BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Institutional sector classifier, a machine learning approach

by Oliver Giudice, Paolo Massaro and Ilaria Vannini

March 2020

Number 548

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Institutional sector classifier, a machine learning approach

by Oliver Giudice, Paolo Massaro and Ilaria Vannini

Number 548 – March 2020

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it.

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

INSTITUTIONAL SECTOR CLASSIFIER, A MACHINE LEARNING APPROACH

by Oliver Giudice[†], Paolo Massaro[*] and Ilaria Vannini[*]

Abstract

We implement machine learning techniques to obtain an automatic classification by sector of economic activity of the Italian companies recorded in the Bank of Italy Entities Register. To this end, first we extract a sample of correctly classified corporations from the universe of Italian companies. Second, we select a set of features that are related to the sector of economic activity code and use these to implement supervised approaches to infer output predictions. We choose a multi-step approach based on the hierarchical structure of the sector classification. Because of the imbalance in the target classes, at each step, we first apply two resampling procedures – random oversampling and the Synthetic Minority Over-sampling Technique – to get a more balanced training set. Then, we fit Gradient Boosting and Support Vector Machine models. Overall, the performance of our multi-step classifier yields very reliable predictions of the sector code. This approach can be employed to make the whole classification process more efficient by reducing the area of manual intervention.

JEL Classification: C18, C81, G21.
Keywords: machine learning, entities register, classification by institutional sector.
DOI: 10.32057/0.QEF.2020.548

Contents

[†] Bank of Italy, Information Technology Directorate.

[*] Bank of Italy, Statistical Data Collection and Processing Directorate.

1. Introduction

This paper explores the possibility of implementing machine learning techniques to obtain an automatic classification by institutional sector of the Italian companies recorded in the Bank of Italy's register of legal entities (*Anagrafe dei Soggetti* – '*Anagrafe'*). Individuals, such as natural persons, public institutions, companies and joint debtors[1] falling under the scope of the institutional tasks of the Bank of Italy, are recorded in this register. It contains a wide range of attributes (name, geographical location, type of economic activity, legal form, institutional sector, etc.) of over 40,000,000 entities and is obtained by harvesting information both from various administrative sources and from financial intermediaries.[2] This database is crucial for the production of aggregate statistics.

All entities included in the *Anagrafe* are uniquely identified by a code assigned by the Bank of Italy, that remains unchanged over time and is shared with other databases supporting all data collections related to surveys on households, firms and persons carried out by the Bank. The code is also linked to other national identifiers to allow the data recorded in the *Anagrafe* to be integrated with data collected from other sources.[3] Lastly, AS data are shared with financial intermediaries that are responsible for their timely updating.

An important piece of information recorded for each legal entity in the *Anagrafe* is its institutional sector, which is typically used to qualify the main activity and the economic function of a company in the market economy. The Bank of Italy classification of the Sector of Economic Activity (SEA) is described in Circular No. 140, 11 February 1991.[4] It should be noted that this classification refers to the sector (not to the economic activity itself) and largely reflects the European System of National and Regional Accounts (ESA 2010),[5] although it is much more detailed.

In the *Anagrafe* different official sources are used to assign the SEA code to each entity depending on its type, in particular:

- the National Institute of Statistics (Istat), for the public sector;
- the Bank of Italy, for the financial entities under its supervision;
- the Insurance Supervisory Authority (IVASS), for insurance companies.

---

[1] Entities that share liabilities on a particular debt.
[2] Financial intermediaries are required to provide the Bank of Italy with information about their customers.
[3] For example, tax codes and national business register codes.
[4] Regulation No. 140, 11 February 1991. https://www.bancaditalia.it/statistiche/raccolta-dati/segnalazioni/normativa-segnalazioni/Circ_140_4_agg.pdf
[5] European System of Accounts – ESA 2010, Luxembourg, 2013. https://ec.europa.eu/eurostat/documents/3859598/5925693/KS-02-13-269-EN.PDF/44cd9d01-bc64-40e5-bd40-d17df0c69334.

For the remaining legal entities (for example natural persons, individual companies, non-financial companies and non-supervised financial entities), the SEA code is provided by financial intermediaries, which are therefore in charge with updating it over time.[6] According to the diagnostic checks carried out by the Bank of Italy, the SEA codes provided by financial intermediaries have a certain degree of inaccuracy; moreover, for a few entities, no code is reported at all and as a consequence it has to be estimated.

Against this background, this paper assesses whether machine learning techniques can help to improve the SEA classification of companies and are able to impute missing values. In the empirical literature there are only a few examples of the adoption of these techniques for a similar purpose: IMF (2014) applies a decision tree based on deterministic rules for the classification of public-sector entities; Noyvirt (2019) compares results from a variety of machine learning methods aimed at classifying financial companies into sub-sectors. More in general, within central banks, the use of such algorithms for purely statistical tasks such as the classification of the institutional sector is a relatively new area of investigation.

The implementation of an automatic SEA classifier SEA has three main goals. First, it would reduce the intermediaries' reporting burden. Second, it would improve the overall quality of the *Anagrafe*. Third, it would improve the Bank of Italy's data quality management activities. Overall, all three reasons support the importance of investing in an innovative statistical approach allowing the most likely SEA to be associated with a company on the basis of its specific features.

The paper is organised as follows. Section 2 describes the dataset. Section 3 illustrates the machine learing solution adopted, from the pre-processing stage to the model selection and evaluation phases. Section 4 concludes.

2. The dataset

The classification of resident entities by SEA Sector of Economic Activity (SEA) follows the Bank of Italy's Circular No. 140/97 (only in Italian), which goes into more detail than the European System of Accounts (ESA 2010) classification. It distinguishes between general government, financial corporations, non-financial corporations, households and non-profit institutions serving households. At the maximum level of detail, the Circular classifies institutional units into 116 SEA groups.

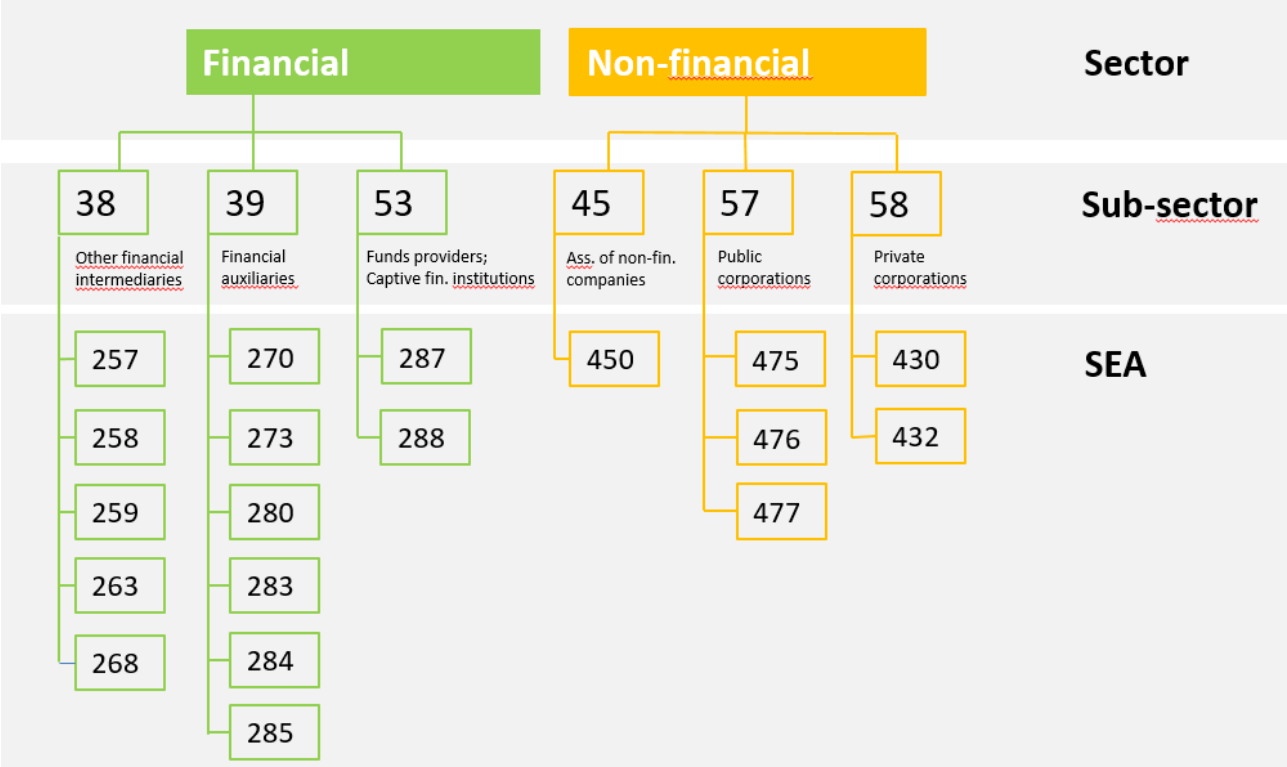
The focus of the paper is the assignment of SEA codes to resident corporations. Hence, our analysis disregards: (i) entities belonging to the public sector, households, non-profit institutions

