NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

DATA SCIENCE FOR THE ASSESSMENT OF OPERATION ERRORS AND THEIR IMPACT

Laura Maria Romeiro Santos

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DATA SCIENCE FOR THE ASSESSMENT OF OPERATION ERRORS AND THEIR IMPACT

by

Laura Maria Romeiro Santos

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics.

Supervised by

Pedro Simões Coelho, PhD, Nova IMS

December, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, December 2nd 2023

ACKNOWLEDGEMENTS

The conclusion of my thesis was a personal achievement, but it was only possible due to the people around me who supported me every day and pushed me forward.

First and foremost, thank my supervisor for his orientation and support, valuable insights, and guidelines. I am grateful for being introduced to this project and for letting me contribute to the first part of the investigation and extend the project to my thesis. Secondly, I want to thank Professor Jorge Mendes, that was also a key person in this project. Thank you for your availability and knowledge sharing during the process of writing my thesis.

Secondly, I want to thank Professor Jorge Mendes, who was also a key person in this project. Thank you for your availability and knowledge sharing while writing my thesis.

To my colleagues, thank you for the time you took to debate some ideas and for introducing an outside perspective, which is always essential.

To my friends, thank you for your friendship, for constantly checking up on me and keeping me motivated, respecting my time, and cheering me up.

To my boyfriend, my rock in this adventure. Thank you for being patient, supporting me every day, especially on the most challenging ones. This would not have been possible without you.

Finally, to my family, I want to leave the deepest “obrigada”. You were always there for me in my academic journey and made the surrounding environment fitting for me to achieve the best results. This milestone is because, and for you.

ABSTRACT

The vast amount financed annually by the European Union for community funds requires rigorous monitoring of the operations allocated to funds. Nowadays, this monitoring is achieved through auditing, which is laborious, costly, and naturally done solely upon a sample of operations, not guaranteeing the conclusiveness of the results. This study explores the feasibility of applying data-science techniques to predict the error of the operations, in complement or, ultimately, replacement of the auditing procedure. We tested two estimation methods based on the data from three funds over two auditing years. We compared them with a benchmark estimation approach in terms of the ability and precision to predict the total error amount and conclusiveness. We experimented with the framework of this study with 1-stage and 2-stage approaches, the last adding a classification step to filter only the operations with error to the prediction stage. Our findings confirm that the assessment of operation error benefits from using a model compared to the traditional estimation method, especially when using the model-assisted estimation. Furthermore, a 1-stage prediction pipeline produces satisfactory results, which are not improved with an additional modeling step.

KEYWORDS

Community Funds; Error prediction; XGBoost; Model-based estimation

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AB	AdaBoost
ANN	Artificial Neural Networks
CF	Cohesion Fund
CI	Confidence Interval
DB	Design-based
DT	Decision Tree
ERDF	European Regional Development Fund
ESF	European Social Fund
ESIF	European Structural and Investment Funds
EU	European Union
GBRT	Gradient Boosting Regression Tree
GNI	Gross National Income
HT	Horvitz-Thompson
IGF	Inspeção Geral de Finanças
LR	Logistic Regression
MUS	Monetary Unit Sampling
MA	Model-assisted
MB	Model-based
ML	Machine Learning
POAT	Programa Operacional Assistência Técnica
RF	Random Forest
RFE	Recursive Feature Elimination
SVM	Support Vector Machine

1. INTRODUCTION

A significant proportion of the European Union's funds are allocated through five different European Structural and Investment funds (ESIF), which support European's employment policy while aiming at promoting a sustainable economy and environment. Among the ESIF there are: the European regional development fund (ERDF) whose goal is to mitigate the unbalanced development across the different regions of the EU; the European social fund (ESF) which is employment-focused and provides investment and support for European citizens looking for a job; finally the Cohesion fund (CF) was developed to uphold the economic, social and territorial cohesion within the EU and render support to states whose gross national income (GNI) per capita is under 90% compared to the average of the other 27 member states (European Commission, 2015).

In Portugal a project called "Programa Operacional Assistência Técnica" (POAT) was designed as a support tool for an improved intervention of the organisms engaged in the management of ESIF (POAT, 2021). One of the focal points is the development of monitorization and evaluation models that ensure quality, innovation and transparency. Aligned to the mission of this project is the authority of structural funds audit - Inspeção-Geral de Finanças (IGF)- who is responsible for the audit of all the operational programs. These audits are paramount to guaranteeing a continuous improvement in internal control systems or to detect expenses that fall outside the scope of the eligibility regulatory framework.

