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

19-22 October 2021, Rome / virtual event

Supervised machine learning for estimating the institutional sectors of legal entities on a large scale[1]

Francesca Benevolo, Thomas Gottron, Ilaria Febbo and Nicolò Pegoraro,
European Central Bank

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

Supervised machine learning for estimating the institutional sectors of legal entities on a large scale

Francesca Benevolo, Thomas Gottron, Ilaria Febbo, Nicolò Pegoraro[1]

Abstract

The Register of Institutions and Affiliates Data (RIAD) is the European System of Central Banks' (ESCB) shared register providing master data for more than 10 million legal entities. One of the key RIAD features is the provision of institutional sector classification according to the ESA 2010 methodology. The distinction between different types of financial institutions, non-financial corporations and private versus public sector is of high importance for several ESCB tasks. In fact, information on the institutional sectors is mandatory for all entities in RIAD and is maintained on an on-going basis by experts at National Central Banks and at the European Central Bank. Though, the process of classifying entities by institutional sector is currently manual and time consuming – as necessary to ensure the requested accuracy – and therefore hardly applicable on a large scale, e.g. when a high number of entities need to be imported from external registers.

To address this use case, we present an automated, high-quality approach for the bulk classification of entities according to their (ESA) institutional sector. The estimates produced serve as good preliminary information supporting the expert assessment and the final entity classification. The approach is based on supervised machine learning with a two-level-approach. It makes use of publicly available information on legal entities, e.g. their name, residence, registration authority or legal form. We use a hierarchical setup of ensemble methods tailored to suit the business needs for the hierarchy of the ESA institutional sectors. Furthermore, we use deep neural networks to create semantic embeddings for company names which have shown to improve classification performance. The approach has been tested and evaluated on a dataset of approximately 550,000 known entities and it was applied to estimate the institutional sector for nearly 1 million entities not yet in RIAD.

Keywords: ESA 2010, institutional sector, Legal Entity Identifier, RIAD, machine learning, estimation

JEL classification codes: C130

[1] Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Contents

Introduction

The European System of National and Regional Accounts (ESA 2010) is an internationally compatible EU accounting framework for a systematic and detailed description of an economy (Eurostat, 2013). At the European Central Bank (ECB), the ESA sector classification is used in several business processes and databases.

The Register of Institutions and Affiliates Data (RIAD) is the ESCB's shared register providing master data for more than 10 million legal entities such as banks, private enterprises, and public institutions. One of the key RIAD features is the provision of institutional sector classification according to the ESA 2010. The information about ESA sectors in RIAD is mandatory for all entities and it is well maintained by experts at National Central Banks (NCBs) and the ECB.

ESA sectors are used to distinguish between types of legal entities, e.g. non-financial, deposit-taking corporations, insurance corporations or governmental bodies. The ESA sector classification in RIAD comprises a code with prefix "S" followed by a number with maximum three digits (cf. Table 1). The nine ESA sectors falling under the financial sector start with "S12". The other seven ESA sectors not belonging to the financial sector start with "S11", "S13", "S14" and "S15".

The process of classifying RIAD entities (especially those residing outside EU) by institutional sector is currently performed manually. The process is time consuming because it needs to ensure a high level of accuracy. This approach is hardly applicable on a large scale, e.g. when a high number of entities need to be imported from external registers such as the Global Legal Entity Identifier Foundation (GLEIF). For these use cases, automated approaches are necessary for providing a preliminary ESA sector estimate. The motivation to work with ESA sector preliminary estimates is to streamline experts' work. In this way, ECB and NCBs experts can prioritise their work and focus on entities of higher relevance for the central banking tasks, i.e. financial entities.

In this paper we present an automated, machine learning based approach to estimate the ESA sector of GLEIF entities based on publicly available reference data (e.g. name, address, legal form, registration authority).

The approach leverages on the overlap between GLEIF and RIAD populations and uses this overlap as labelled training data. The labelled data provides the basis for training a two-level supervised machine learning model based on a Random Forest classifier.

We make the following contributions in this paper:

- We investigate the potential of using supervised machine learning for estimating the legal entity ESA sectors solely based on entity reference data. To the best of our knowledge there is little prior work dealing with ESA sector estimation and no prior work performing this task only via reference data.
- We investigate methods for feature engineering of reference data (e.g. feature selection, semantic embeddings, one-hot encoding).
- We systematically test and evaluate different supervised machine learning methods and identify the best solution for the task.
- We demonstrate that the final solution is of high quality and can safely be integrated in the RIAD production environment.

Table 1: ESA sector classification

ESA sector	Description
S11	Non-financial corporations
S121	Central banks
S122	Deposit-taking corporations except the central bank
S123	Money Market Funds (MMFs)
S124	Non-MMF investment funds
S125	Financial corporations other than MFIs, non-MMF investment funds, financial auxiliaries, captive financial institutions and money lenders, insurance corporations and pension funds
S126	Financial auxiliaries
S127	Captive financial institutions and money lenders
S128	Insurance corporations
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

Related work

There is a wide range of research addressing the estimation of economic activity codes or institutional sectors of business units. A recent investigation by ONS (Noyvirt, 2021) looking into machine learning approaches to solve this task concluded that achieving a high accuracy remains a challenge. Approaches for estimating the economic activity or sector of an entity mainly differ in terms of input data, methods and target codification schemes.

In the context of classifying counterparties in the EMIR dataset (Lenoci & Letizia, 2021) external sources were used to provide context. This context information helped in identifying the type of activity of an entity, e.g. because its information was obtained from the ECB's list of monetary financial institutions. The overall solution then involved a knowledge-based classification system.

