missing values treatment

Most variable selection methods are regression or classification models that do not accept missing values as input, at least in the libraries used. This is the case with the Random Forest model, chosen as one of the variable selection/elimination methods.

Therefore, when using certain models, it becomes necessary to impute missing values. For this project, the decision has been made to replace missing values with zeros since, in the majority of variables, this is the true meaning of a missing value. The final xgboost model can handle missing values, so the temporary imputation is undone for the final model, which is trained using the original variables.

Imputation has also been attempted to train machine learning models that do not allow missing values as input, such as Regression Trees, Random Forest, and Logistic Regression. Ultimately, due to the good classification results of Xgboost and the fact that it does not require missing value imputation, it was chosen as the final model.

Additionally, an additional method for imputation was tested, based on the k-nearest neighbors method (using the KNNImputer module from the sklearn library, (Pedregosa, et al., 2011)). However, it was ultimately discarded due to the difficulty in interpreting some of the imputations it made.

The chosen temporary imputation method could introduce a very slight deviation when selecting the best variables or the best model. Nevertheless, the metrics of the final model are satisfactory, and the results have been validated through business analysis. Therefore, the chosen model, with the selected variables, meets the requirements of this project.

5.1.4 Variable selection and importance ranking using Random Forest and SHAP values

A Random Forest model is executed for variable selection, following the previous methods of collinearity elimination and removal of variables with constant values or many null values. The final result is a datamart with the best variables, sorted by importance. After this initial variable selection, pruning is performed using Shapley values, obtaining the optimal set of variables, as those variables are the most influential in the model.

5.1.5 Selection of the number of variables

A grid of variables is created, ranging from 5 to 20 variables. In the final phase, the business sector model with the highest accuracy had 7 explanatory variables, while the institutional sector model had 14; thus, those were the selected models. During these processes, some features were manually discarded by analyzing their lack of coherence from a business perspective.

5.2 Preliminary steps carried out prior to model construction

In this section, some of the paths taken to reach the final model are explained. Some of the ideas have been discarded for different reasons, in order to get to the best model.

5.2.1 Data partitioning and first models with training-test split and cross-validation

First, as is customary and necessary in the construction of machine learning models, a partition was made into a training set (where the model is trained and tuned) and a test set, where the metrics of the model are validated. The proportion of the training and test samples is 85% and 15%, respectively. This proportion was chosen through empirical methods, testing ranges from 80%-20% to 90%-10%. In the former case, the training set could still be increased with a corresponding improvement in the model, without affecting the test sample. In the latter case, the model trained well, but the test dataset was insufficient to validate with complete certainty.

The model is trained using cross-validation on the training set, choosing the optimal number of folds or subsets from the 4-6-8 grid.

Table 13
Comparison of models in early stages of the Business sector project

%

Model	Accuracy	Precision	Recall	F_1 Score	ROC-AUC
Extreme gradient Boosting	99.8	100	98.6	99.3	99.5
Random Forest	99.8	99.5	98.6	99.1	99.5
Logistic Regression	97.9	84.8	99.5	91.6	97.3
Decision Trees	99.7	100	97.7	98.8	98.6

SOURCE: Central Balance Sheet Data Office.

5.2.2 Decision Trees and Random Forests

The decision tree is used as a supervised classification model in multiple cases, and its usefulness lies in its simplicity and high interpretability.

The random forest is another classification model that utilises information from multiple decision trees and combines them through bagging techniques and random feature selection. Hyperparameter tuning is performed to find the best model from a parameter grid. Some of the values to be determined include the total number of trees in the model and the number of features for each tree.

The tree models helped gain a better understanding of some of the variables in the model, and random forests provide good classification metrics. However, as will be seen in the section, the model ultimately selected is Extreme Gradient Boosting.

5.2.3 Application of other classification models

Apart from the decision trees and random forests mentioned in the previous sections, with the same dataset and variables, logistic Regression Models and Extreme Gradient Boosting were trained. The results, along with the random forest, are shown in Table 13.

The three most commonly used metrics for model selection are accuracy, F1-score, and area under the ROC curve (ROC-AUC). In this case, the algorithms ranked from best to worst are Extreme Gradient Boosting (xgboost), Random Forest, Decision Trees, and Logistic Regression. Extreme Gradient Boosting slightly outperformed the others in terms of F1-score, while Extreme Gradient Boosting and Random Forest performed the best in terms of Accuracy and ROC-AUC. Therefore, Extreme Gradient Boosting was chosen. Bear in mind that the metrics indicated in this table are higher than the final ones as the labels are different; difficult entities to classify were reviewed and added to the training sample.

5.2.4 Sample balancing

Throughout both projects, accuracy has been used as the classification metric, as it provides a more intuitive understanding of the model's performance. In both projects, the labels were reasonably balanced. Otherwise, if the labels were imbalanced, using accuracy would not have been possible, and other metrics such as F-score would have had to be used or the sample would have needed to be balanced.