Upon the impossibility of including all observations of a population when conducting statistical testing in auditing procedures, data are collected on a subset of the population also known as samples. This is referred to as sampling audit. The samples can then be used to make statistical inferences about the population through two different principal approaches: design-based and model-based (Lohr, 2019; Särndal et al., 1978). In the design-based approach, inference depends on randomly assigning a subset of population observations to be in the sample and the probabilistic component is introduced by the sampling plan. In the model-based approach, there is an extra probabilistic component because the inference relies on the distributional assumptions about the stochastic process for data generation (superpopulation) which are specified by a model. There is an extensive historical context on both theories and their advantages and disadvantages (Lohr, 2019).

Even with the momentous audits, they can incur significant costs, human resources, and durations, especially given the scale in which they are performed. The estimated errors at the program level have severe repercussions which can include significant financial corrections, including the application of flat-rate corrections by the European Commission and lead to the need for developing improvement plans. Considering the prejudicial role these errors play in developing public policies and the substantial financial volume associated with the corrections, ensuring reliability in the produced estimations while reducing the required effort

for audits is paramount. Moreover, sampling-based audits can lead to inconclusive results, which brings about additional audit necessities.

Data science can be the answer the limitations imposed by currently applied methods in auditing operations. Therefore, this study aims to apply data science-based techniques such as Machine Learning (ML) to estimate and predict the number of errors and their impact in the operational programs supported by community funds in Portugal. These methodologies are expected to replace or complement traditional audit activities exclusively based on sampling audits.

The contribution of this paper is two-fold. First, it extends the growing research regarding the application of data science focused on auditing (Huang et al., 2022; Mabelane et al., 2023) - but still inexistent when narrowed down to the error prediction paradigm- by applying ML algorithms to the prediction domain in alternative to parameter estimation used in statistical sampling. Secondly, it provides a comparison between methods applied to operations at the national level, which EU can leverage to bridge the gap between the way audit procedures are conducted at the moment and the advanced data science methods that are currently available.

The paper is structured in the following way. In section 2 the previously studied Survey Sampling theories as well as the application on data science in auditing are reviewed. This is followed by the proposed method and the research methodology in sections 3 and 4 and the results and discussion in section 5. Finally, we present the conclusions and outline open topics for future research in section 6.

2. LITERATURE REVIEW

2.1. SURVEY SAMPLING THEORIES: DESIGN-BASED AND MODEL-BASED APPROACHES

Boosted by several technological changes and the increase of available data, which precluded the examination of the entire population, audit procedures have not been applied to the entire population for a long time, requiring the application of statistical sampling methods to make inferences about the total amount of errors in the population (Roberts, 1978).

Therefore, the literature regarding finite population inference is highly connected with Survey Sampling (Smith, 1976). Survey Sampling Theory[1] relies on the use of probability sampling[2] and assumes that the observation y_i made on the i^{th} unit is accurate. The estimate error occurs only due to the random sampling variation that arises when n units are measured instead of the entire population of dimension N. Moreover, an estimator represents the method used to calculate an estimate of a population characteristic μ based on the sample results. In contrast, the word estimate is the value obtained from a specific sample. Furthermore, an estimator $\hat{\mu}$ of μ given by a sampling plan is named unbiased if the mean value of $\hat{\mu}$, taken over all possible samples from the plan, is equal to μ (Cochran, 1977)

The first studies related to this theory go back to the early 90's when Bowley (1906) evaluated the accuracy of estimates derived from large random samples that were drawn from large finite populations.

Following Bowley's work, Neyman (1934) investigated the foundations of inference from finite populations, resulting in an influential paper that set the grounds for the design-based approach. The paper provided a statistical framework for sampling theory. It proposed the definition of a representative sampling method along with a consistent estimation method, also referred to as Classical Survey Sampling[3]. He proved that random sampling provided representative samples since it was possible to construct confidence intervals from them, providing consistent inferences for all samples repeatedly drawn following a specific random sampling method.

Horvitz & Thompson (1952) developed a method to obtain unbiased estimates known as the Horvitz-Thompson estimator to complete the Classical Survey Sampling Theory. Given that the inclusion probability is known to the researcher, the method can be used to obtain a reliable estimate, independently of the probability sampling plan.

[1] The aim of Survey Sampling Theory is to enhance the effectiveness of sampling by devising techniques for selecting samples and making estimations that are accurate enough for the intended purpose (Cochran, 1977).

[2] Probability sampling is a method characterized by known selection probabilities assigned to each possible sample, selection through a random process based on these probabilities, and a defined method for calculating a single estimate for any specific sample, allowing for the development of Sampling Theory and the calculation of frequency distribution of estimates (Cochran, 1977).

