BANCO DE ESPAÑA
Eurosistema

PUBLIC

Dirección General de Economía y Estadística

09/10/2023

Business Sector Classification And Beyond Using Machine Learning

Alejandro Morales Fernández (*)
Central Balance Sheet Data Office Division
Statistics Department

In this document, the classification and sectorization projects of holdings using machine learning in the Statistics Department of the Bank of Spain are summarized. This work has been presented at the WSC (World Statistics Congress) in Ottawa in July 2023

(*) The author is grateful to Rafael Arroyo Juan and Encarnación Colodrás Lozano for their contributions to this project

Executive summary

The objective of the presented work consists of two parts: firstly, obtaining an automated procedure that helps distinguish companies as either Holding or Head Office (also named headquarters) economic activities. In other words, the purpose is to detect companies with possible CNAE (National Classification of Economic Activities, equivalent to international NACE codes) 6420 or 7010 by verifying if those declaring such activities show indicators (economic and financial ratios) of being one, and vice versa, among those not declaring those activities, their data (mainly their annual financial statements) indicate the potential of being so. Secondly, the objective is to perform an institutional sectorization (that is, the classification needed by the systems of National Accounts, different to the mere economic activity) of Holding/Head Office companies, i.e., classify them into Financial/Non-Financial sectors. To achieve this, the model and information generated by the first part of the project are used as a starting point.

To fulfill both tasks, Artificial Intelligence is used, in particular supervised machine learning models for classification. A supervised model requires a prior set of labeled companies, meaning it needs companies categorized in advance and with total certainty as Holding/Headquarters/other or Financial/Non-Financial. In the databases available in the Central Balance Sheet Data Office Division (from now on, CB) of the Statistics Department, there is a wide range of companies previously processed by the business personnel, and this has resulted in having labeled information, an essential factor for building the model.

In addition, other essential tasks have been performed for the creation of the final machine learning model. Among them, is the integration of various data sources from the CB and the subsequent adaptation to the necessary structure for model creation. This includes the selection, elimination, and transformation of variables using statistical methods, as well as the selection and/or elimination of variables for business reasons.

Finally, after constructing and evaluating the model, a quality control is proposed. The proposed CNAEs sometimes differ from the originally recorded CNAEs. In such cases, two independent actions are proposed as a result of the model's application: the automatic assignment of over 8,500 companies where the model's result aligns with the business rules, and the suggested review, manually, of approximately 5,300 other companies. As for the institutional sectorization model, it provides a smaller set of entities to review its sector and therefore saves human effort.

In the Appendix: Technical Details of the Model, the steps taken to reach the proposed model are thoroughly described, along with other technical details.

INDEX

1 Introduction

1.1 Initial motivations

The Central Balance Sheet Data Office Division (CB), within the Statistics Department in Banco de España, 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 CB (CBA: annual central balance) and annual accounts obtained from the financial statements filed compulsorily by companies in the Mercantile Registries (CBB). The information available in CBA is more detailed as it includes additional complementary information compared to the annual accounts filed, but there is a much smaller number of companies available, 10,000 compared to the 1,000,000 in the CBB database. The non-financial information obtained from both sources - with varying levels of detail - is essential for categorizing companies; among this information, for example, are the number of employees, geographic location, and the 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 standardized by requesting that companies declare their activity according to the National Classification of Economic Activities (CNAE). This CNAE is a standardized classification for Spain and internationally (NACE, in this case, both 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 therefore is treated and filtered applying automated methods that eliminate a 20% of the filled annual financial statements.

The objective of the first work summarized in this document has been to obtain an automated procedure that assists in detecting companies of two specific branches of activity, namely Holding and Head Offices (HC + HO), which have certain specific characteristics. These branches of activity correspond to business sectors 6420, 7010. If an entity is not classified into the two previous types, which is the majority of the cases, a label called “Other” has been assigned, thereby providing an initial classification of the companies in the Central Balance Sheet Office. To achieve this, Machine Learning has served as an additional component in the classification process of this type of companies, aiding in the initial classification of Holdings or Central Offices. The algorithm used to classify, due to its good performance is Xgboost (Extreme gradient Boosting).

On the other hand, the second project originates in the Directories and Publications Unit of the Bank of Spain due to the need to label a group of primarily small-sized companies (total assets less than 50 million euros) for which there is no information available about their shareholders and other crucial information in the current database (but could be in the entity documents), as explained in section 3.1. Therefore, the usual business rules used in the unit to classify this type 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 explained throughout the document. The final model is also an xgboost model, but with different variables. Additionally, techniques for interactive visualizations have been used for model interpretation and validation.

1.2 Previous work carried out by other central banks.

It has been a key point to research the previous works other Central Banks have done to have a good State of the Art on this analysis.