Many approaches leverage the availability of national codifications for economic types of activity for assigning NACE codes (Eurostat, 2008) to entities. Different machine learning techniques are used in settings where no one-to-one translation between different codification systems is available. The techniques range from the use of multi-level classification systems (Giudice, Massaro, & Vannini, 2020), matching pre-processed textual descriptions (Colasanti, Macchia, & Vicari, 2009) tokenising web texts and generating descriptive features (Kühnemann, van Delden, & Windmeijer, 2020) or supervised solutions like Naïve Bayes, Random Forest, Support Vector Machines, k-Nearest Neighbours or voting ensemble methods (Roelands, van Delden, & Windmeijer, 2018).

A general survey of how to model a probabilistic approach for capturing overlaps of text tokens for coding the occupation sector of survey respondents is discussed in (Gweon, Schonlau, Kaczmirek, Blohm, & Steiner, 2017).

The paper at hand presents an approach that is new in respect to the work available in the literature. To the best of our knowledge there is little prior work addressing the estimation of ESA sector classification. Our work offers a new method of estimating the ESA sector classification, based on machine learning techniques such as text analytics, neural networks, and random forests.

Methodology

A thorough analysis of the classification task and users' needs led to specific methodological choices, i.e. the use of a two-level supervised machine learning model and semantic embeddings.

The reason for selecting a two-level supervised machine learning model derives from the primary need to distinguish financial versus non-financial entities. The first step identifies financial entities, i.e. S12 (independently from the ESA sector detail). The second step estimates the full three-digit ESA sector code (S122, S123, etc.).

The aim of exploring semantic embeddings is to make entity legal name information more manageable and valuable. The entity legal name constitutes the most valuable resource for identifying an entity's nature as well as the largest challenge in processing it. To verify the statistical relevance of the legal name we analysed the words frequency (to find the most explicative for ESA sectors) and generated semantic embeddings using neural networks (with the scope of retrieving hidden meaning from the legal names).

In this section we provide further insights into these two choices.

A supervised learning approach

We modelled the task as a supervised learning approach because we could leverage on the existing overlap between GLEIF and RIAD, i.e. 548,464 legal entities belong to both databases. The common entities were used to align GLEIF features with the target RIAD variables. This aligned data was needed to train and test our models.

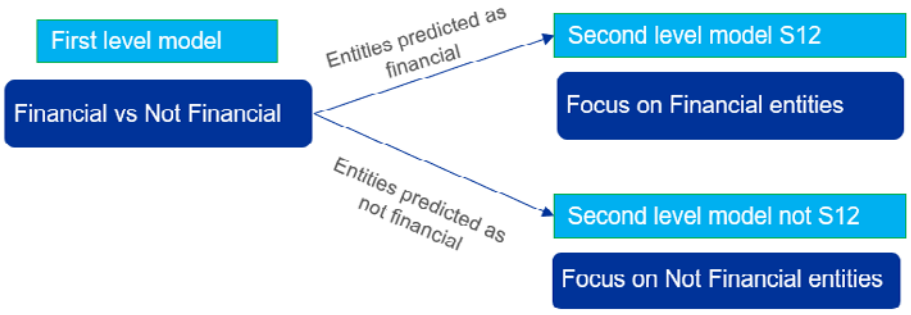
The main steps undertaken to build the model are illustrated in Figure 1 and can be summarised as follows:

1. **Building a first level model to classify observations into financial and non-financial entities.** This task is a binary classification problem aiming to predict if an entity belongs to the financial class. Accordingly, the target variable was re-shaped into a binary variable with a value of 1 if the ESA sector started with "S12" (financial) and a value of 0 otherwise. The decision of this high-level distinction was driven by the business need to primarily distinguish between financial and non-financial entities. The use case of prioritising a further manual assessment implied a conservative approach which is rather biased towards assigning uncertain entities to the financial class. We addressed this requirement by using weighted classes in the classification task. The weights reflected priorities in the

outcome and the corresponding preference for different types of errors in the classification.

2. **Building the two second level models which subsequently predict the detailed ESA sector class.** The second level consists of two sub-models. The first sub-model aims to detail the financial entity class (i.e. distinguish among credit institutions, money market funds, pension funds, central banks, etc.). The second model aims to detail the different types of non-financial entities (e.g. non-financial corporations, governments, households). The financial and non-financial domains are quite heterogeneous. The advantage of having two separate sub-models is to better fine-tune our algorithms and deal with such heterogeneity.

Figure 1: Process design



Legal name analysis

In this paragraph we discuss the impact of the entity legal names on our work, i.e. the analysis of individual words used in legal names as well as the benefit of using semantic embeddings.

We analysed the words frequency in the legal names. The most frequent words were slightly different when considering only entities belonging to the S12 classes (financial). Our assumption was that some words in the legal names were more likely to be used by financial entities since the company name may include indicative hints on the main business of an entity. Legal names containing terms like "bank", "fund" or "insurance" provide good evidence for a financial entity. Instead, if the name contains terms like "manufacturing", "travel" or "transport", the entity probably belongs to the non-financial sector. We decided to include the most frequent words as features for our models, using one-hot encoding. We also included some additional words that were defined as very relevant by the RIAD experts. The business expertise on certain words being indicative for S12 or not S12 was a very good motivation to consider individual words as features.

In the semantic embeddings work, the challenge lies in the text complexity of company names. Information in text is encoded on the semantic level and involves a deeper understanding of concepts and their relations expressed by words. While to a certain extend this knowledge can be encoded in a rule-based static knowledge base it is tedious to set up and maintain such a knowledge base. As an example, the multilingual context was challenging in the sense that we had to deal with words such "bank", "banca" or "banque" that refer to the same concept. Semantic embeddings are a technique to represent texts as high dimensional, numeric vectors, where similar

vectors represent similar concepts or even meaning. It is a well established approach for multiple use cases of text processing.

We computed semantic embeddings using a deep recurrent neural network. For our use case we made good experiences using a relatively simple topology of a bi-directional recurrent neural network composed of LSTM neurons neurons (Hochreiter & Schmidhuber, 1997), following by a multi-layer dense network of ReLU nodes. The network was trained on a large number of legal names taken from GLEIF. The values observed in the last layer of the dense network served as the semantically rich representation of the legal names. These embeddings were included as features in the actual ESA sector classification task.