[3] Classical Sampling Theory and Sample Survey Theory differ in their assumptions about the measurements made on sampling units in the population. The former assumes that these measurements follow a frequency distribution, such as the normal distribution. On the other hand, the latter assumes very little information about this frequency distribution, making the approach model-free or distribution-free (Cochran, 1977).

Godambe (1955) was one of the first statisticians to present a limitation to the design-based approach regarding the impossibility of a unique linear unbiased estimate with the least variance for an entire class of linear estimates. Therefore, he proposed an alternative procedure for estimation, which requires formulating assumptions about the population and, when implemented repetitively, can guarantee the least variance on the average. Further work from the author contributed to developing the design-based inference theory (Godambe, 1965).

In light of this limitation and others, statisticians demonstrated a growing interest in a new sampling theory approach based on Fisher's inferential framework (Fisher, 1922), which developed into what today is known as the model-based theory. Cassel et al. (1976) later established this theory's foundations.

Further inference analysis has been conducted by Hansen et al. (1983), who provided some principles for inferences, including the dependence of the best model-dependent estimators and confidence intervals on the validity of the model, the beneficial usage of a "good" enough estimator instead of the "best" model-dependent estimators under certain conditions, and the inappropriate use of models for inference when samples have a large size, cases in which models should only be used to assist the design of probability samples.

In the audit context, despite the information being restricted to the audit values of the sample units, in some cases, when information about the book values – recorded euro values in an account or unit- of those units and the population is also available, the auxiliary information[4] can be used to improve the precision of the estimates. The aim is to use the correlation between the response variable and the auxiliary variable of each sample unit to increase the precision, as described in the work by Cochran (1977).

A research work by Neter & Loebbecke (1975) compared the performance of eleven statistical estimators in sampling accounting populations, some of which are to be used when auxiliary data is available, such as difference and ratio estimators. A section of this study also examined the performance of an estimator when dollar unit sampling is used. This approach, which has evolved into monetary unit sampling (MUS), started to appear in the 70's (Kaplan, 1975) and considers an individual dollar to be a sample unit, meaning that higher valued amounts in the population have a higher selection probability.

The conclusions drawn from Neter & Loebbecke's research revolved around two fundamental aspects of statistical procedures, namely the precision and the reliability of the nominal confidence coefficient (indicator used by auditors to measure the degree of certainty that the estimator provided correct confidence intervals for a given sample). For dollar unit sampling, the dollar unit mean-per-unit estimator outperformed the analogous stratified estimator in terms of precision and similar precision compared to the auxiliary information. Given the high

[4] In most cases, the auxiliary variable is either associated with the observed variable or a crude approximation of the primary variable of interest. However, in the context of auditing, the auxiliary variable - the recorded book value of each item in the population - is highly correlated to the primary variable or even equal to it.

performance of estimators using MUS, the method has been widely used in auditing since its appearance, particularly in the context of error estimation in European funds in Portugal.

Neter & Loebbecke's paper was extended by Baker & Copeland (1979), who analyzed a stratified version of the regression estimator in the auditing context. Like the auxiliary information estimators analyzed in the previous paper, the regression estimates require both the audit values of a random sample and the population book values. This estimator had superior to equal results to another estimator regarding precision but needed a better reliability level. Therefore, the author suggests not using this technique with a low expected error rate. Beck (1980) further analyzed the regression estimator in populations with more significant errors, whose results showed that regression estimates are robust even when the predictive linear model assumptions are violated.

Despite the extensive use of statistical sampling, new developments in technologies and study domains such as data science and data analytics are turning the page for auditing procedures, enabling once unfeasible methods. Instead of focusing on parameter estimation, where the goal is to produce estimates from β that underlie the relationship between the output value (y) and input variables (x) (Mullainathan & Spiess, 2017), statistical models can be applied to data to predict new or future observations, particularly in cases where there are new observations whose y we are interested in predicting, based on their x (Shmueli, 2010).

2.2. DATA SCIENCE IN AUDITING AND ERROR PREDICTION

Data science is a multidisciplinary approach that studies the generalizable extraction of knowledge from data (Dhar, 2013) by applying math and statistic methods, machine learning algorithms, and artificial intelligence.

Machine learning (ML) is a research field that addresses the development of computer programs that can improve automatically with experience (past information available) concerning a set of tasks (Mitchell, 1997). Designing accurate and efficient algorithms is the core of machine learning. The data plays a crucial role in determining a learning algorithm's effectiveness; hence, ML is intrinsically related to statistics and data analysis. ML can be applied to several domains, from Computer Vision, text, and document classification to Natural Language Processing using different learning tasks, including classification, regression, or clustering (Mohri et al., 2018).