In 2018, the Bank of England (Noyvirt, 2018) published a paper for the classification of Financial Entities. They achieved good results in some 3 and 5 SIC digits (Standard industrial classification of economic activities): “Financial leasing” and 6420-2 “Holding Companies in Production Sector”

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

In the presentation made by the Austrian central bank (Oesterreichische Nationalbank, 2019), they first conducted data exploration and unsupervised analysis on data from companies in their equivalent of the Central de Balances. They concluded that the Holding 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 on the same population, using various machine learning models to discriminate these branches of activity.

The German central bank (Raulf & Schürg, 2019), on the other hand, 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 Central Balance Sheet Data Office Division conducted a study based on exploratory data analysis and other visualization techniques, including dimensionality reduction. They concluded that using appropriate machine learning techniques could lead to the automatic categorization 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 Contains a representative sample of the total population (in this case, the population of entities from the databases of the CB) with which an ML model will be trained for the corresponding extrapolation of Holding / Head Office / Rest. This sample must contain a label of Holding, Head Office, or Rest based on human reviews. The labels have been encoded as 1 for Holding, 2 for Head

Office, and 0 for the rest of the companies. In the Institutional Sector project, non-Financial is encoded by 0 and Financial is encoded by 1.

2 Contains a set of explanatory variables (also known as features) that meet certain characteristics, primarily: have a certain relationship with the target or objective (Holding / Head Office / Rest, Financial / non-Financial), be a reduced set without duplicates or high correlations, and be numerical (and if they are not, they are transformed using Feature Engineering methods) and, 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.