[6] Financial intermediaries granting loans are requested to report the reference data of their customers.

serving households and institutions resident in countries different from Italy, (ii) quasi-corporations, (iii) supervised institutions, and (iv) all the institutions for which it is possible to determine the SEA on the basis of their being on a specific register or list (for example insurance companies and pension funds).[7]

The hierarchical classification of resident corporations by institutional sector includes 2 main sectors (financial and non-financial), 6 subsectors and 19 SEAs (Figure 1).

Figure 1. Structure of the classification by institutional sector of corporations, according to Bank of Italy Circular No. 140/97



On February 2018, the number of Italian corporations included in the population of interest had reached 1,756,006 entities.

The implementation of machine-learning algorithms to obtain an automatic classification by SEA of the Italian companies recorded in the *Anagrafe* requires a relatively large sample of correctly labelled companies to be available first. To this end, we used the Cerved database.[8] The sample for

[7] They are listed in specific registers maintained by the Institute for Insurance Supervision (IVASS) and the Board of Supervisors on pension funds (COVIP), respectively.
[8] A provider that collects information about Italian companies' annual accounts and their related reference data, including sector of economic activity (SEA), https://www.cerved.com/it

our empirical analysis contains companies that belong to both the *Anagrafe*'s and Cerved's databases[9] and that are identified by the same SEA.

Table 1. Sample distribution of Italian corporations by institutional sector

SEA code	Description	Sample		Population	
		Frequencies	%	Frequencies	%
257	Merchant banks	26	0.002	54	0.003
258	Leasing companies	188	0.013	653	0.037
259	Factoring companies	63	0.004	208	0.012
263	Consumer credit companies	60	0.004	620	0.035
268	Other financial corporations	1,252	0.089	7,870	0.448
278	Associations between financial and insurance undertakings	0	0.000	9	0.001
280	Insurance brokers, agents and consultants	2,710	0.192	4,246	0.242
283	Investment bankers	24	0.002	75	0.004
284	Other financial auxiliaries	1,200	0.085	2,434	0.139
285	Head offices of financial corporations	29	0.002	602	0.034
287	Financial holding companies whose principal activity is owning the group	64	0.005	1,869	0.106
288	Non-financial holding companies whose principal activity is owning the group	6,795	0.481	9,394	0.535
289	Captive financial institutions different from the holding company	0	0.000	13	0.001
430	Non-financial corporations	1,395,961	98.779	1,711,958	97.495
432	Head office of non-financial corporations	1,402	0.099	3,941	0.224
450	Associations of non-financial companies	0	0.000	5,180	0.295
475	Government-owned companies	129	0.009	775	0.044
476	Companies owned by local public authorities	3,266	0.231	5,412	0.308
477	Companies owned by other general government bodies	51	0.004	625	0.036
Total		**1,413,220**	**100**	**1,755,938**	**100**

In the *Anagrafe*, a limited number of companies belonging to the population of interest is classified according to three SEA codes - 278 ("Associations between financial and insurance undertakings"), 289 ("Captive financial institutions different from the holding company"), and 450 ("Associations of non-financial companies") – these companies are not represented in the selected sample since their SEA codes are not available in the Cerved database. (Table 1). Overall, 1,413,220 entities, belonging to 16 institutional sectors, are considered in the sample.

[9] The overlapping between the *Anagrafe* and Cerved gives a result of 1,625,636 corporations. This number differs from the total amount of our population of interest because of time lags in the registration of new corporations in the Cerved database, according to the deadlines for the presentation of balance sheets.

It is important to note that 98.7 per cent of entities are classified as "Non-financial" (i.e. SEA code 430). This is unsurprising due to the large number of small and medium-sized enterprises active in the manufacturing and agricultural sectors in Italy. Not counting these entities, the most frequently occurring SEA codes are:

- 288, "Non-financial holding companies whose main activity is owning the group";
- 476, "Companies owned by local public authorities";
- 280, "Insurance brokers, agents and consultants";
- 432, "Head office of non-financial corporations";
- 268, "Other financial corporations";
- 284, "Other financial auxiliaries".

The remaining sectors account for less than 0.05 per cent of the sample (Table 1).

A set of characteristics of the companies included in our sample that could be related to the institutional sector are extracted from different sources, namely:

- The Bank of Italy's *Anagrafe*: name of the corporation and its legal form;
- Cerved: balance sheet data;
- Agenzia delle Entrate,[10] (Italy's Revenue Agency): economic activity by ATECO code;
- Infocamere,[11] the National Business Register: number of employees, status of activity and notes to the financial statement;
- Ministry of Economy and Finance:[12] public (State or municipal) share in the equity and related costs.

The various sources of information are merged into one database and are used to assign the appropriate institutional sector to the companies that are not included in our sample.

The next two sections describe the distribution of all the attributes available in the database, distinguishing between "structured" data (regarding legal form, activity status, number of workers, ATECO and balance sheets, see Section 2.1) and "unstructured" data (text data derived from the name of the company and the notes to the financial statement; see Section 2.2).

[10] https://www.agenziaentrate.gov.it
[11] https://www.registroimprese.it/infocamere
[12] http://www.mef.gov.it

2.1 Structured data

Legal form and status of activity - According to the distribution of corporations in the *Anagrafe*, 90 per cent of the sample refers to "Limited liability companies", 6 per cent to "Cooperatives" and about 3 per cent to "Joint-stock companies" (Table 2).

Table 2. Distribution of Italian corporations by legal form

Legal form	Sample		Population	
	Frequencies	%	Frequencies	%
Other companies in the Italian Business Register	6,213	0.440	13,022	0.742
Cooperatives	89,190	6.311	111,469	6.348
Companies limited by shares	71	0.005	157	0.009
Foreign companies	178	0.013	2,798	0.159
Joint-stock companies	38,307	2.711	58,732	3.345
Limited liability companies	1,279,170	90.515	1,569,760	89.397
Not available	91	0.006	0	0.000
Total	**1,413,220**	**100**	**1,755,938**	**100**

According to the Italian Business Register, the majority of corporations analysed (62 per cent) is active (Table 3); inactive companies (removed from the Italian BusinessRegister, under liquidation or legal proceedings) also belong to our population of interest and as such they are included in the sample.

Table 3. Sample distribution of Italian corporations by activity status

Status	Frequencies	%
Active	875,291	61.9
Inactive	537,929	38.1
Removed from the Italian Business Register	*349,566*	*24.7*
Under liquidation	*125,592*	*8.9*
Under legal proceedings	*62,771*	*4.4*
Total	**1,413,220**	**100.0**

Employment - Information about the number of workers is collected by Infocamere (a consortium of Italian chambers of commerce) (Table 4); it is worth specifying that for about 400,000 companies in our dataset the number of workers is either "not available" in our database (since this piece of information is not reported by the companies) or it is 0 ("none" in Table 4). This is probably due to the fact that Infocamere does not collect data on workers with some types of contracts, for example e.g. supply contracts and casual work contracts.