However, an attempt was made to increase the balance of the holding companies/ head offices sample compared to the rest of the companies in the CBA and CBH population combined. The original ratio is 1,482 holding companies / head offices versus 10,993 non-holding companies/ non-headquarters, which is 13.45%.

Different ratios were tested, including 1:5 and 1:3, but they did not yield improved results. In conclusion, the natural proportion of holding companies / headquarter companies is suitable for the business sector project. This proportion is also sufficiently good for the institutional segmentation project.

5.3 Retraining the business sector model with corrected training data

A review was conducted by the Treatment Units of the Statistics Department and more details can be seen in section 3.7. In total, both large and small companies were analysed. These companies shared the following characteristic: the model suggested a CNAE code of holding companies /head offices, while the CNAE code declared by the company or stated by the business worker was different. The results of the model on this set of companies were:

- 12% accuracy for large companies
- 25% accuracy for SMEs

The reasons for this low accuracy are as follows:

- Large entities: the model was mostly trained on small and medium-sized companies, as they were the most abundant. In subsequent modifications, it was retrained with a small sample of large companies included.
- Large number of unrevised companies in the training set: it was later learnt that companies have a certain bias in declaring their CNAE codes. That is why the final models focus on revised companies.
- As the validation is based on the revised sample (CBA source), there is less propensity for CNAE code changes. Therefore, revisions should be done on both revised and unrevised samples.

Other lessons learnt during the review were:

- Including absolute variables instead of relative (ratios) variables can be helpful.
- There is a small number of holding companies with slightly more than 5 employees and headquarters with slightly less than 5 employees. That is why subsequent models became multi-class (holding companies, head offices, other).

Similarly, a smaller review of the institutional sector model has been performed, leading to good results. Therefore, further modifications have been applied to this model for now.

5.4 Parameter grid

The xgboost models were trained with 4-6-8 cross-validation subsets and a grid of parameters:

- Min_child_weight: minimum sum of weight required in a child node.
- Subsample: subsample ratio of the training data for each iteration.
- Max_depth: maximum depth of each tree.
- Learning_rate: learning rate. Helps prevent overfitting.
- N_estimators: number of trees. Equivalent to the number of boosting iterations.

5.5 Business rules taught to the algorithm

During the business review described in the previous chapter, it was concluded that certain business indicators could help the algorithm learn. To achieve this, in the unrevised training sample, certain labels were changed based on the surpassing of certain business thresholds. After this action, the desired objective was achieved:

- If Long-term investment ratio in group companies to total assets < 35%, then, low probability of holding companies/central headquarters.
- If Equity instrument ratio over total assets (if the regular questionnaire for the entity is available) < 35%, then, low probability of holding companies / head offices.
- If Employment greater than 5 employees, then, low probability of holding company.
- If Employment less than 5 employees, then, low probability of head office.
- If Employment greater than 5 employees. Then, low probability of holding company.

- If Employment greater than 150 employees - Discarded as holding company, then, low probability of head offices.
- If Turnover - Holding-related income > 50,000,000, then, low probability of holding companies / head offices.

5.6 Interpretation and impact of variables in the model

To properly assess the impact of variables in the model and gain feedback on their behavior, Shapley values have been utilised, specifically the SHAP library (Lundberg & Lee, 2017).

Such analyses aids in understanding the variables and their influence on the model. Even some variables were eliminated manually using this tool. Finally, this method was used as the final variable pruning method. The Shapley values for both models can be seen in figures 2 and 4.

REFERENCES

ECB, E. O. (2013). *Final Report by the Task Force on Head Offices, Holding Companies and Special Purpose Entities (SPEs).*

Galli, S. (2021). Feature-engine: A Python package for feature engineering for machine learning. *The Journal of Open Source Software.* September.

Lundberg, S. M., and S.-I. Lee (2017). A Unified Approach to Interpreting Model. *31st Conference on Neural Information Processing Systems.*

Malato, G. (2021). A Python library to remove collinearity. Retrieved from https://github.com/gianlucamalato/collinearity. June

Noyvirt, A. (2018). *Machine learning for classification of financial services companies in the.* Bank of England.

Oesterreichische Nationalbank. (2019). *Which Sector is your Business dealing in? Can Machine Learning Tools predict the Business Sector Classification from Balance Sheet Data?*

Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, . . . Duchesnay. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12,* pp. 2825-2830.

Pegoraro, N., F. Benevolo, T. Gottron, I. Febbo and ECB. (2021). Supervised machine learning for estimating the institutional sectors on a large scale. *IFC-Bank of Italy Workshop on "Machine learning in central banking".*

Raulf, F., and C. Schürg. (2019). *Classifying Holding Companies in the Individual Accounts Statistics at Deutsche Bundesbank.* Deutsche Bundesbank.