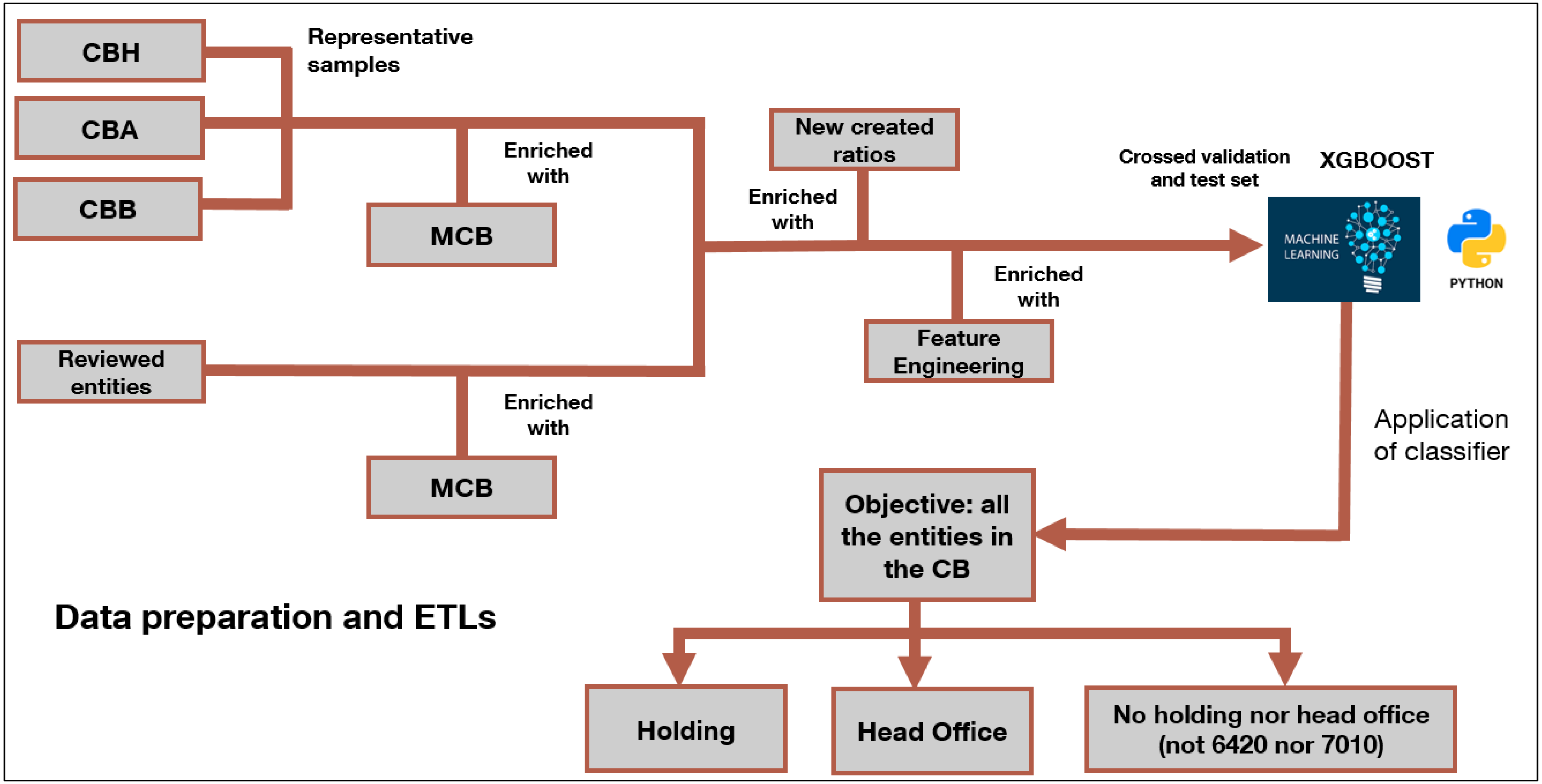
1.3.1 Data Engineering

We started with 3 different data sources:

- CBA (Annual Central Balance Sheets)
- CBH (Holding Companies' Central Balance Sheets)
- CBB (Individual Questionnaire from the Central Balance Sheets obtained from the Mercantile Registers)

Additionally, company data has been enriched with MCB concepts (Microdata from the Central Balance Sheets). This source contains ratios and values calculated from the microdata contained in the three previous databases.

The CBA, CBH, and CBB Questionnaire keys have been matched using their identifiers so that the model only has access to the keys common to the 3 mentioned sources. In total, there are 982 common keys. As for the MCB concepts, there are always 397 common concepts for all the sources. A schematic summary can be seen in Figure 1.



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

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).
  - 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, standardized 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.
- Prioritization of current year variables over the previous year. In case of doubt, the variable from the previous year is always prioritized 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 whose objective is 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, Head Office, or Rest or whether a Holding is Financial or not) based on previous learning from labeled 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, head office, or not (either because the declared CNAE by the company has been accepted as valid, or because after a review by CB personnel, the most appropriate one has been chosen).

2.1 Business rules associated to Holdings and Head Offices

The standard business criterion for classifying companies as Holding or Head Office is as follows:

- *Percentage of Equity Instruments in Group and Long-term Associated Companies over Total Assets greater than or equal to 50%.*

In the CB, the above percentage or ratio is calculated by dividing Equity Instruments in Group and Long-term Associated Companies per Year over the Total Assets. The numerator of the previous variable is only available in normal CB questionnaires. If it is not that type of questionnaire, the equivalent definition is applied:

- *Percentage of long-term investments in group companies greater than or equal to 50%.*

The above percentage or ratio is calculated by dividing the long-term Investments in Group and Associated Companies over the 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 Percentage of Long-term Investments in Group Companies.*

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

2.2 Final distinction between Holdings and Head Offices through the employment business rule

The business rule 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 (CNAE 6420). It is classified as a Head Office (CNAE 7010) otherwise*

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 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 to Holdings, among other factors.

Therefore, the distinction between Holding 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 (Rest), 1 (Holding), and 2 (Head Office)."

The distinction of Financial Holdings from non-Financial Holdings is still binary, as Head Offices are not considered in the analysis as it will be shown later.

2.3 Final Results for the Business Sector Model

In total, for the development of the model, data from 1,682 entities from the years 2019 and 2020 have been used, distributed as follows:

Year	Used for model training	Total available
2019	1,083	850,984
2020	599	827,014

Table 1: number of entities by year

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 pandemic may have had.

This data come from different sources in different extractions throughout the last quarter of 2022. The origins of these companies and their volumes are as follows:

Source	Used for model training	Total available records (including two years)
CBA	597	23,972
CBB	250	1,651,357
CBH	564	2,669
Sampled reviewed by business staff	271	306

Table 2: number of entities by source

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

The 1,411 entities from the standard sources have been most of them analyzed, but not of them for this specific purpose. Nevertheless, they include entities from different sources from which the model should learn and have been applied quality control business rules in order the algorithm does not learning 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 below:

Training	Test
1,429	253

Table 3: number of entities by set

In terms of the model's performance, the results in both the training and test samples are as follows:

Sample	*Accuracy*
Training	98,0%
Test	95,7%

Table 4: accuracy accomplished for each set

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 generalizations do exist). Instead, it seems more intuitive to show the confusion matrices for the training and test samples, respectively:

Real \ Predicted	Rest	Holding	Head Office
Rest	727	3	4
Holding	0	448	5
Head Office	5	12	225

Table 5: confusion matrix for the training dataset

Test dataset:

Real \ Predicted	Rest	Holding	Head Office
Rest	125	2	4
Holding	0	79	1
Head Office	1	3	38

Table 6: confusion matrix for the test dataset

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 of 7010, but the model predicts that it does not have that CNAE, nor the 6420. It can be observed that the model is fairly balanced in terms of errors, with no one 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 analyzed by the business staff, which are carefully analyzed by several CB units. As a result of this work, a high-quality model is obtained, whose main objective will be extrapolation to not reviewed companies. As will be seen in section 2.7, there are a number of companies that meet certain conditions and consistently assign themselves an incorrect CNAE.