Table 4. Sample distribution by firms' staff numbers and activity status

Number of workers	Activity status				Total
	Active	Removed from the Italian Business Register	Under liquidation	Under legal proceedings	
None	14,810	11,762	3,266	593	30,431
1-2	269,474	108,056	51,729	19,794	449,053
3-5	147,720	44,593	20,414	12,267	224,994
6-10	103,281	25,349	10,630	9,736	148,996
11-20	71,722	16,062	5,686	7,249	100,719
21-50	39,619	9,824	2,972	4,746	57,161
51-100	11,956	2,910	805	1,249	16,920
101-250	6,671	1,629	325	585	9,210
251-500	1,852	414	73	130	2,469
> di 500	1,338	202	19	39	1,598
Not available	206,848	128,765	29,673	6,383	371,669
Total	**875,291**	**349,566**	**125,592**	**62,771**	**1,413,220**

With regard to the distribution by "type of worker", 885,913 companies have payroll employees, 196,475 self-employed workers, and 53,405 other types of workers. (Table 5).

Table 5. Sample distribution by type of worker (descriptive statistics)

Types of workers	Number of enterprises	Median	Total
Payroll employees	885,913	3	11,503,075
Self-employed workers	196,475	1	283,118
Other	53,405	1	142,850
Total	**1,011,120**	**3**	**11,929,043**

Economic activity – A company's economic activity is a useful input for identifying its institutional sector. This information is provided by the Italian Revenue Agency, which every three months sends the updated classification of Italian companies by economic activity (ATECO 2007)[13] to the Bank of Italy. Its breakdown ranges from one digit (21 general categories) to six digits (480 categories).

[13] The 2007 Classification by Economic Activity (ATECO) adopted by ISTAT is the national version of the European Nace Rev.2; see Regulation No 1893/2006 of the European Parliament and of the Council of 20 December 2006.

The ATECO code entered in the *Anagrafe* is taken from the declaration of the company to the Italian Revenue Agency; according to our experience, such information is not fully reliable since sometimes it may refer to only one of the many activities of the company.[14]

On the basis of the ATECO codes recorded in the *Anagrafe*, in the sample under consideration here, companies were distributed among the following sectors: 19 per cent are in "Wholesale and retail trade", 16 per cent in "Construction", 15 per cent in "Manufacturing", and 11 per cent in "Real estate" (Table 6).

Table 6. Sample distribution by economic activity (ATECO code, 1st digit)

Classification of economic activities	# companies	%
Manufacturing	217,084	15.4
Electricity, gas, steam and air conditioning supply	11,971	0.8
Water supply; sewerage, waste management and remediation activities	8,074	0.6
Construction	225,701	16.0
Wholesale and retail trade; repair of motor vehicles and motorcycles	278,927	19.7
Transportation and storage	47,849	3.4
Accommodation and food service activities	79,290	5.6
Information and communication	56,211	4.0
Financial and insurance activities	12,291	0.9
Real estate activities	155,572	11.0
Professional, scientific and technical activities	66,103	4.7
Administrative and support service activities	76,997	5.4
Public administration and defense; compulsory social security	302	0.0
Education	7,845	0.6
Human health and social work activities	21,086	1.5
Arts, entertainment and recreation	22,609	1.6
Other service activities	17,224	1.2
Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	18	0.0
Activities of extraterritorial organisations and bodies	94	0.0
Not available	107,972	7.6
Total	**1,413,220**	**100**

Institutional sector and balance sheet data - The classification by institutional sector classifies companies according to the degree and type of control by government entities. On the basis of the list of shareholdings of governmental agencies (central, local or other) made available for 2017 on the website of the Ministry of Economy and Finance,[15] in our sample agencies belonging to the public sector participate in 5,351 companies (directly, indirectly or both).

[14] The same company could be involved in different types of economic activities and declare only one of them to the Italian Revenue Agency.

[15] http://www.dt.mef.gov.it/it/attivita_istituzionali/partecipazioni_pubbliche/censimento_partecipazioni_pubbliche/ (only in Italian)

Looking at the balance sheet data, we selected those items that are likely to be closely related to the institutional activity of the company (Table 7) for example:

(1) "Total equity investments" and "Financial income from investments" are very significant for financial and non-financial "Holding companies whose principal activity is owning the group";

(2) "Manufacturing expenses" - specific costs related to productive activity such as raw materials, consumables used and services, "Payables to suppliers", and "Total receivables" are quite important for non-financial enterprises;

(3) "Total financial fixed assets", "Finance costs and revenues", and "Interest and other financial expenses" are quite significant in financial companies' balance sheets.

Table 7. List of items selected from the balance sheet

Balance sheet items
Total financial fixed assets
Total equity investments
Total receivables
Share capital
Payables to suppliers
Value of production
Manufacturing expenses
Raw materials and consumables used
Cost of services
Employee benefit expenses
Finance costs and revenues
Financial income from investments
Interest and other financial expenses
Balance sheet total

2.2 Unstructured data

Name of the company – Useful information about the institutional sector of a company can be derived from the name of the company itself. For example, if a company name contains the words "leasing" or "factoring", then it could be similarly classified with SEA codes "258 - Leasing companies" or "259 - Factoring companies".

Notes to the financial statements - The "Notes to the financial statements" contain additional information on the budget figures, explanations of the accounting methodologies used for recording and reporting transactions, details on pension plans and information on stock option compensations. These notes are very detailed for large enterprises; instead small and medium enterprises (SMEs) are

only required to publish a condensed version[16] and micro enterprises[17] are not obliged to publish them at all.

The introduction to the notes to the financial statements can also contain some useful information to identify the SEA. Overall, we use this information for around 600,000 companies and analyse over 100 million words (corresponding to about 200 words for each note).

3. A machine learning (ML) approach to predict the SEA

In this section we propose an ML approach in order to assign the SEA code to a company on the basis of the vast amount of data described in the previous Sections. A supervised ML method is used in order to learn how to use the best classification model through the following four steps:

1. data pre-processing;
2. feature encoding;
3. model selection and training;
4. model evaluation.

The method must be able to deal with the following three main characteristics of our database:

1. data is of a mixed type, that is both structured and unstructured where: (i) structured data is both categorical and numerical; and (ii) unstructured (textual) data have different lengths and semantics;
2. the sample under consideration is imbalanced across the response class;
3. the potential presence of outliers.

The task is inherently difficult, since the part of the input data that contains the relevant information is unknown beforehand. Moreover, part of the information set is encoded in natural language, which is inherently ambiguous and might contain typos or include domain-specific initialisms, idioms and jargon. Finally, the main language in the input data is Italian, which is a language that is currently less investigated than English in Natural Language Processing.

As mentioned above, an additional problem is related to the imbalanced sample distribution for each SEA in the dataset, since most companies (about 98 per cent) belong to "Manufacturing" and consequently receive the SEA code 430. It is important to deal with this issue since, as pointed out by Longadge: "most of the classifiers are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major

[16] Companies with: (1) balance sheets with an asset size of less than €4,400,000, (2) income from sales of less than €8,800,00 euros, and (3) less than 50 employees can publish a condensed set of financial statements and the related notes.
[17] A micro enterprise is defined as a small business employing up to nine people and having a balance sheet or turnover of less than €2,000,000.

class and ignores the minor class" (Longadge *et al.*, 2013, p. 1). To overcome this problem we select two very efficient state-of-the-art Imbalanced Learning methods (He, Garcia, 2008), as explained in Section 3.2.

3.1 Data pre-processing

The classification takes place in a context that is made complex due to the presence of data with different features. For a given company, the available information can be summarised as follows:

1. numerical:
    a. balance sheet data (x_B, 14-dimensional) as listed in Table 7;
    b. the organisation and structure of the company (x_C, 15-dimensional)[18];
2. categorical:
    a. legal form (x_{lf}, 1-dimensional);
    b. ATECO (x_A, 3-dimensional)
3. textual:
    a. name of the company (x_N);
    b. notes to the financial statements (x_n);

As described in Section 2, not all features are present for all the companies in our sample. The absence of a specific feature is treated in two different ways: as regards x_n, its presence or absence leads to a completely different classification model, as described in Section 3.3, whereas the absence of any other feature is encoded with a specific value (-inf).