Data

Our analysis leverages two data sources: GLEIF and RIAD. GLEIF contains a rich feature set for legal entities covering basic reference data. Among others, this feature set contains information on the name, address, legal identifiers, registration authorities or relations to parent companies and subsidiaries. Some GLEIF information items, like the address information, are structured further, e.g. to distinguish between legal address, headquarter address, other addresses, or transliterated addresses. All these features are potential raw input variables for an ESA sector prediction model.

When we started our exercise, it was possible to establish a link between records in RIAD and GLEIF via LEI for approximately 550,000 entities. The first step was to partition the dataset into training, testing and validation data. We decided to take 20% out as blind holdout data before creating the models. The blind holdout data provides the basis for a final evaluation of the machine learning model's performance after it has been trained and tested. It was not used to make decisions about which model to use or for improving or tuning algorithms. The remaining 80% of the labelled data was used to build, train, and test the models. We used a simple random split for partitioning the data.

We started to analyse the data in an explorative phase to get a better understanding of the task. As extensively discussed, this paper aims to estimate the ESA sector for entities in GLEIF. The distribution of the ESA sector classes is unbalanced in our dataset (cf. Table 2). The financial entities (starting with S12) represent the 21.7% of the total.

The non-financial entities (S11, S13, S14, S15) represent the 78.3% of the overall volume. Among those, the non-financial corporations (S11) play the main role: with 415,426 units, they cover 75% of the entities. We took these percentages as benchmark of our model, at least for the binary problem of predicting if an entity was financial or not. Our goal was to build a model that could estimate the ESA sector significantly better that an approach based on flipping a coin with probability 0.783.

Table 2: ESA sector frequency in the data

ESA sector	Count	Frequency
S11	415,426	75.74%
S124	51,226	9.34%
S127	28,321	5.16%
S125	14,065	2.56%
S126	11,990	2.19%
S15	8,840	1.61%
S122	6,471	1.18%
S128	3,404	0.62%
S129	2,754	0.50%
S1313	2,463	0.45%
S14	1,381	0.25%
S1311	804	0.15%
S1314	481	0.09%
S123	444	0.08%
S1312	311	0.06%
S121	83	0.02%

Experiments

First level model

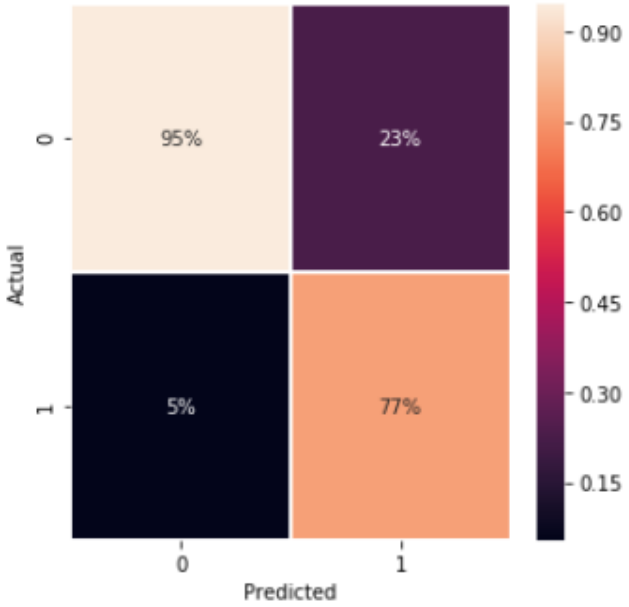
The first level model was a Random Forest Classifier with 100 trees, optimised parameters (from parameter search techniques) and weighted classes. The model was trained and tested on a dataset of 438,771 records with cross validation methods. It was finally validated on our blind holdout dataset of size 109,693. Being a binary classifier, the first level model must deal with two types of errors: false positives and false negatives. The RIAD experts wanted to reduce the error of classifying entities as not financial when they were financial. We translated the business constraint by reducing the false negatives.

Methodologically, we accepted higher false positive, while keeping as low as possible the false negative, i.e. the error of predicting as "Not Financial" the entities that were "Financial" in reality. Therefore, we included a higher weight for the class "Financial". After several experiments, we decided that a triple weight was a good compromise among keeping the percentage of false negatives under 5% and having an acceptable percentage of false positives (<30%).

The confusion matrix in Figure 2 shows the percentage of the real entity type versus the prediction of being financial or not in the validation set. Among the entities predicted as "Not Financial" (Predicted = 0), only 5% (5 480) were "Financial" (Actual=1). The number of false negatives resulted to be quite low, as requested by the RIAD experts. As consequence, we accepted a higher error for the false positives.

In our data, 23% of the entities 79,863) predicted as "Financial" (Predicted = 1) were "Not Financial" (Actual = 0) in reality.

Figure 2: Results on blind hold out data Weighted confusion matrix for first level model



We used the accuracy metric to evaluate the models. The accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. Accuracy is used as a statistical measure indicating how well a binary classification test correctly identifies a condition.

We selected the Random Forest Classifier with triple weights on financial classes, all variables, and semantic embeddings. The selected model was validated on the blind holdout dataset with 109,693 labelled records, which was not used in the test and training phase to guarantee the test integrity. The overall accuracy score of the first level model was 90% on the blind holdout dataset. In other words, in 90% of the cases the first level model could predict correctly if an entity was financial or not. Please note that this value of accuracy must be taken with a grain of salt, as it also reflects the bias of the model to declare entities as financial entities. This tendency of reducing false negative results has a slightly negative impact on the overall accuracy of the model.