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 the appendix 5.1, the final model incorporates the following 7 variables:

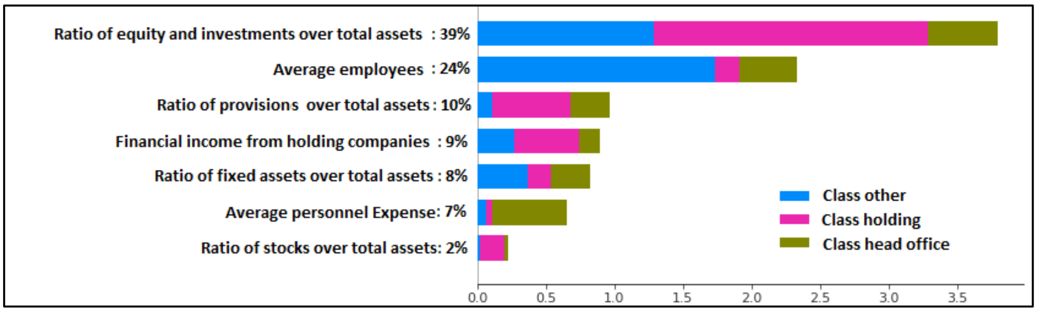
Variable	Description	Type
Ratio of equity instruments and investments over total assets	The numerator of the ratio is the long-term equity instruments in group and associated companies, if available. In the case of a reduced questionnaire, it is imputed as long-term investments in group companies: shares, 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 over total assets	Ratio of company's inventories in the current year divided by total assets	Calculated ratio
Ratio of Fixed Assets over 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 divided by total average employment	Calculated ratio
Financial income from holding companies	Financial income from holding companies	Questionnaire Key
Ratio of Stock over Total Assets	Inventory ratio of the company in the current year divided by total assets	Calculated ratio

Table 7: variables selected for the final model and their description

The influence of the variables in the model has been interpreted using the Shapley values (whose results are shown in Figure 2) 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 an inverse influence in the model. The colors indicate the value of the target variable. Blue indicates low values, and red indicates high values. The explanation follows next:

1 **Ratio of investments over total assets:** It is the variable that contributes the most to the determination of the Holding and Headquarters branches of activity, with greater accent on the Holding companies.
2 **Average number of employees:** higher in Headquarters and other companies.
3 **Ratio of provisions over total assets:** higher in Head Offices and Specially in Holdings
4 **Ratio of Fixed Assets over Total Assets:** very low in Holding companies
5 **Average Personnel Expense:** very high in head offices.
6 **Financial income from holding companies:** almost definitive for it to be classified as a holding, but with numerous missing values.
7 **Ratio of Stock over Total Assets:** values are really low in holdings

The previous interpretations have been validated from a business perspective and are coherent to the knowledge of accounting in the Central Balance Sheet Office.



2 Variables of the final model

2.5 Review tasks performed by business staff

Based on the model results and discrepancies with respect to CNAEs, various review actions have been taken on the selected companies. This constitute an additional quality control to the ones usually performed on the CB database.

For the integration, familiarization, 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 written, the first stage was carried out.

The model was applied to the entire set of companies in the CB of both years 2019 and 2020 (a total of 1,677,998 entities counting the duplicates for both years), and from its execution, a series of actions can be derived and 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 6420 and 7010 assigned, was sent to the treatment units. The SME Unit analyzed a total of 172 entities, with the model achieving a 25% accuracy rate, while the Large Unit analyzed 146 companies, achieving 0% accuracy for companies with more than 100 employees and 53% accuracy for entities with less 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 review –on CBB- is of vital importance, for several reasons. CBA companies already had a previously revised CNAE, which means there is less propensity for CNAE 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 specified as 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 summarized can be seen in the following table. The columns indicate what the conclusions of the model are, whereas the rows indicate the official classification:

Source		Other	Holding	Head Office
CBA	**Other**	11,347	80	24
CBA	**Holding**	3	266	2
CBA	**Head Office**	19	9	267
CBB	**Other**	795,783	17	5,197
CBB	**Holding**	8,341	3,363	10
CBB	**Head Office**	91	3	890
CBH	**Other**	98	95	3
CBH	**Holding**	28	1,027	12
CBH	**Head Office**	2	1	32

Table 8: confusion matrix of the revision by source in 2020

The previous decision to change a CNAE has been though thoroughly, as we have to give some value to what an entity has reported as its CNAE. A 90% quality control threshold of probability is chosen to change the CNAE 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 Bank of Spain, as it will impact the creation of different data aggregates produced by the Balance Sheet Central 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, if its main activity is carried out by the entity itself or if, on the contrary, its activity is subordinate to the decisions made by the direct or indirect parent company of that entity.