Both numerical and categorical features are very different in terms of their magnitude and statistical distribution. This is not taken into account at this stage; below, we deal with this issue by resorting to a robust classification model. Textual features need specific pre-processing in order to be evaluated by any classification model, as explained in the following paragraph.

3.1.1 Textual Pre-Processing

Textual features x_N and x_n must be represented in a numerical form in order to be exploited through any ML technique, but from a semantic point of view they are very different: x_n is a long well-written and well-organised text (hundreds of words); x_N contains few words that are, typically,

[18] In detail: number of employees, number of independent collaborators, number of other type of workers, share of direct government equity participation in the company (central, local and other governmental agencies), share of indirect government participation equity in the company (central, local and other governmental agencies), costs due to the equity investments (distinguishing central, local and other governmental agencies), and further divided into costs borne and costs incurred.

not put into sentence form and sometimes refer to acronyms or words that cannot be found in the dictionary.

We start by dealing with x_n. It can be pre-treated by natural language processing. Standard normalisation operations (Jurafsky and James, 2009) are applied to x_n in order to transform the original text as follows:

1. lower-casing;
2. removing punctuation;
3. removing stop words;[19]
4. tokenisation (for numbers, references to laws, amounts of money, etc.);[20]
5. removing words not present in a mixed Italian/English dictionary;[21]
6. lemmatising;
7. stemming;
8. representing everything in a list of words.

Turning to x_N, a simplified approach is taken due to its specificity: only a lower-casing operation is employed in order to prevent the risk of losing too much information.

The textual data are represented in the form of a "list of words"; then, it is necessary to encode them into a numerical form in order to feed an ML model. We adopt the well-known *Bag of Words* technique (Jurafsky and James, 2009). In particular, the *Bag of n-grams*[22] representation is implemented, with 1-2-grams and TF/IDF normalisation[23] in the case of x_n and with 1-2-3-grams with TF/IDF for x_N. The vectors of features thus obtained have dimensions of $6\text{x}10^5$ and $1{,}7\text{x}10^4$, respectively.

These vectors are then evaluated through a Principal Component Analysis (PCA; Word *et al.* 1987) in order to reduce dimensionality. Through this kind of analysis, the number of principal

[19] Commonly used words not corresponding to any particular subject matter.
[20] Reserved words, not present in any dictionary that are commonly known as "tokens", are used to replace textual information that is not in any dictionary but must be normalised. For instance, a reference to a law such as Dlgs. 194/96 is converted into [lex-194-96]: a token able to encode uniquely the referencing information. Numbers and amounts of money are replaced with tokens representing their magnitude such as €12.40 is replaced with [eur-10] and a generic number 1420 is replaced with [1000].
[21] This is due to the presence of English terms that have entered the Italian corpus.
[22] An n-gram is a contiguous sequence of n items (for example phonemes, syllables, letters, words) from a given sample of text.
[23] TF/IDF is a two-fold normalisation so that each document is normalised to length 1. This is equal to taking the relative frequencies instead of the absolute term counts.

components decreases up to 15 for x_N and 175 for x_n and so an explained variance[24] of 75 per cent for both such vectors is obtained.

3.2 Imbalanced Learning

In the literature on ML, "Imbalanced Learning" has received great attention in order to overcome the problems related to datasets with imbalanced response variables (Chen and Breiman, 2004; Japkowicz, 2000; Chawla *et al.*, 2002; Chawla *et al.*, 2003; Weiss and Provost, 2003).

In general, the use of sampling methods consists in modifying a set of imbalanced data through resampling mechanisms in order to provide a new balanced distribution. This is crucial for ML since many studies have already demonstrated both the weakness of the classification models developed in such a condition of imbalancing and the substantial improvement in performance obtained by means of sampling methods (He and Garcia, 2009; Weiss and Provost, 2003).

Resampling techniques can easily be applied to any ML method, since they act as a further processing phase to be done after feature encoding: this is why samples need to be well defined into the correct feature-space. Several resampling methods have been proposed, the most used are oversampling and undersampling.

Let S be a dataset with n observations ($|S| = n$), where x_i is an observation in the n-dimensional space, and y_i is the response class (label) associated with x_i and let $C = 2$ be the total number of classes. We can define:

- $S_{min} \subset S$, the subset of S containing all the samples for the minority class;
- $S_{max} \subset S$, the subset of S containing all the samples for the majority class;

so that $S_{min} \cap S_{max} = \{\emptyset\}$ and $S_{min} \cup S_{max} = S$.

The new samples generated by re-sampling procedures on S are indicated with R and can be divided into the disjoint subsets, R_{min} and R_{max}. In this framework, oversampling is a method that aims at balancing the distribution of classes through the random replication of minority class examples. On the one hand, the random oversampling mechanisms include the addition of a set R obtained from observations belonging to the minority class, randomly perturbed. In this way, the number of total observations in S_{min} is increased by $|R|$ and the distribution of the response class of S is balanced. On the other hand, random undersampling is a method that aims to balance the class

[24] If $\hat{y}$ is the estimated target output, y is the corresponding (correct) target output, and $Var(\cdot)$ is the well-known variance function, then the explained variance is estimated as $1 - \frac{Var\{y-\hat{y}\}}{Var\{y\}}$.

distribution through the random elimination of observations belonging to the majority class. In particular, a series of observations belonging to S_{max} are removed in a way such that $|S| = |S_{min}| + |S_{max}| - |R|$. As argued by Liu *et al.* (2009), this process not only yields balancing but also accelerates the learning process of the algorithm.

At first glance, the methods of random oversampling and random undersampling might appear functionally equivalent, since they both alter the volume of the original data and can actually provide the same proportion of balance. However, each method raises a few problems that can potentially hinder learning (He and Garcia, 2009). Several authors agree that random oversampling can increase the likelihood of overfitting, since copies of observations from the minority class are replicated. Furthermore, it can introduce an additional computational cost if the imbalanced dataset is large enough. The main drawback of random undersampling is that this method can discard potentially useful data, which could be important for the learning process (Kotsiantis *et al.* 2005).

Differently, with the "Synthetic Minority Over-sampling Technique-SMOTE) (Chawala *et al.* 2002), the majority class is re-sampled by removing samples on the basis of a specific optimisation metric trying to minimise the risk of removing useful data.

However, the SMOTE algorithm is not suitable for categorical predictors, as in our scenario. Thus, we resort to the algorithm known in the literature as SMOTE-NC, which relies on a variation of the metric distance by adopting the Value Distance Metric (VDM), using it to calculate the nearest neighbours for datasets having both a nominal and a continuous predictor (Chawala *et al*. 2003).

In sum, the imbalancing issue in our original dataset is tackled by applying a combination of under-sampling and over-sampling: the minority class is over-sampled by using the SMOTE-NC algorithm and the majority class is randomly undersampled.

3.3 Model selection and training the classifier

The novelty of the working context and the complex data scenario described in the previous sections imply a complex and multi-step iterative model selection. Several learning models have been tested with many combinations of parameters and architectural compositions. The investigated solutions can be summarised by five different classifiers that we call "Generations" (from G0 to G4). In order to better understand the classification process, let us formally define the SEA classification problems.