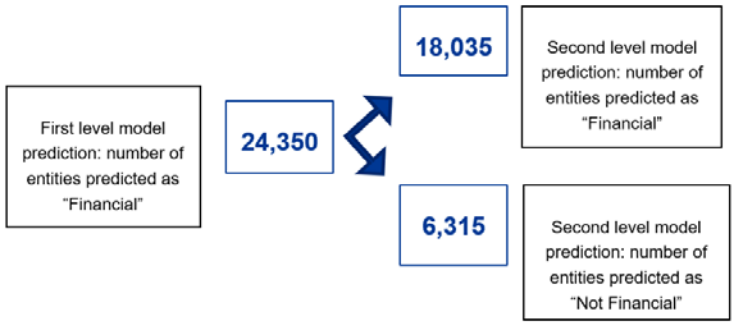
Second level models

The second level model for financial entities aimed to assign the ESA sector to entities that were estimated as "Financial" by the first level model. We trained this second level model on entities that were predicted as S12 (financial) in the first level model. The selected model for financial entities was a Random Forest Classifier for multi-class target with 100 trees and selected parameters. We run some experiments considering XGBoost. Like the first level model, we excluded the XGBoost classifier due to an observed accuracy lower than 90%.

The selected model was validated on 24,350 records from the blind holdout dataset that were predicted as "Financial" by the first level model.

In Figure 3, the number of well predicted "Financial" entities by the second level model is 18,035 on the blind holdout dataset, corresponding to the 73% of the data. This means that the three digits ESA classification was correctly predicted as financial in 73% of the cases. The accuracy dropped significantly from the 90% of the first level model. The reason is related to the business choice of reducing false negatives. Among the 24,350 entities that were predicted as "Financial", we have a 23% of wrongly classified entities from first level model. This percentage is amplified in the second level model, as the errors made earlier cannot be corrected by the second level models.

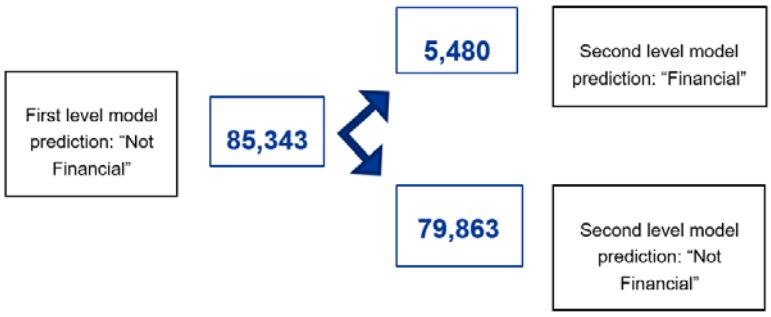
Figure 3: Second-level model for financial entities - Results on blind hold out data



The second level model for not financial entities aimed to assign the ESA sector classification for entities that were predicted as "Not-Financial" by the 1st level model. The input data includes 85,343 records. The experiments motivated us to select a Random Forest Classifier with 100 trees and ad-hoc parameters. As in the previous models, alternatives to the Random Forest Classifiers demonstrated inferior performance.

The model was validated on records from the blind holdout dataset that were predicted as "Not Financial" by the first level model. In Figure 4 the number of well predicted "Not-Financial" entities by the second level model is 79,863 on the blind holdout dataset, corresponding to the 93% of the data. This means that the three digits ESA sector was correctly predicted as not financial in 93% of the cases.

Figure 4: Second-level model for not financial entities - Results on blind hold out data



The model performance was measured by the accuracy scores on the blind holdout dataset, that resulted to be in line with the accuracy scores of the test phase. This means that the model's performance on new data was accurate as much as on the test data.

Overall, we compared the accuracy scores calculated on the validation set with those obtained in the testing phase. The results were very promising and made us confident in the model generalising well. The accuracy scores on the test data were very similar to those calculated on the blind holdout dataset, as shown in Table 3. Our models were predicting the ESA sectors well, compared to the performance during the test phase. The models did not show overfitting or underfitting problems.

Table 3: Accuracy score on test and blind data

	Accuracy score	
Model	Test data	Blind holdout dataset
First level: S12 vs not S12	0.9018	0.9018
Second level: S12	0.7326	0.7301
Second level: not S12	0.9874	0.9367

Conclusions

In this paper we estimated the institutional sector classification for legal entities according to the ESA 2010, based on basic entity reference information publicly available on the GLEIF website. The aim was to assign the correct institutional sector to GLEIF entities, in view of their potential registration in RIAD.

The proposed solution was leveraging data on approx. 550,000 entities from GLEIF which are already recorded in RIAD with their respective ESA sector classification. Such data was used to build a gold standard for training, testing and validating a supervised machine learning approach. The approach is composed of a two-level design, with a first level model to distinguish between financial and non-financial entities and two second level model to perform the fine-grained classification into the final ESA sectors. We employed semantic embeddings methods, feature engineering and selection, and random forest to address the task. The solution was tested and evaluated on a blind holdout dataset. The evaluation showed that the solution achieves high accuracy: 90% for the first-level model, 73% for the second-level model on financial entities and 93% for the second-level model on not financial entities. The adopted strategy also considered business-specific needs to bias the first level classifier to ensure a high recall for identifying financial entities.

The resulting automated, high-quality process for the bulk classification of entities according to their (ESA) institutional sector can be directly reused in the future. The produced estimates will serve as high-quality preliminary information supporting the expert assessment and the final entity classification. As a result, the institutional sector of nearly one million entities from GLEIF have been made available for the assessment of the RIAD experts before the potential recording in RIAD.

References

Colasanti, C., Macchia, S., & Vicari, P. (2009). The automatic coding of Economic Activities descriptions for Web users. *New Techniques and Technologies for Statistics*.

Eurostat. (2008). *NACE Rev. 2 - Statistical classification of economic activities*. Eurostat.

Eurostat. (2013). *The European System of Accounts — ESA 2010. Official Journal as Annex A of Regulation (EU), 549. doi:10.2785/16644*.