3.1 Business rules associated to Financial Holdings and Financial Head Offices

The criteria defined for the institutional categorization of financial holdings and financial head offices, in the Task Force (TF) on Head Offices, Holding Companies and Special Purpose Entities (SPE's) 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 they possess decision-making autonomy:

o Employment > 5, it is considered an IU, therefore, it would be a head office.

o Employment <= 5, it is not considered an IU, unless its parent is non-resident or those in which none of its shareholders holds a stake of more than 50%, autonomy of decision is assumed by agreement, and therefore, they are IU, it would be a holding company.

o Employment <= 5, it is not considered an IU and consolidates with its parent, excluding those from the previous point. Only if the parent is financial, it will be analyzed whether it should be categorized 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.

• And 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 Holding and Head Offices considerably more abstract than those of Holding and Head Offices. Therefore, throughout the project, we have doubted and learned quite considerably about what type of variables should be influential for this.

3.2 Challenges and focus

For this project, only the xgboost model has been used, given its good performance in the business sector project. 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.

To distinguish the Financial and Non-Financial sectors in companies from different industries is statistically easier than distinguishing Financial Holdings and Non-Financial Holdings. That is why a sample of excellent quality has been a key requirement, i.e. a sample where the labels Financial-Non Financial are 100% sure. A second problem arose given the small size of the Holdings to analyze, some variables had not been filled out or have been filled out with low reliability. To solve this problem, we used only well-informed variables (assets, net amount, investments, fixed assets, etc.) and no other calculated concepts. The variables must be well-informed in both financial, non-financial, and target population. Also, we try to make the model not very sensible to the entity size, that is why ratios have been widely used.

The initial models for institutional sectorization 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

Just like in the Business sector project, we have started with the same three data sources from the Business sector model:

- CBA variables (Central de Balances Anual -Annual Central Balance Sheet Office)
- CBH variables (Central de Balances Holding -Holding Central Balance Sheet Office))
- CBB variables (Individual questionnaire from the Mercantile Registers' Deposit in the Balance Sheet Central)

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 3**¡Error! No se encuentra el origen de la referencia.**.

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 Balance Sheet Central, 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 really 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 analyzed, and also running an automatic institutional sectorization software using R, has provided an initial Dataset. After some quality checks, the following values for this dataset:

Sectorized – Objective	Financial – Non- Financial	Volume
Labelled	Non-Financial	1,875
Labelled	Financial	2,256
Objective	Objective	26,627
Head Offices	Not considered	4,517

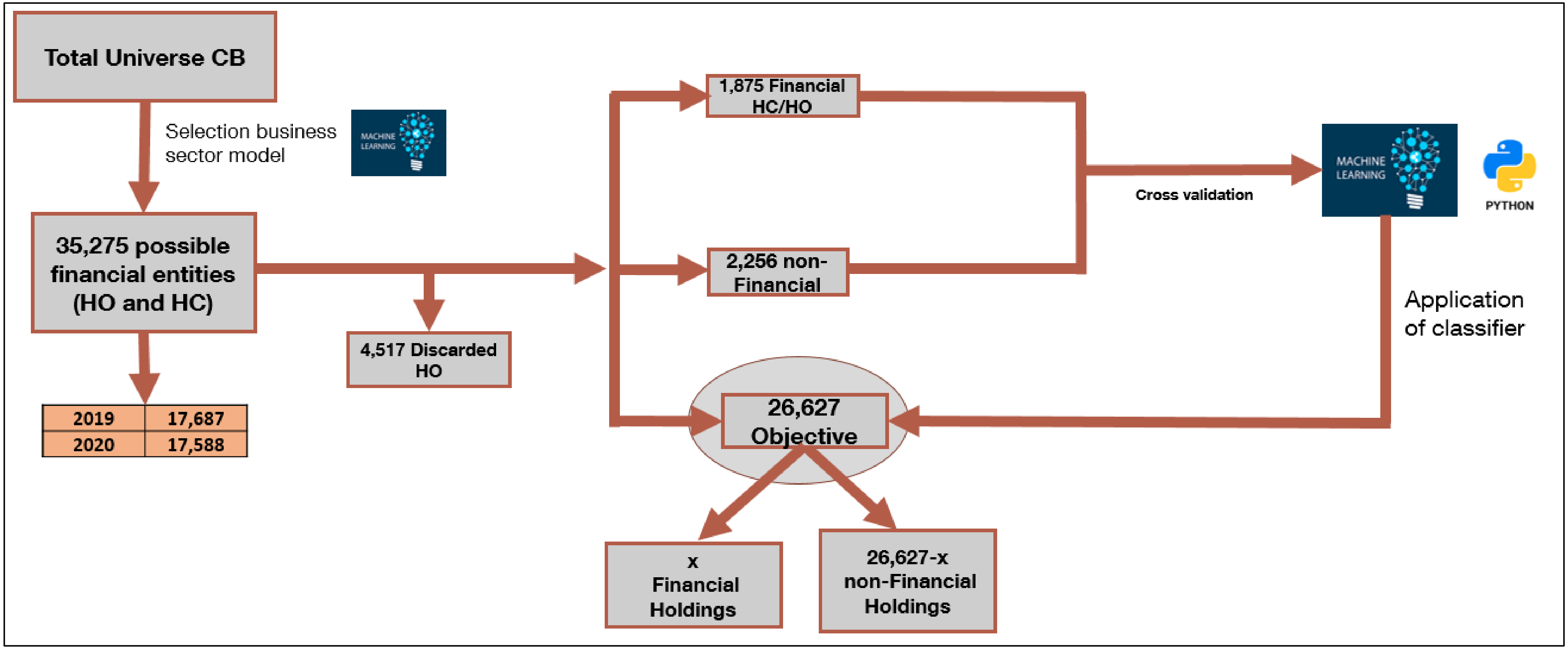
Table 9: classification of entities in the institutional sector project