We start by defining an entity E in the *Anagrafe*, as a sample described by all of its features:

$$E \coloneqq x = \{x_B, x_C, x_{lf}, x_A, x_N, x_n\} \quad (1)$$

For each E we know the corresponding SEA label assigned correctly on the basis of a process involving the direct assessment of experts. In order to make the whole process automatic, given an entity E, we look for a classification function f_C defined as follows:

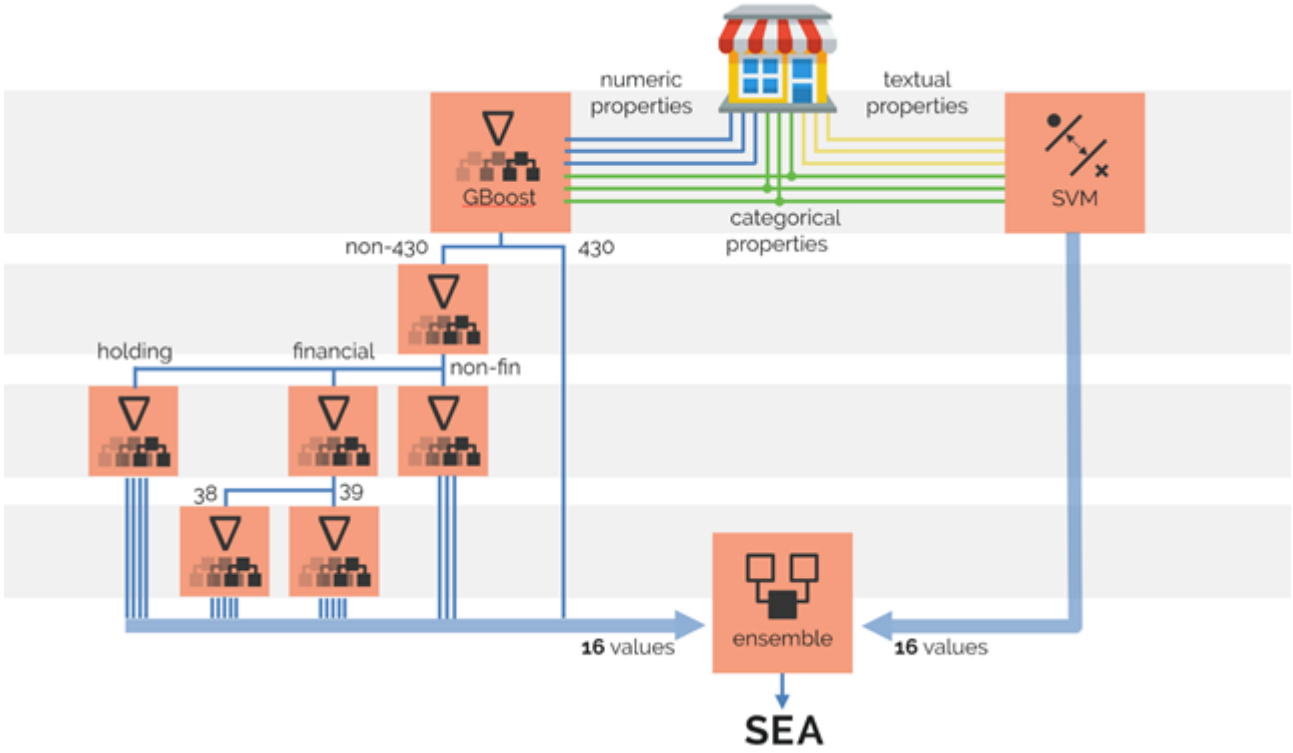
$$f_C : \mathbb{R}^{223} \rightarrow SEA \quad (2)$$

where SEA is a categorical response class composed by 16 SEA codes.

For non-textual features, due to their heterogeneity in terms of origin, significance, magnitude and dimensionality, a normalisation process would compromise the classification results (Salama *et al.*, 2010). This is why instead of finding the best normalisation parameters manually, we select the most robust classification model for this kind of issue, known as the Gradient Boost technique - GBoost (Friedman, 2002; Chen and He, 2015). Instead, for textual features we select Support Vector Machine (SVM) models (Cortes and Vladimir, 1995) with RBF kernel, given that $x_N, x_n \in [0,1]$.

Although SMOTE-NC and undersampling have been proved to be very effective, the degree of imbalancing in *Anagrafe* data is so high that it can only be overcome by adopting a hierarchical classification solution (Gordon *et al.*, 1987), such that at the first stage a classifier is applied to correctly discriminate between the most represented class and all the others. It should be noted that if 430 is predicted, an almost zero probability is evaluated for the entire hierarchy for non-430 classes. Figure 2 (left-hand side) shows how the classifiers are organised into a hierarchical model of sequential GBoosts. For textual features (Figure 2, right-hand side), the same considerations apply, but using SVM models instead.

Figure 2. The proposed hierarchical classification model (1)



(1) The left-hand side gives the hierarchy for the numeric and categorical features, the right-hand side shows the SVM models for the textual and categorical features.

To combine the results, a simple ensemble method (Tordoff and David, 2002) is exploited:

$$SEA = \text{argmax} \sum_{j=1}^{2} w_j \, P_j(x) \tag{3}$$

where $P_j(x)$ is a 16-dimensional vector representing the probability distribution for each of the 16 SEAs obtained given input x. The probability distribution is obtained as a result of all the predictions of the classifiers in the two hierarchies (j=1 for the left-hand side of Figure 2 and j=2 for the right-hand side); w_j is an arbitrarily assigned weight (obtained through the validation set). Thus, the predicted SEA is the argument that maximises the weighted sum of all SEA probability distributions obtained by each classifier in the two branches.

The hierarchical technique described is developed into two slightly different versions:

- G1: $f_{G1}(x)$ with $x = \{x_B, x_C, x_{lf}, x_A, x_N\}$;
- G2: $f_{G2}(x)$ with $x = \{x_B, x_C, x_{lf}, x_A, x_N, x_n\}$, thus including the notes to the balance sheets.

Hierarchical techniques help to overcome the imbalance in our data; however, they rely on *a priori* information on the distribution of such data that is not necessarily suitable to capture the actual hierarchical organisation of SEA codes. For this reason, we built a fully inductive technique where

the following eight peer classifiers are trained on a subset of SEA codes, grouped only on sample cardinality considerations and with as little *a priori* knowledge as possible:

- $C_{430}(x) \rightarrow p(\{0,1\})$ a probability distribution where 1 is SEA=430 and 0 denotes any other SEA;
- $C_{2-sec}(x) \rightarrow p(\{sec4, sec23\})$: a probability distribution where the labels can be 2 out of the three sectors in the SEA hierarchy;
- $C_{3-sec}(x) \rightarrow$ p({holding, sec4, sec23}): a probability distribution where the labels can be all of the three sectors in the SEA hierarchy;
- $C_{is-holding}(x) \rightarrow$ p({0,1}): a probability distribution where 1 denotes a holding company, 0 otherwise;
- $C_{holdings}(x) \rightarrow$ p({285,287,288,432}): a probability distribution where the labels can be one of the four SEAs belonging to the holding Sector;
- $C_{b4}(x) \rightarrow$ p({476,280,288,432}): a probability distribution where the labels can be one of the four most represented SEA labels (in terms of samples);
- $C_{b5}(x) \rightarrow$ p({476,280,288,432,430}): a probability distribution where the labels can be one of the five most represented SEA labels (in terms of samples);
- $C_{b7}(x) \rightarrow$ p({476,280,288,432,430,284,268}): a probability distribution where the labels can be one of the seven most represented SEA labels (in terms of samples).

Given the eight classifiers described above it is possible to define a new feature vector:

$$\xi(x) = \{C_i(x), x_A, x_{n6}\}, i = 1..8 \quad (4)$$

where C_i are the eight classifiers, x_A is the 3-dimensional ATECO representation and x_{n6} represents the 6-principal components of the notes to the financial statements.

This 36-dimensional feature vector is used as input on a neural network (Demuth *et al.*, 2014) in a very simple architecture composed of three layers:

- L1: Fully connected, 40 neurons, RELU, dropout 0.1;
- L2: Fully connected, 20 neurons, RELU, dropout 0.1;
- L3: Fully connected, 15 neurons, RELU, dropout 0.1;
- Output layer, 16 SEA codes, softmax.