Giudice, O., Massaro, P., & Vannini, I. (2020, March). Institutional sector classifier, a machine learning approach. *Occasional Papers (Questioni di Economia e Finanza)*(548). Retrieved from https://www.bancaditalia.it/pubblicazioni/qef/2020-0548/QEF_548_20.pdf

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three Methods for Occupation Coding Based on. *Journal of Official Statistics, 33*(1), 101-122.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780.

ISTAT. (2009). *Classificazione delle attività economiche Ateco 2007*. Roma: Istituto nazionale di statistica. Retrieved from https://www.istat.it/it/files//2011/03/metenorme09_40classificazione_attivita_economiche_2007.pdf

Kühnemann, H., van Delden, A., & Windmeijer, D. (2020). Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. *Statistical Journal of the IAOS, 36*(3), 807-821.

Lenoci, F., & Letizia, E. (2021). Classifying counterparty sector in EMIR data. In *Data Science for Economics and Finance*. Springer.

Noyvirt, A. (2021). *FinBins – granular classification of the UK's financial sector*. (Office for National Statistics - Data Science Campus) Retrieved April 28, 2021, from https://datasciencecampus.ons.gov.uk/project/finbins-granular-classification-of-the-uks-financial-sector/

Roelands, M., van Delden, A., & Windmeijer, D. (2018). *Classifying businesses by economic activity using web-based text minin*. Statistics Netherlands.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Estimating the institutional sectors of legal entities on a large scale

A supervised learning approach

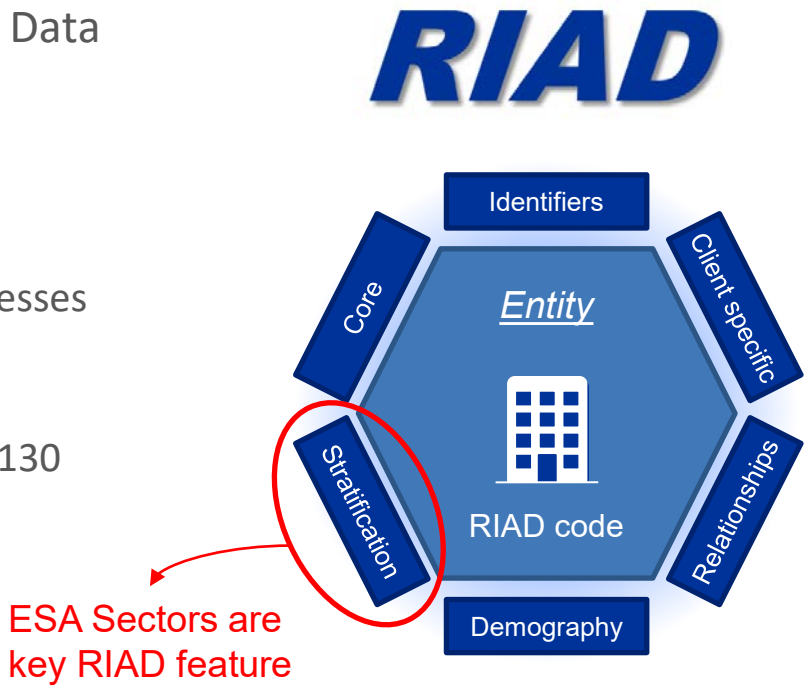
Francesca Benevolo, Thomas Gottron, Ilaria Febbo, Nicolò Pegoraro

European Central Bank

20th October 2021

RIAD: shared master dataset on legal entities

- The Register of Institutions and Affiliates Data (RIAD):
  - is a shared master-dataset
  - supports several clients and business processes across the ESCB[1], SSM[2] and EBA[3]
  - Has more than 12 Mn entities, more than 130 attributes



[1] European System of Central Banks
[2] Single Supervisory Mechanism
[3] European Banking Authority