Similarly to the Business sector project, an 85% training sample and a 15% test sample have been used. Also, cross – validation is utilized.

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 labeled-target subset (labeled 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 Central Headquarters by a large margin.



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

3.5 Final model

The model achieves 83% of accuracy in the test sample and the main metrics can be seen in the following table:

Sample	*Accuracy*	$F_1 score$
Training	87%	88%
Test	83%	85%

Table 10: performance of the model

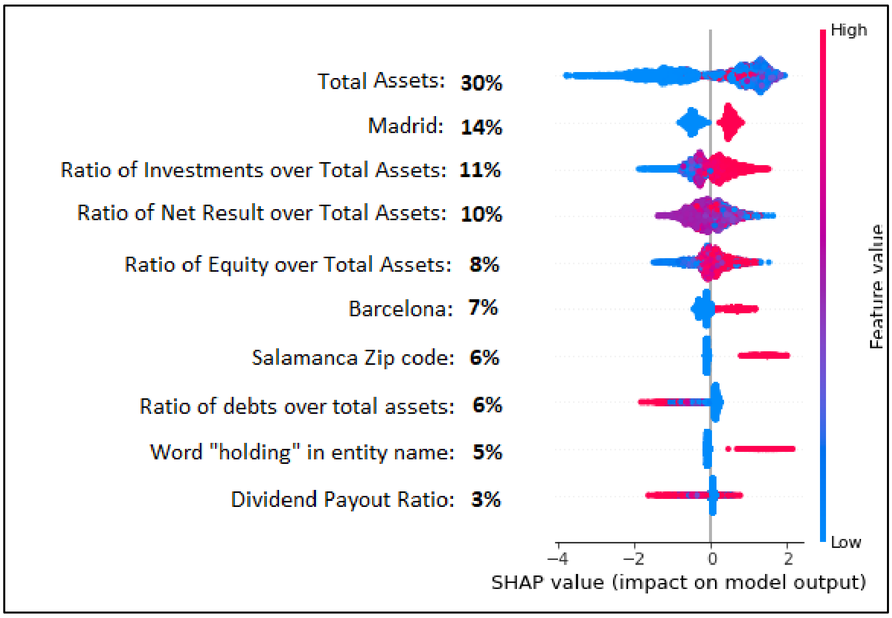
The final model trained on a grid described in Section 5.4 led to a model having the 10 variables in the following table:

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 Equity of the company in the current year divided by total assets.	Calculated ratio
Equity to Total Assets Ratio	Ratio of Equity of the company in the current year divided by total assets.	Calculated ratio
Barcelona Associated 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
Salamanca District	ZIP code associated with Salamanca district (28001) 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

Table 11: variables selected for the final model and their description

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 figure 4.