Given the classification model described so far, the third and fourth approaches can be defined as:

- G3: $f_{G3}(\xi(x))$ with $\xi(x) = \{C_i(x), x_A, x_{n6}\}, i = 1..8 \; and \; x = \{x_B, x_C, x_{lf}, x_A, x_N, x_n\}$;
- $G4: \begin{cases} G1, if \; G1 \; predicts \text{ "430"} \\ G3, \qquad\qquad otherwise \end{cases}$.

The next Section reports and discusses the experimental results.

3.4 Model evaluation and experimental results

Experimental tests are conducted on the dataset with all the trained models from all the generations described in the previous sSection. The dataset is split into two parts: 60 per cent of the samples are used to build the training set to which the classifiers are fitted; for the remaining samples, 20 per cent are used to build a validation set to find the best hyper-parameters and model-architecture; and the residual 20 per cent is used as test set for evaluating of the effectiveness of the various techniques.

We exploit a standard ML evaluation metric known as “confusion matrix”, a specific table layout where each row represents the actual class while each column shows the predicted class. A confusion matrix C is such that the element $C_{i,j}$ is equal to the number of observations known to be in group *i* and predicted to be in group *j*. A normalised confusion matrix is such that $C_{i,j}$ is divided by the number of elements that are actually in the true class *i*. The following figures show the normalised confusion matrices for the models from G1 to G4.