ESA 2010 sector classification*

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

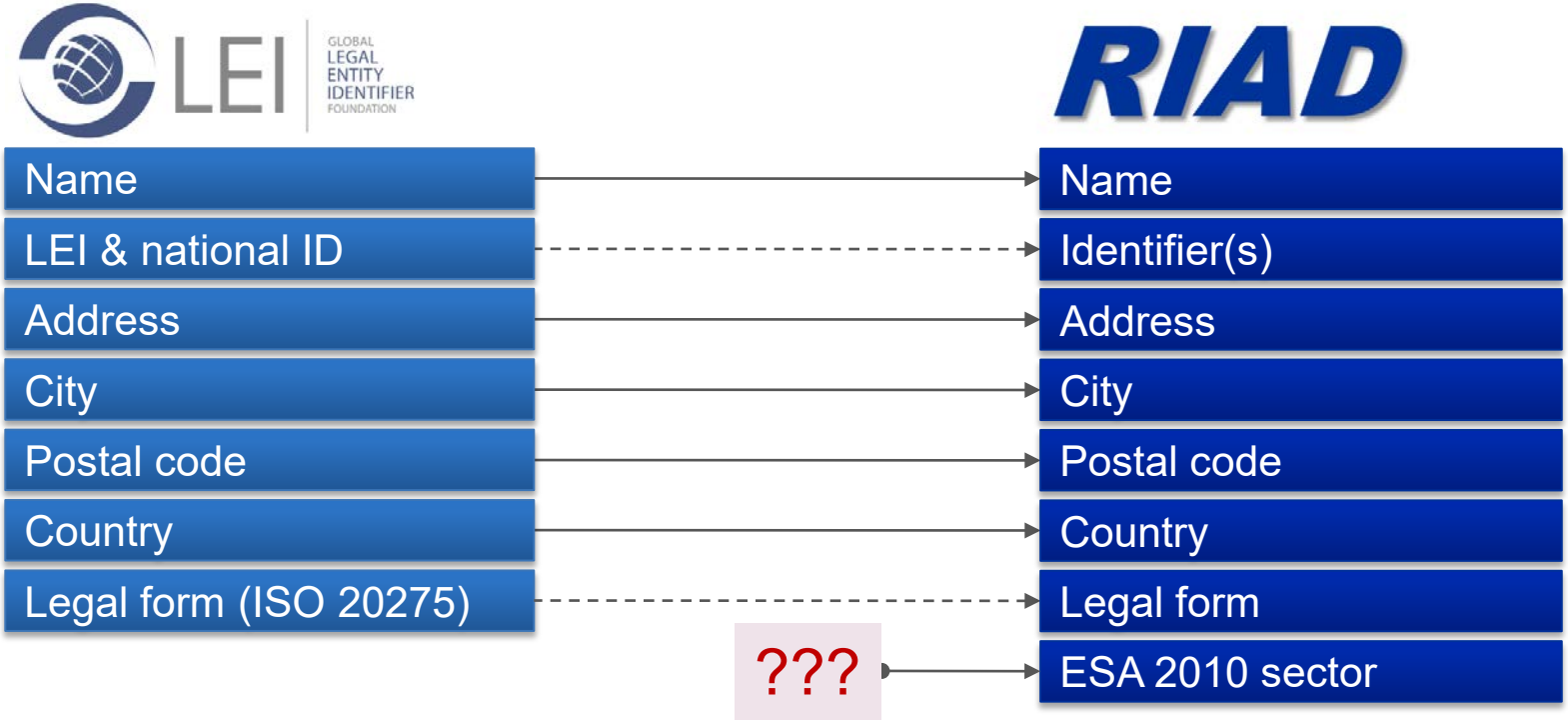
	ESA sector	Description
	S11	Non financial corporations
Financial sector	S121	Central banks
	S122	Deposit-taking corporations except the central bank
	S123	Money Market Funds (MMFs)
	S124	Non-MMF investment funds
	S125	Financial corporations other than MFIs, non-MMF investment funds, financial auxiliaries, captive financial institutions and money lenders, insurance corporations and pension funds
	S126	Financial auxiliaries
	S127	Captive financial institutions and money lenders
	S128	Insurance corporations
	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

GLEIF: Legal Entity Identifiers and reference data

- GLEIF is a non-profit organisation
- GLEIF provides legal entity identifiers (LEI) for corporations and other organisations
- Contains information on approx. 1,6 Mn entities
- Entities involved in financial transactions need to have an LEI

Attributes in GLEIF & RIAD



Business case

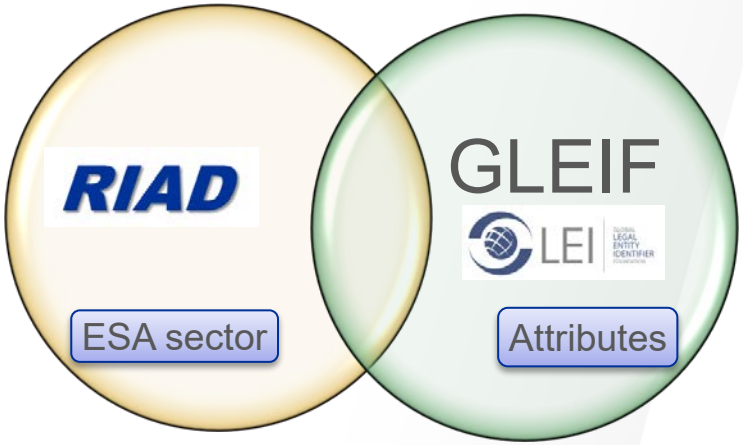
Problem: Assigning ESA sectors to newly recorded RIAD entities is a manual and time consuming process (esp. for non-EU).

Scope: Creating a model based on public GLEIF information able to automatically estimate ESA sectors so to streamline RIAD experts' work.

Question: How to estimate the ESA sector classification based only on GLEIF data starting from a simple LEI code?

Solution: A supervised learning approach trained on RIAD data applicable for present and future needs.