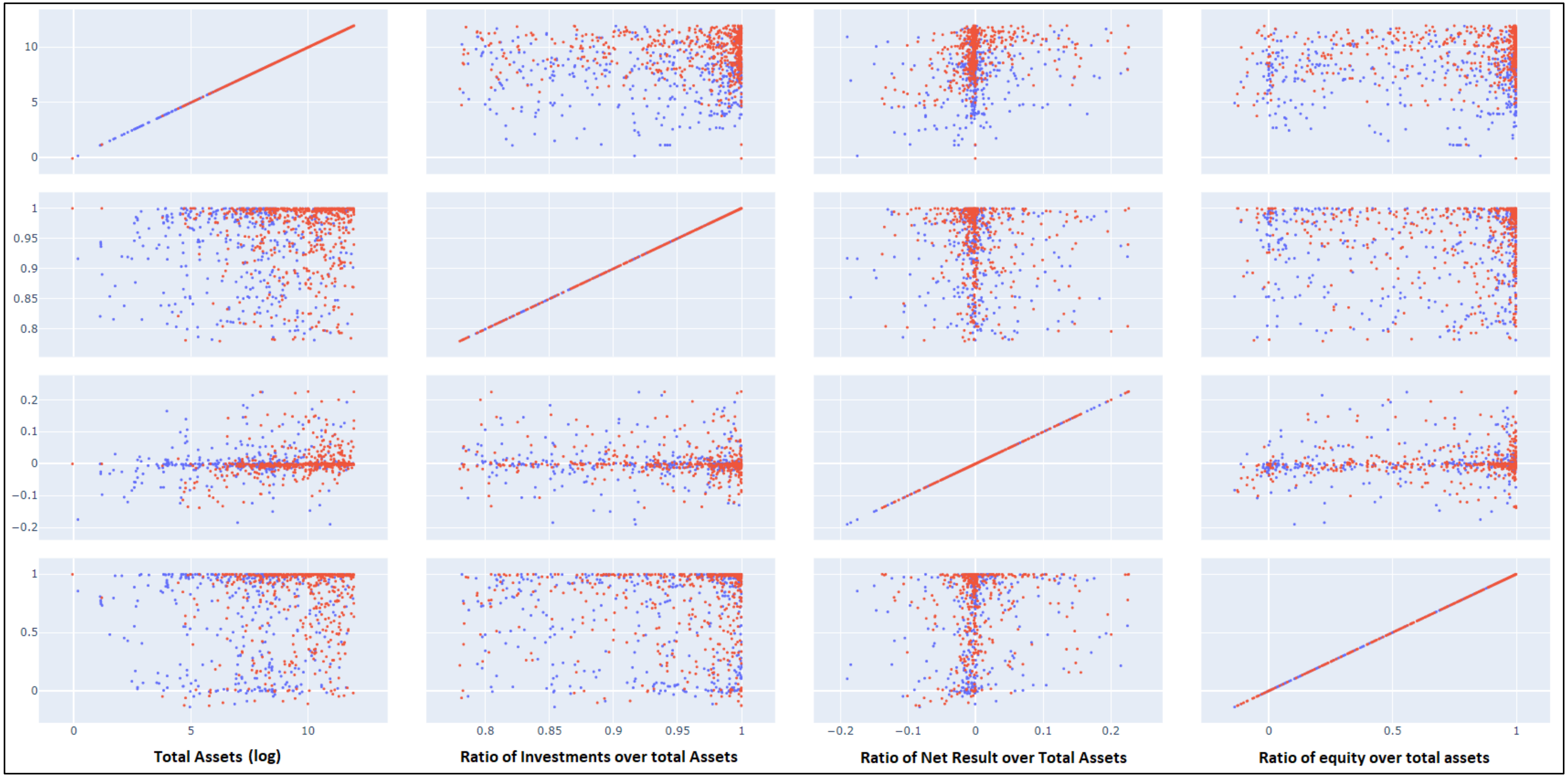
4 Variables of the institutional sector model

The interpretations of the values, the percentage of impact on the model, and additional details of the most interpretable variables follows next:

- **Total Assets (30%)**: This is the only non-binary variable in the model in absolute value (not a ratio), and it represents the Entity's Size. It's also the denominator for most of the ratios. In a very general sense, low values of this variable indicate a positive impact on the allocation of Non-Financial Holding.
    - Additional Detail: Its average value is lower in Non-Financial Holdings (€13 million) compared to Financial Holdings (€20 million).
- **Madrid Associated Postal Code (14%)**: High values of this variable (i.e., residence in Madrid) indicate a positive impact on the allocation of Financial Holding.
    - Additional Detail: 67% of Financial Holdings are located in Madrid, contrasting with 32% of Non-Financial Holdings with Madrid Residence.
- **Investment to Total Assets Ratio (11%)**: Low values of this variable indicate a positive impact on the allocation of Non-Financial Holding.
    - Additional Detail: Its average value is higher in Financial Holdings (~92%) compared to Non-Financial Holdings (~84%).
- **ROA, Net Income to Total Assets Ratio (10%)**: Low values of this variable indicate a positive impact on the allocation of Non-Financial Holding.
    - Additional detail: Its average value is higher in Financial Holdings (~8%) compared to Non-Financial Holdings (~0%).