Figure 3. Normalised confusion matrix for G1

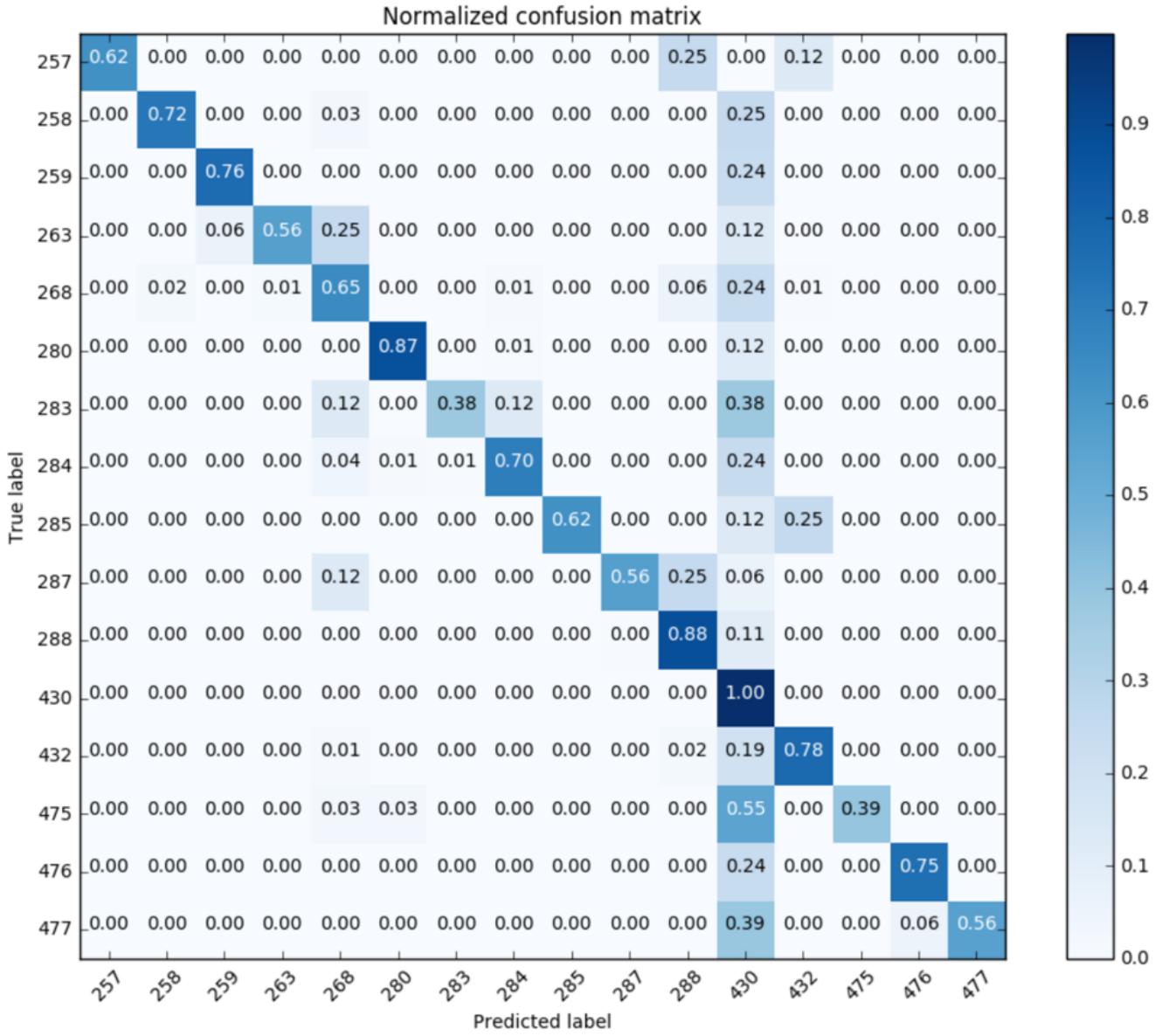


Figure 4. Normalised confusion matrix for G2

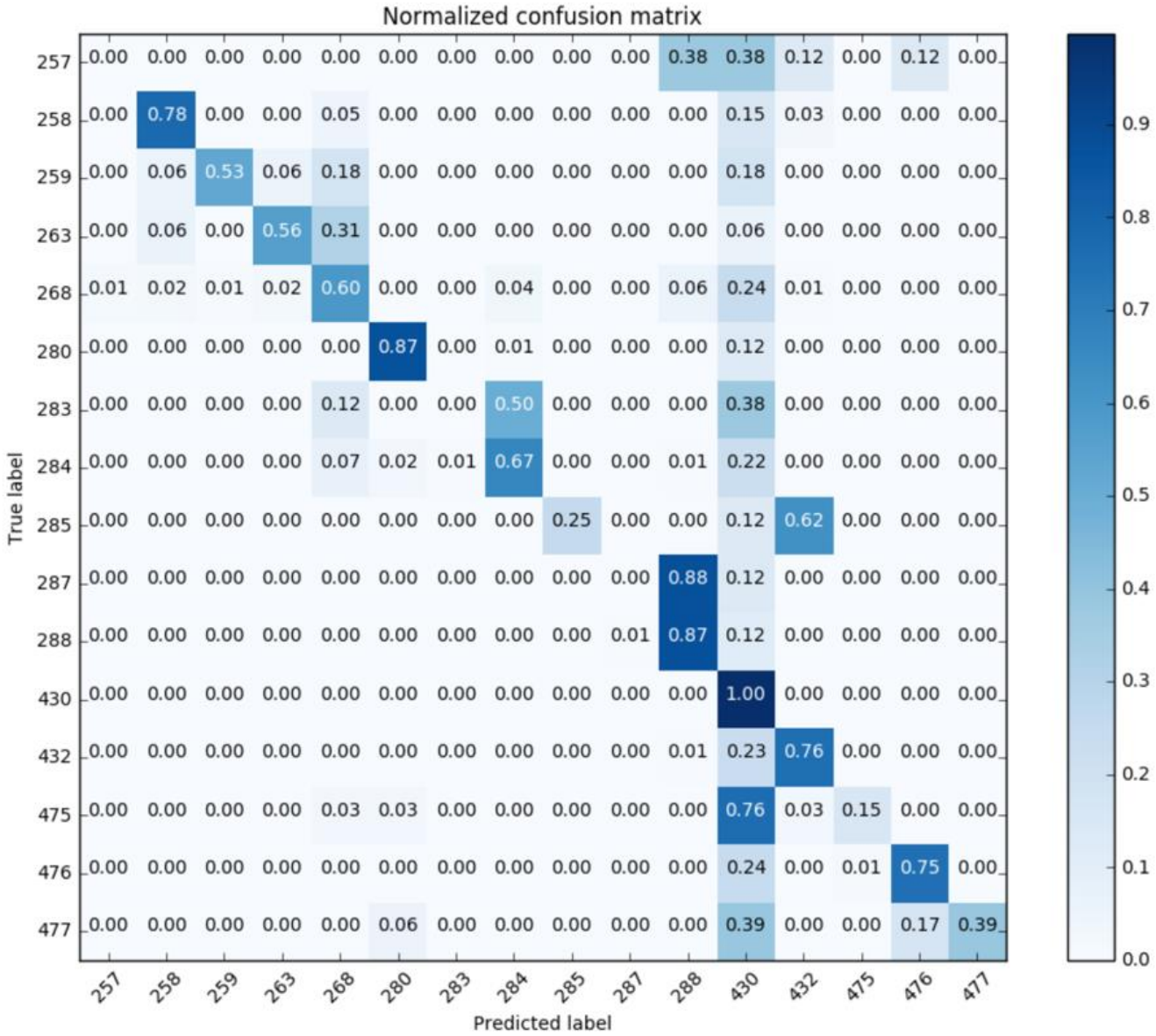


Figure 5. Normalised confusion matrix for G3

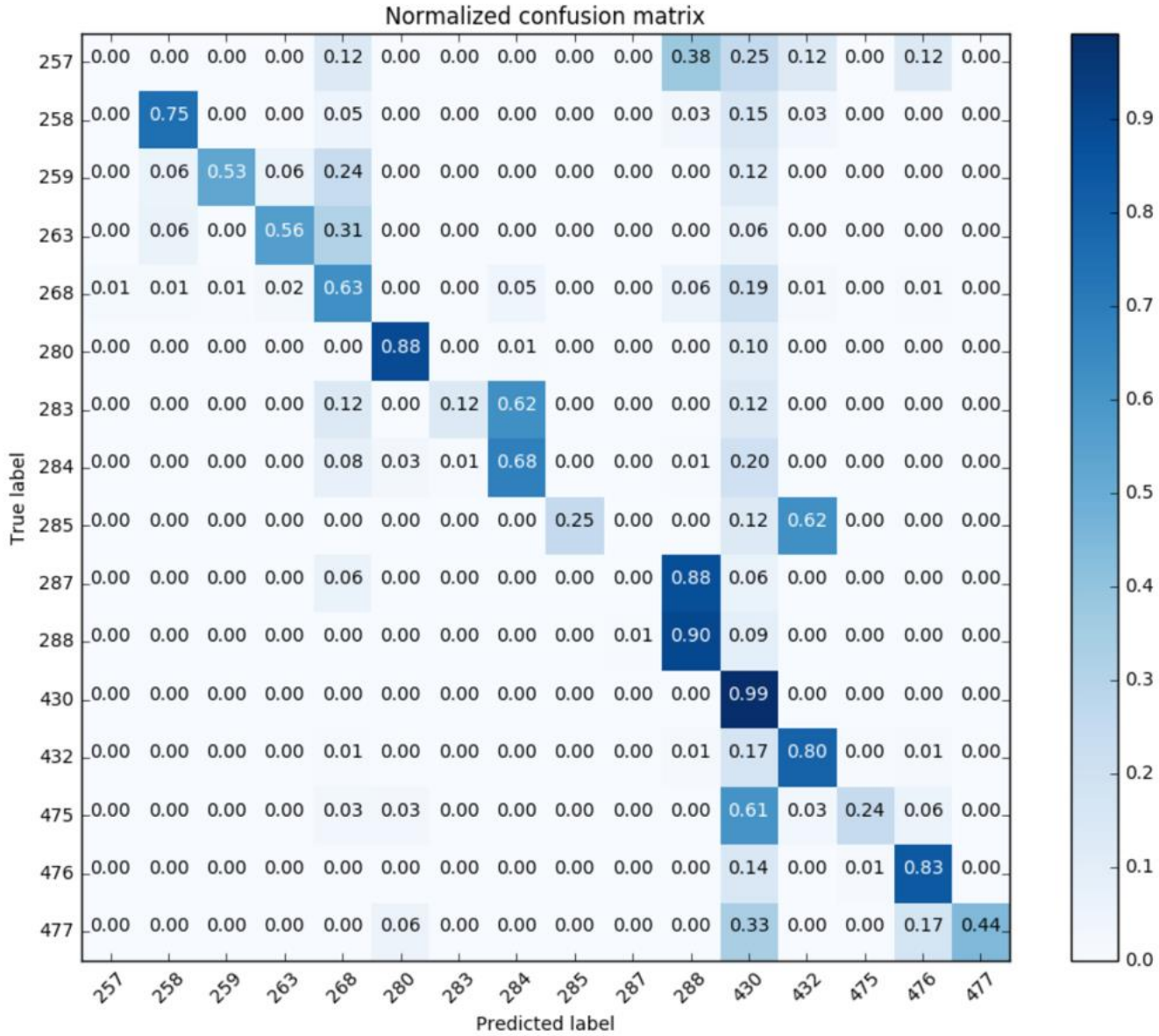
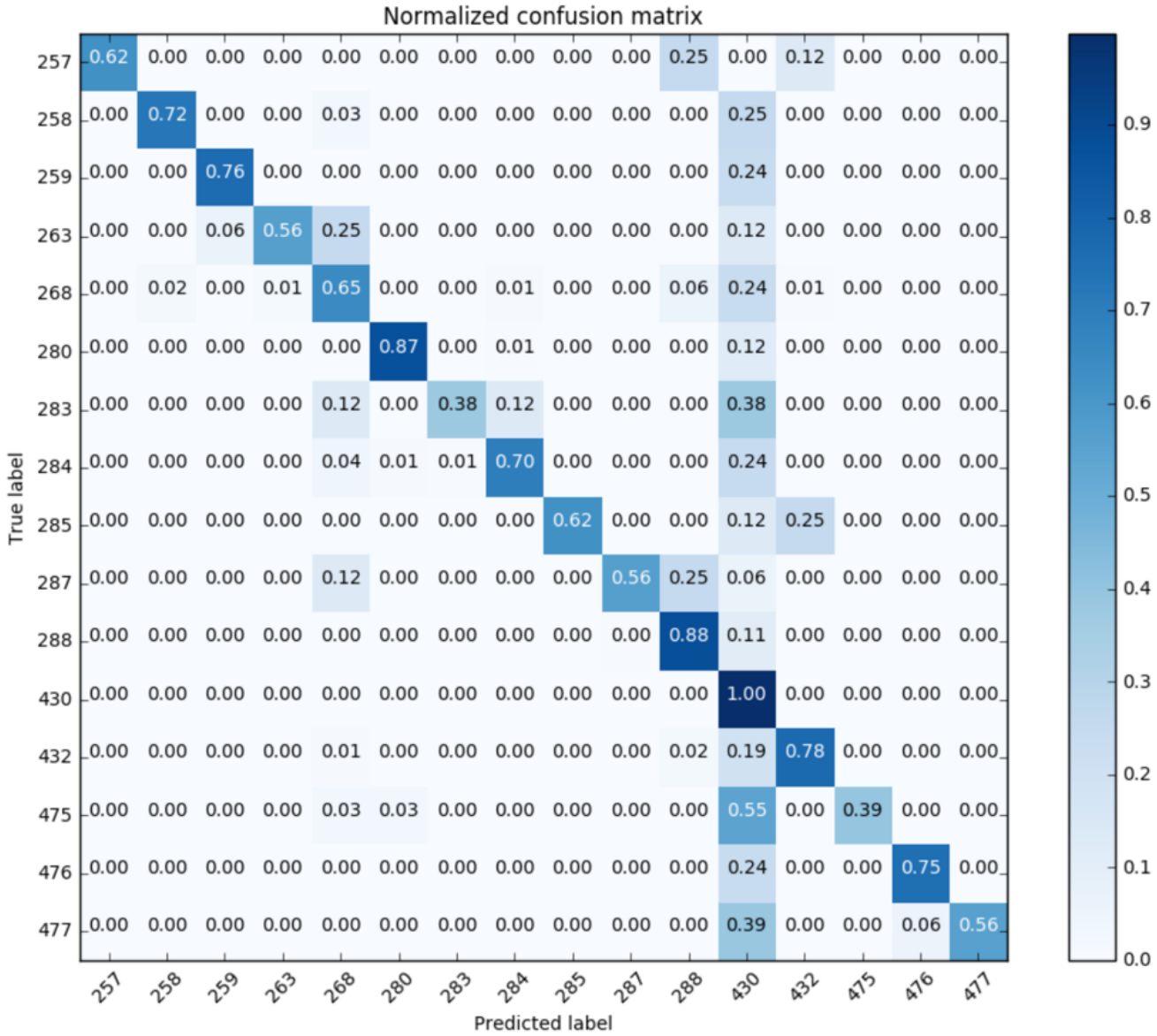


Figure 6. Normalised confusion matrix for G4



Visual inspection of Figures 3-6 shows that the best results on the test set are obtained by implementing G4[25] that extends the already well-performing G1 to situations in which further information, such as the notes to the financial statements, is available.

