Estimating the institutional sectors of legal entities on a large scale

A supervised learning approach

15 December 2020

Francesca Benevolo
European Central Bank

Overview

Machine Learning team at ECB

- The team is part of the Directorate General Statistics of the ECB
- It is a small team that aggregates a bigger network of ML practitioners at the ECB

The Machine Learning team contributes to projects with different degrees of involvement

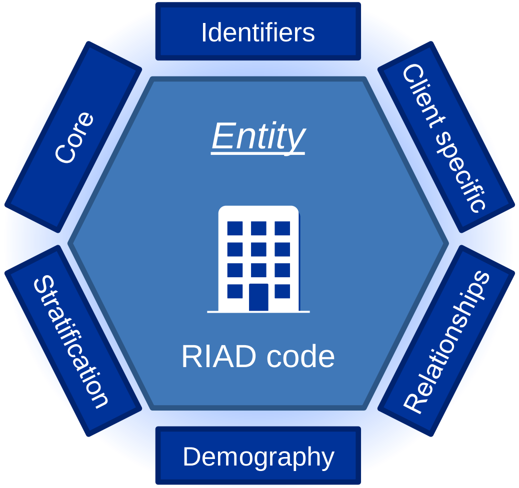
1. Advise
2. Collaborate
3. Lead

less — *Level of involvement* — more

RIAD: shared master dataset on legal entities

- Register of Institutions and Affiliates Data (RIAD) is a ESCB-wide register of legal entities and foreign branches
- Contains information on more than 10,000,000 entities
- Basis for publication of official lists and support several key processes.

RIAD



GLEIF: Legal Entity Identifiers and reference data

GLOBAL LEGAL ENTITY IDENTIFIER FOUNDATION

- GLEIF is a non-profit organisation
- GLEIF provides legal entity identifiers (LEI) for corporations and other organisations
- Contains information on approx. 1,600,000 entities
- Entities involved in financial transactions need to have an LEI

Attributes in GLEIF & RIAD

LEI – Global Legal Entity Identifier Foundation	RIAD
Name	Name
LEI & national ID	Identifier(s)
Address	Address
City	City
Postal code	Postal code
Country	Country
Legal form (ISO 20275)	Legal form
???	ESA 2010 sector

ESA 2010 sector classification*

European System of Accounts (ESA) is internationally compatible accounting framework for a systematic and detailed description of a total economy

	ESA sector	Description
	S11	Non financial corporations
Financial sector	S121	Central banks
Financial sector	S122	Deposit-taking corporations except the central bank
Financial sector	S123	Money Market Funds (MMFs)
Financial sector	S124	Non-MMF investment funds
Financial sector	S125	Financial corporations other than MFIs, non-MMF investment funds, financial auxiliaries, captive financial institutions and money lenders, insurance corporations and pension funds
Financial sector	S126	Financial auxiliaries
Financial sector	S127	Captive financial institutions and money lenders
Financial sector	S128	Insurance corporations
Financial sector	S129	Pension funds
	S1311	Central government (excluding social security funds)
	S1312	State government (excluding social security funds)
	S1313	Local government (excluding social security funds)
	S1314	Social security funds
	S14	Households
	S15	Non profit institutions serving households

* As represented in RIAD

Business case

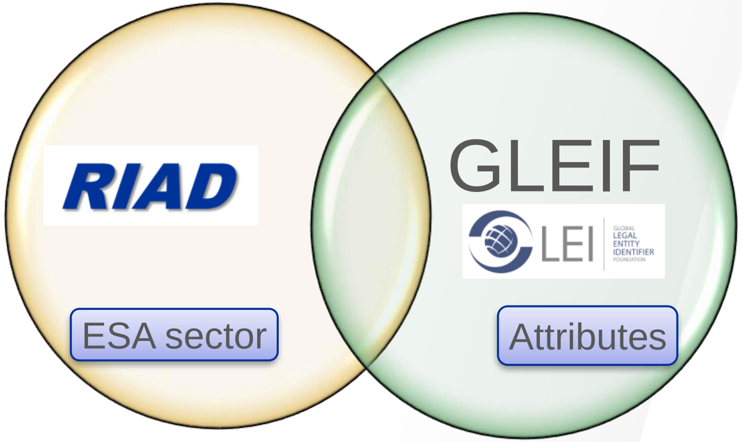
Scope: On-boarding of legal entities from GLEIF* into RIAD**.

Problem: ESA*** sector classification is a mandatory attribute in RIAD,
but it is missing for entities in GLEIF.

Question: How to estimate the ESA sector classification for GLEIF data?