TECHNICAL GLOSSARY

Term	Description
Accuracy	Proportion of correctly predicted data (in this case, companies) out of the total
Bagging	Repeated retraining of the model designed to improve stability and accuracy of algorithms. Reduces variance and prevents overfitting
Batch	An automated execution process. In this case, it would involve running the Python prediction scripts for the Business Branch, either upon user request or triggered by an event
Boosting	Combining the results of multiple (typically weak) classifiers to obtain a robust classifier. Reduces bias and variance
Data Engineering	Also known as data preprocessing or ETL (Extract, Load, Transform), it refers to a set of techniques for transforming data into its final and suitable format
Datamart	A clean and specifically created subset of data to meet specific business needs
Feature Engineering	Set of techniques related to the treatment of features (explanatory variables) prior to building a machine learning model
FN	False negative
FP	False positive
F_1 score	$2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}$
Machine Learning	Branch of Artificial Intelligence that creates systems capable of learning automatically
Missing	Missing values or data points that are not available in the dataset, which would be useful for model training in this case
One-Hot Encoding	Method for converting categorical variables into dummy variables, necessary in most machine learning models
Performance	The performance or effectiveness of the machine learning model. There are various metrics to evaluate this performance, such as accuracy, F1 score, etc.
Precision	$\frac{\text{TP}}{\text{TP} + \text{FP}}$
Pre-processing	Data preparation for training a machine learning model, including ETL tasks and feature engineering.
Random Forest	Ensemble of decision trees combined with modified bagging
Recall	$\frac{\text{TP}}{\text{TP} + \text{FN}}$
ROC	Acronym for Receiver Operating Characteristic. It is a graphical representation of two-dimensional metrics of a binary classifier system (usually sensitivity vs. specificity) as the discrimination threshold varies
ROC-AUC	Acronym for Receiver Operating Characteristic - Area Under the Curve. It is a metric in supervised models whose value equals the area under the ROC curve
Dummy Variables	Artificial binary variables created prior to a machine learning model. For example, the dummy variable sect09_64 takes the value 1 if sect09 is 64, and 0 otherwise
TN	True negative
TP	True positive
Xgboost	Extreme gradient boosting: Ensemble of decision trees combined with modified boosting

STATISTICAL NOTES PUBLISHED

1 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Registro de los Servicios de Intermediación Financiera en Contabilidad Nacional a partir de 2005. (Publicada una edición en inglés con el mismo número.)

2 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Valoración de las acciones y otras participaciones en las *Cuentas Financieras de la Economía Española.* (Publicada una edición en inglés con el mismo número.)

3 DEPARTAMENTO DE ESTADÍSTICA Y CENTRAL DE BALANCES: Registro de los Servicios de Intermediación Financiera en Contabilidad Nacional a partir de 2005. Adendum. (Publicada una edición en inglés con el mismo número.)

4 LUIS GORDO MORA Y JOÃO NOGUEIRA MARTINS: How reliable are the statistics for the stability and growth pact?

5 DEPARTAMENTO DE ESTADÍSTICA: Nota metodológica de las *Cuentas Financieras de la Economía Española.*

6 DEPARTAMENTO DE ESTADÍSTICA: Nota metodológica de las *Cuentas Financieras de la Economía Española.* SEC-2010.

7 DEPARTAMENTO DE ESTADÍSTICA: *Holdings* y sedes centrales en el marco del SNA 2008/ SEC 2010.

8 DEPARTAMENTO DE ESTADÍSTICA: Presentación de los resultados de la encuesta de satisfacción de los usuarios de las estadísticas del Banco de España.

9 DEPARTAMENTO DE ESTADÍSTICA: Los cambios en la Balanza de Pagos y en la Posición de Inversión Internacional en 2014.

10 DEPARTAMENTO DE ESTADÍSTICA: Impacto de la revisión *benchmark* 2019 sobre la capacidad/necesidad de financiación y la Posición de Inversión Internacional de la economía española.

11 DEPARTAMENTO DE ESTADÍSTICA: La estimación de los ingresos por turismo en la Balanza de Pagos.

12 DEPARTAMENTO DE ESTADÍSTICA: Revisión extraordinaria de las *Cuentas Financieras de la Economía Española* (2019).

13 DANIEL SÁNCHEZ MENESES: The advantages of data-sharing: the use of mirror data and administrative data to improve the estimation of household external assets/liabilities (2020).

14 Ana Esteban e Ignacio González: Efectos de la aplicación de la NIIF16 sobre arrendamientos en los grupos cotizados españoles no financieros (2020).

15 ROBERTO BADÁS ARANGÜENA: La inversión exterior directa en España: ¿cuáles son los países inversores inmediatos y cuáles los últimos? (2021).

16 JAVIER JAREÑO MORAGO: Notas estadísticas relativas a las series históricas de los tipos de interés del Banco de España 1938-1998 (2022).

17 BORJA FERNÁNDEZ-ROSILLO SAN ISIDRO, EUGENIA KOBLENTS LAPTEVA Y ALEJANDRO MORALES FERNÁNDEZ: Micro-database for sustainability (ESG) indicators developed at the Banco de España (2022).

18 ALEJANDRO MORALES FERNÁNDEZ: Business Sector Classification And Beyond Using Machine Learning (2024).

ISSN 2530-7495 (edición electrónica)