Supervised Learning Approach



545,541 entities in both GLEIF and RIAD

➡ These entities have the ESA sector available

TRAIN, TEST, VALIDATE

963,652 entities only in GLEIF

➡ ESA sector to be predicted for these entities

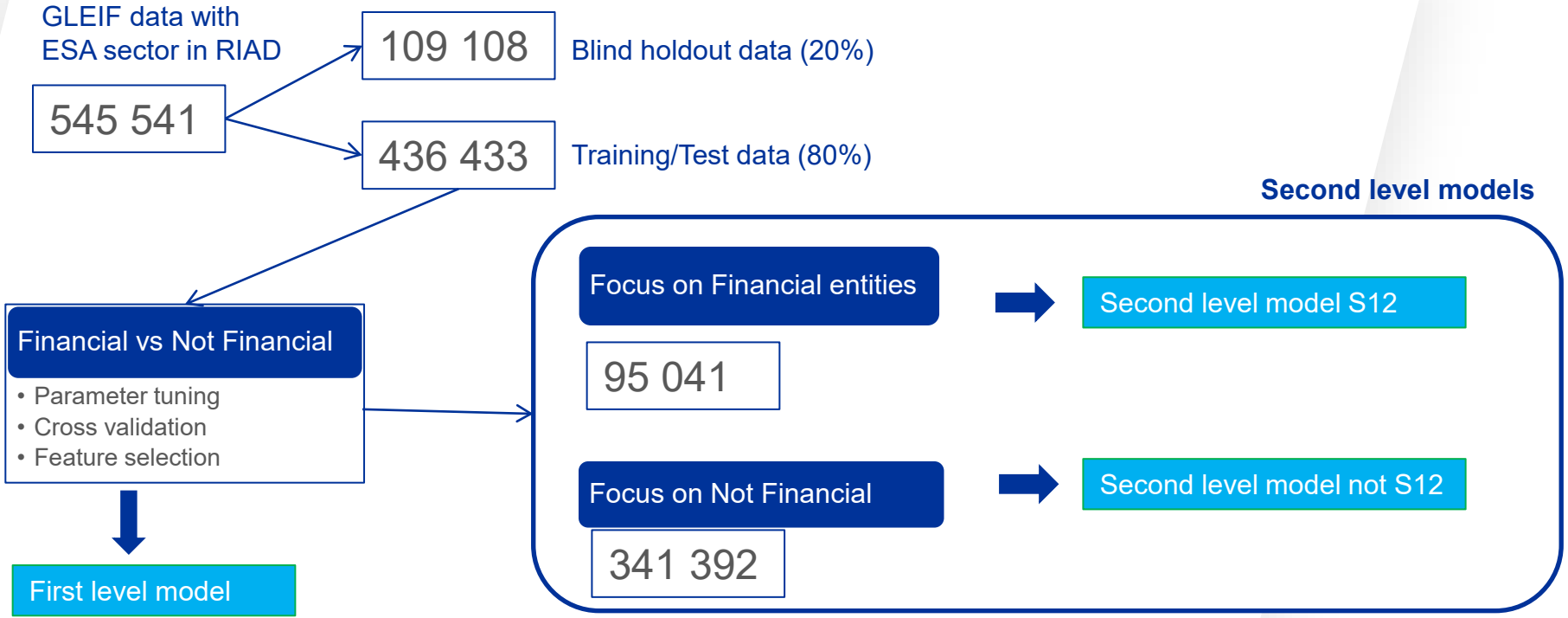
REAL DATA

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

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	118,554	**22%**
Not financial S12	426,987	**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

Conclusions

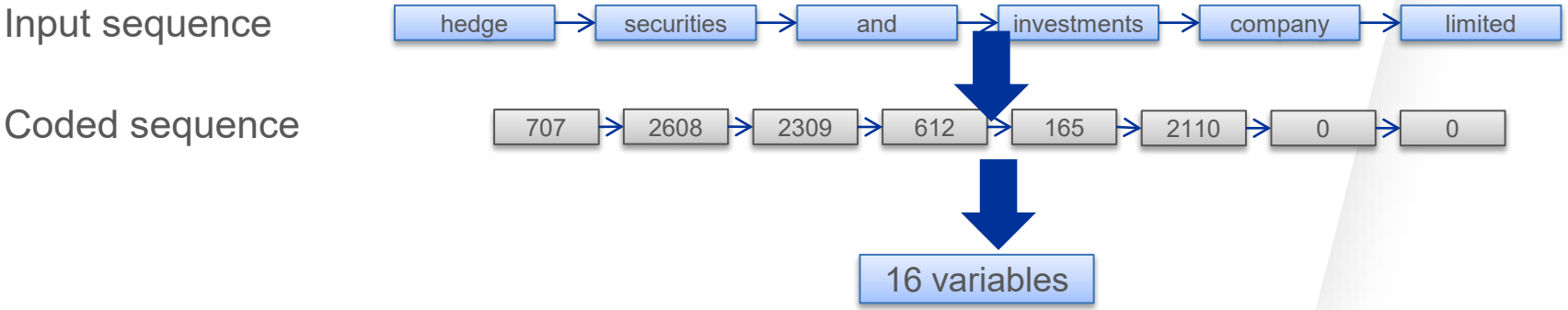
- The application estimated the ESA sector for 963,652 GLEIF entities.
- The semantic analysis on legal names added value to the models.
- The parameters fine tuning and cross validation search helped to find the best model.
- Benefits for the business areas (efficiency gain, prioritisation, data availability)
- The innovative aspect of our work was to estimate missing data using reference data only.

Appendix

Appendix: Embedded variables

The Legal Name was encoded with semantic embedding.

Results: 16 embedded variables were generated and used as models predictors, improving the overall accuracy.



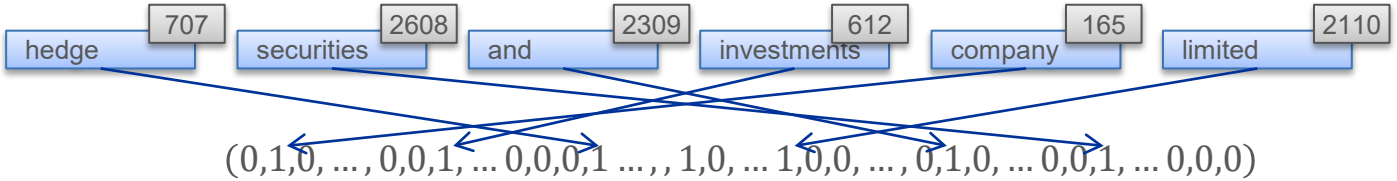
Appendix: Embedded variables

Embedded variables: incorporate name information in the classification task.