Solution: A supervised learning approach…

Supervised Learning Approach



545 541 entities in both GLEIF and RIAD

➡ These entities have the ESA sector available

TRAINING/TEST DATA

963 652 entities only in GLEIF

➡ ESA sector to be predicted for these entities

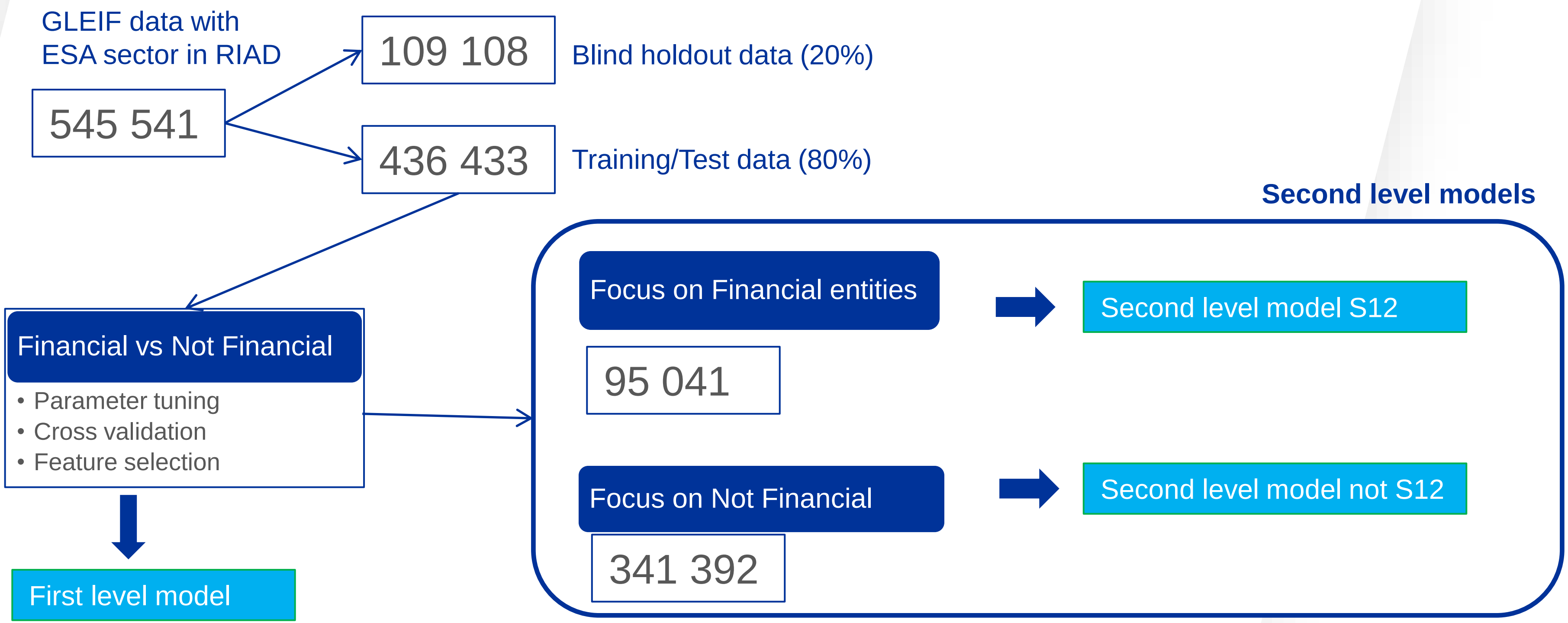
REAL DATA

Training/Test data: entities in both databases.

Target variable: ESA sector.

Predictors: GLEIF attributes.

Process design



Methodology

FEATURE ENGINEERING

Legal Name was encoded with semantic embedding to improve the predictions

PARAMETERS TUNING

Comparison of Random Forest input parameters to find the best combination

CROSS VALIDATION

The best model was selected among 72 options based on the accuracy.

BLIND HOLDOUT DATA

Additional 100 000 entities used to confirm the quality of the models in the end

Methodology

Top features used to predict the ESA sector:

- Category FUND
- Embedded variables from the semantic analysis of legal name
- Luxemburg as legal basis
- Legal form
- Presence of words HOLDING, INVEST, BANK, FUND in the legal name
- Registration authority

Two levels model

Priority: to find all Financial entities (S12)
→ reduce false negative

First level model:
Predict if an entity is financial (S12) or not.

Second level models:
Predict ESA sector 3-digits code.

Distribution of ESA sector in the data

	Frequency	Percentage
Financial S12	118554	**22%**
Not financial S12	426987	**78%**

First level model accuracy score: 90%

Improvement from baseline (78%): the first level model distinguishes financial and not-financial entities with 90% probability

Second level model accuracy score: 73%

Improvement from baseline (43%): the second level model predicts the 3-digits ESA sector for financial entities with 73% probability

Benefits

- Prioritisation - The RIAD team and the National Central Authorities can focus their work on the (predicted) financial entities first.

- Data availability - The predicted ESA sectors will be available for RIAD users much faster than before.

- Efficiency gain - The process is fully automatic to predict the ESA sector for new entities in the future, without any intervention from RIAD experts.

Conclusions

- The close collaboration with the business unit was fundamental to incorporate business needs into the models (example: higher importance to financial entities).
- The semantic analysis on legal names was added value for the models.
- The parameters fine tuning and cross validation search helped to find the best model.
- As result, the application predicted the ESA sector for 963 652 entities in GLEIF.

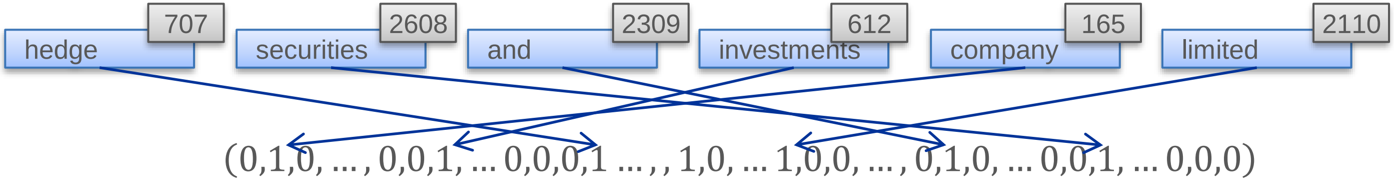
Appendix: Embedded variables

Embedded variables: incorporate name information in the classification task.

HEDGE SECURITIES AND INVESTMENTS COMPANY LIMITED

Traditional approach: Bag-of-words

- Each word corresponds to an index number (using a dictionary)
- Vector setting the index entry to 1 if the word is present.



Drawback:

- Space of words is of very high dimension and sparsely populated
- Word order is lost in this representation

Appendix: Embedded variables