- **Equity to Total Assets Ratio (8%)**: High values of this variable indicate a positive impact on the allocation of Financial Holding.
  - Additional Detail: Its average value is higher in Financial Holdings (~68%) compared to Non-Financial Holdings (~46%).
- **Barcelona Associated Postal Code (7%)**: High values of this variable (i.e., residence in Barcelona) indicate a positive impact on the allocation of Financial Holding.
  - Additional Detail: Among Holdings not residing in Madrid or Barcelona, only 27% of them are Non-Financial.
- **Postal Code 28001 - Barrio de Salamanca (6%)**: High values of this variable (i.e., residence in Barrio de Salamanca, Madrid) indicate a positive impact on the allocation of Financial Holding.
  - Additional detail: 16% of all financial holdings are located in this district.
- **Debt (Long and Short-term) to Total Assets Ratio (6%)**: High values of this variable indicate a positive impact on the allocation of Non-Financial Holding.
  - Additional detail: Its average value is higher in Non-Financial Holdings (~6%) compared to Financial Holdings (~2%).
- **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 allocation of Financial Holding.
  - Additional detail: Higher presence in Financial Holdings (~11%) compared to Non-Financial Holdings (~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 allocation of Non-Financial Holding
  - Additional detail: In large entities, the dividend ratio is 14.8% for Financial Holdings, compared to 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 figure 5.



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

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 in both years, 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 90% of probability required is a quite restrictive condition.

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 of them have enough information in our sources (questionnaires, documents and other internal and external sources) to determine the financial sector, explained in section 3.1.

Model/Business	Financial Holding	Non-Financial Holding
Financial Holding	11	12
Non-Financial Holding	0	8

Table 12: business revision of the institutional model

The 8 companies are businesses that are not located in Madrid and have low assets (generally less than €100,000). In contrast, the 11+12 companies that the model identifies as financial are mostly located in Madrid and have slightly higher assets. Most of them also meet a ratio of investments in group companies to assets and equity to assets very close to 100%. It's 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 learned

4.1 Conclusions

A machine learning model has been created to automatically detect Holdings 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 sectorization AI model.

The institutional sectorization ML model serves as a powerful tool in the institutional sectorization to validate, select and filter financial holdings.

4.2 Lessons learned

To achieve a good performance and interpretable model, it has been indispensable to use a high quality sample to train. In this case, a set of entities reviewed by business staff has been essential in other the model learns correctly.

It is important to give certain value to the NACE 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 Lessons learned

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

Additionally, this work is covered under a project of sectorization with machine learning within the Statistics Department within Banco de España. The next project to be covered involves early sectorization 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, utilizing NLP, semantic embeddings, and/or Large Language Models.

5 Annex: technical details of the models

The objective of this chapter is 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 that have been performed 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 due to 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.

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 utilizes 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 ultimately selected model 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 below:

Model	***Accuracy***	***Precision***	***Recall***	***F1-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%
Regresión Logística	97.9%	84.8%	99.5%	91.6%	97.3%
Árboles de Decisión	99.7%	100%	97.7%	98.8%	98.6%

Table 13: comparison of models in early stages of the Business sector project

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 / Headquarters sample compared to the rest of the companies in the CBA and CBH population combined. The original ratio is 1,482 Holding / Headquarters versus 10,993 non-Holding / 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 / Headquarters 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 analyzed. These companies shared the following characteristic: the model suggested a CNAE code of Holding/Central Headquarters, while the CNAE 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 listed below:

- Large entities: the model was mostly trained on small and medium-size 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 learned that companies have a certain bias in declaring their CNAEs. 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 changes. Therefore, revisions should be done on both revised and unrevised samples.

Other lessons learned during the review were:

- Including absolute variables instead of relative (ratios) variables can be helpful.
- There is a small number of Holdings with slightly more than 5 employees and headquarters with slightly less than 5 employees. That's why subsequent models became multi-class (Holding, Central Headquarters, Others).

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 over total assets < 35%. Then, low probability of Holding/Central Headquarters.

-If Equity instrument ratio over total assets (if the regular questionnaire for the entity is available) < 35%. Then, low probability of Holding/Central Headquarters.

-If Employment greater than 5 employees. Then, low probability of Holding.

-If Employment less than 5 employees. Then, low probability of Head Office.

-If Employment greater than 5 employees. Then, low probability of Holding.

-If Employment greater than 150 employees - Discarded as Holding. Then, low probability of Central Headquarters.

-If Turnover - Holding-related income > 50,000,000. Then, low probability of Holding/Central Headquarters.

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 utilized, specifically the SHAP library (Lundberg & Lee, 2017).

Such analyses aid 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, September). Feature-engine: A Python package for feature engineering for machine learning. *The Journal of Open Source Software*.

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

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

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*, 2825--2830.

Pegoraro, N., Benevolo, F., Gottron, T., Febbo, I., & 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., & Schürg, C. (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.
$\mathrm{F_1}$score	$2\frac{precision \cdot recall}{precision + recall}$
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	$\frac{TP}{TP + FP}$
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	$\frac{TP}{TP + FN}$
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.