HEDGE SECURITIES AND INVESTMENTS COMPANY LIMITED

Traditional approach: Bag-of-words

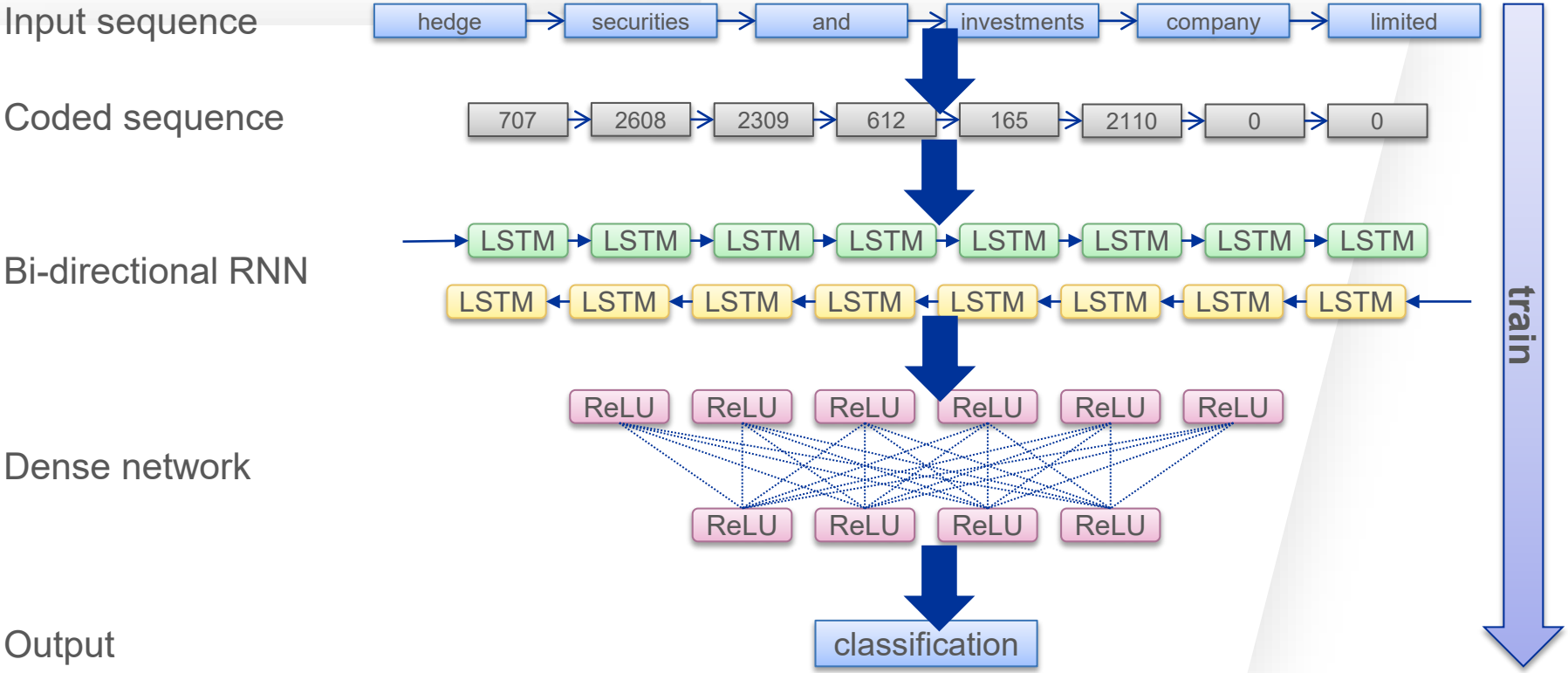
- Each word corresponds to an index number (using a dictionary)
- Vector setting the index entry to 1 if the word is present.



Drawback:

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

Appendix: Embedded variables



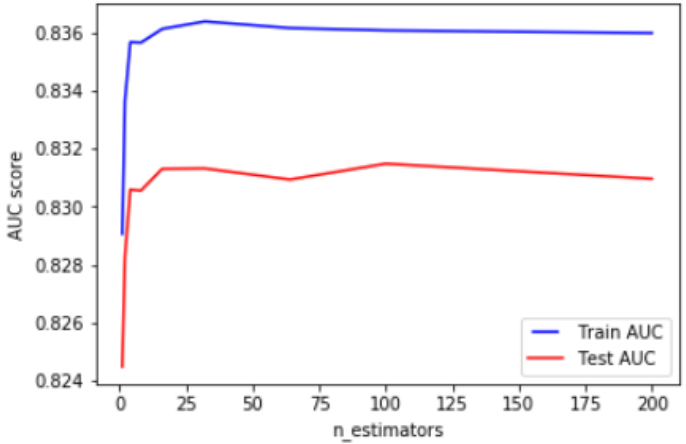
Appendix: Methodology

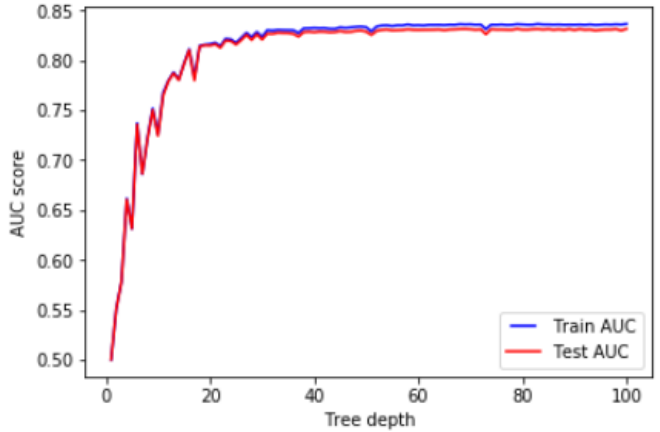
Top features used to predict the ESA sector:

- 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

Methodology: Parameters fine tuning

Random Forest parameters: N estimators, tree depth, minimum sample leaf, min sample split, max features.





Methodology: Cross validation

Random Forest parameters grid with 3 folds for 24 combinations, totalling 72 fits

```
from sklearn.model_selection import GridSearchCV
# Create the parameter grid based on the results of random search
param_grid = {
    'bootstrap': [True],
    'max_depth': [40, 60],
    'max_features': [10, 20],
    #'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [10, 20, 30],
    'n_estimators': [10, 100]
}

# Instantiate the grid search model
rf = RandomForestClassifier(random_state = 42)
grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                           cv = 3, n_jobs = -1, verbose = 2)
```

Best model from Grid Search CV

```
{'bootstrap': True,
 'max_depth': 40,
 'max_features': 20,
 'min_samples_split': 20,
 'n_estimators': 100}
```

Methodology: Blind holdout dataset

The models were evaluated on a blind holdout dataset to verify the accuracy.

Result: The accuracy scores on the blind holdout dataset were very close to the accuracy scores on test data.

The models are stable