Although results can be visually explained by confusion matrices in a simple way, a numerical metric known as the "error rate" (defined as $e = 1 - \frac{H}{T}$ where H stands for all the correct classifications and T is the total number of samples in the test set) is used in order to select the best performing generation. However, this metric is strongly related to label cardinality and the presence of imbalanced classes produces the wrong perception of the values of the error rate. To better appreciate this, a simple classification model (G0) such that the SEA is always set equal to 430 is

[25] Rounding off to two digits after the decimal point makes Figures 3 and 6 identical.

introduced as a benchmark for all the other models; its accuracy is equal to the ratio between the number of SEA "430" samples and the total number of all samples in the test set (98.7 per cent), corresponding to an error rate of 1.30 per cent (Table 8).

Table 8. Accuracy performances of 5 different classification models

	Errors	Error rate (%)	Error rate not considering 430 (%)	Error rate 430 vs. others (%)[1]
G0	4993	1.30	100	1.30
G1	1658	0.41	20.89	0.40
G2	3459	1.01	17.24	1.00
G3	3691	1.02	16.63	1.00
G4	1651	0.41	20.90	0.40

(1) Computed as the error rate for the binary test 430 vs. others.

According to the results reported in Table 8, the classification models G1 and G4 perform best in terms of error rate. As for G2 and G3, they give a more accurate classification of classes different from 430 (Table 8, column "Error rate not considering 430") at the price of making significantly more errors on 430 predictions compared with G1 and G4.

Overall, according to these results G1 and G4 are the best performing models. It is important to note that G1 is more parsimonious than G4 in terms of number and type of features used, making the empirical analysis significantly less demanding from a computational point of view; hence, it represents the best solution for our business environment.

4. Conclusions

The paper explores the possibility of improving the quality of the classification of Italian companies by institutional sector in the entities register (*Anagrafe*) of the Bank of Italy through the implementation of a machine learning solution to be employed in a business process pipeline.

In particular, a hierarchical classification solution is adopted in order to overcome the high degree of imbalancing in as the *Anagrafe*, which is due to the presence of a very high proportion of companies with the same SEA code. A sequence of GBoosts (for non-textual features) and SVM models (for textual features)are estimated. Their results are then combined using a simple ensemble method.

Since data might not follow the SEA hierarchy, which is based on a lot of *a priori* knowledge, a full inductive technique that trains six peer classifiers is built on a sub-set of SEA codes, grouped only on sample cardinality considerations. The classifiers are further evaluated by means of an artificial neural network, thus improving, for the most part, the baseline results in non-430 cases.

The hierarchical technique is developed into four different versions: one model (G1) includes all the numerical and categorical features and the names of the companies; the other three models (G2, G3 and G4) add the introductions to the notes to the financial statements.

The performances of the models are evaluated in terms of error rate and compared using a naive classification model (G0) as a benchmark; the prediction of G0 is always the class having the largest number of instances (430). Therefore models G1 and G4 are the best performing in terms of error rate and G2 and G3 give a more accurate classification of classes different from 430 but at the price of making significantly more errors on 430 predictions compared with G1 and G4.

In conclusion, G1 and G4 are the models that perform best, but the computational complexity of G1 is much lower. Hence, G1 is the best solution for implementation in a business environment.

Future works might consider adding deductive techniques to the hierarchical and ensemble model based on Automated Reasoning, which would assign the SEA code on the basis of rules extracted from *a priori* knowledge.

References

Chawla, N. V., Lazarevic, A., Hall, L. O. and Bowyer, K. W. (2003), *SMOTEBoost: Improving Prediction of the Minority Class in Boosting*. In: Lavrač, N., Gamberger, D., Todorovski, L. and Blockeel, H. (eds) Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science, vol 2838. Springer, Berlin, Heidelberg, pp. 107-19.

Chawla, N. V., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P. (2002), *SMOTE: Synthetic Minority Oversampling Technique*, "Journal of Artificial Intelligence Research", 16, pp. 321-57.

Chen, C. and Breiman, L. (2004), *Using Random Forest to Learn Imbalanced Data*. University of California, Berkeley.

Chen, T. and He, T. (2015), *Xgboost: extreme gradient boosting*, R package version 0.4-2, pp.1-4.

Cortes, C. and Vladimir, V. (1995), *Support-vector networks*, "Machine learning", 20(3), pp. 273-97.

Demuth, H. B., Martin, T. H., Beale, M. H. and Orlando, D. J. (2014), *Neural network design*. Ed. Martin Hagan.

Eurostat (2013), *European System of Accounts – ESA 2010*, Luxembourg.

Friedman, J.H. (2002), *Stochastic gradient boosting*, "Computational statistics & data analysis", 38(4), pp. 367-78.

Gordon, A. D. (1987), *A review of hierarchical classification*, "Journal of the Royal Statistical Society: Series A (General)", 150(2), pp. 119-37.

He, H. and Garcia, E. A. (2009), *Learning from imbalanced data*, "IEEE Transactions on Knowledge & Data Engineering", 9, pp. 1263-84.

IMF (2014), *Government finance statistics manual 2014*, Washington, D.C, p. 31-33

Istat (2009), *Classificazione delle attività economiche - Ateco 2007*, "Collana Metodi e norme", 40.

Japkowicz, N. (2000), *The class imbalance problem: significance and strategies*, Proceedings of the 2000 International Conference on Artificial Intelligence.

Jurafsky, D. and James, H.M. (2009), *Speech and language processing: an introduction to natural language processing, speech recognition, and computational linguistics*, 2nd edition, Prentice-Hall.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2005), *Handling imbalanced datasets: A review*, "GESTS International Transactions on Computer Science and Engineering", 30, pp. 25-36.

Longadge, R., Snehalata, S. D. and Lates, M. (2013), *Class imbalance problem in data mining review*, arXiv preprint arXiv:1305.1707.

Liu, X.Y., Wu, J. and Zhou, Z.H. (2009), *Exploratory undersampling for class-imbalance learning*, Cybernetics, 39(2), pp. 539-50.

Noyvirt, A. (2019), *FinBins – granular classification of the UK's financial sector*, https://datasciencecampus.ons.gov.uk/project/finbins-granular-classification-of-the-uks-financial-sector

Salama, M. A., Hassanien, A. E. and Fahmy, A. A. (2010), *Reducing the influence of normalization on data classification*, "Proceedings of International Conference on Computer Information Systems and Industrial Management Applications", pp. 609-13.

Word, S., Esbensen, K. and Geladi, P. (1987), *Principal component analysis*, "Chemometrics and intelligent laboratory systems", 2(1-3), pp. 37-52.

Tordoff, B. and David, W. M. (2002), *Guided sampling and consensus for motion estimation*, "European conference on computer vision", pp. 82-96.

Weiss, G., and Provost, F. (2003), *Learning when training data are costly: the effect of class distribution on tree induction*, "Journal of Artificial Intelligence Research", 19, pp. 315–54.