There are several ML methods, including but not restricted to supervised, unsupervised, semi-supervised, and reinforcement learning methods. In supervised learning, algorithms apply a mapping from a numerical representation of observations to target values or categories in regression and classification tasks, respectively. Apart from being more prediction-driven, "supervised" ML focuses more on developing computation models able to efficiently perform the inference compared to the field of statistical inference, where the focus is on developing

a data model able to describe the probabilistic distribution of the feature and target values and their relation (Hao & Ho, 2019). Unsupervised learning uncovers meaningful input representations, such as associations and correlations, by using unlabeled data and considering assumptions about its structural properties (Jordan & Mitchell, 2015). In semi-supervised learning, the training data consists of labeled and non-labeled data, and the prediction is done for unseen observations. Reinforcement learning interlaces the training and testing phases, and the objective is to iteratively maximize the reward gained from the interactions with the environment. A critical aspect of ML is generalization as the objective, particularly in the supervised task, is to make predictions about unseen observations based on a finite sample of observations and, consequently, the generalization error, which measures how accurately the algorithm can perform such task (Mohri et al., 2012). Developments in ML algorithms, triggered by the need to reduce the prediction generalization error, lead to the appearance of ensemble modeling. This process uses multiple different models to predict an outcome and aggregates the prediction of each base model to make the final prediction of the unseen example. Under the condition that the base models are independent of each other, this approach leads to increased prediction accuracy (Kotu & Deshpande, 2015).

Within the scope of auditing, ML algorithms such as logistic regression (LR), back-propagation neural networks, decision trees (DT) and support vector machines (SVMs) have been used for fraud risk assessment (Song et al., 2014) and artificial neural networks (ANN), bagging and stacking algorithms for financial statement fraud detection (Ashtiani & Raahemi, 2022; J. Perols, 2011; J. L. Perols et al., 2017). Gradient boosted regression tree (GBRT) has been studied for misstatement detection (Bertomeu et al., 2021). The Zhang et al. (2021) have suggested using five popular ML models in this domain such as LR, random forest (RF), SVM, ANN and AdaBoost (AB) to identify informative indicators of audit quality. An algorithm that has attained state-of-the-art results in multiple real-world scenarios ranging from store sales prediction and high energy physics event classification to customer churn prediction (Hanif, 2020) is XGBoost. Within the scope of auditing and accounting, the algorithm has already been studied for financial statement fraud detection (Ali et al., 2023). Similar to the results achieved in other tasks, in this study, XGBoost performed best among six models, including LR, DT, SVM, AB, and RF.

In recent years, an increasing number of papers have been addressing the use of ML to tackle audit sampling issues and leverage the power of algorithms to analyze the entire population (Y. Chen et al., 2022), a procedure known as "audit-by exception". Despite the limitations regarding the lack of explainability of ML models, difficulty in documenting the use of analytic techniques, and need for clarity on how to draw the line for auditor's responsibility in complete population testing, this approach is becoming increasingly significant in today's macroenvironment. In this regard, a research paper has demonstrated how to use techniques for Explainability Artificial Intelligence (XAI), including Local Interpretable Model-agnostic

Explanations (LIME) and Shapley Additive exPlanations (SHAP) to fulfil the requirements of audit principles (Zhang, Cho, et al., 2021).

However, to date, we have not found any literature focusing on employing ML predictive models in the auditing procedure for error prediction.

3. PROPOSED METHOD

This section describes the two primary types of inference used for finite populations: design-based and model-based. Moreover, each type of inference relates to an estimator used for parameter estimator, which will also be covered in this section. ML models are employed in the model-based approach and in a type of estimation that will also be covered below, that combines the two main approaches.

3.1. TYPES OF INFERENCE AND ESTIMATORS

It is possible to distinguish between direct and indirect estimators among the estimators. While both estimators can use information regarding auxiliary variables inside and outside the study domain, the direct estimators can only use the values within the study domain and consider the period when it comes to the variable of interest. Considering that the study domain of the present project corresponds to an operational program or group of operational programs in a given accounting year, indirect estimators can be defined as estimators that use data from different funds or programs to estimate the errors of a given program.

Indirect estimators can be divided into synthetic, direct modified, and combined estimators. Synthetic estimators are indirect estimators whose properties depend on the assumptions of a postulated model. Direct modified estimators are a version of the synthetic estimators to which is added a factor of bias correction. From a sampling plan view, synthetic estimators can be biased and inconsistent, while direct modified estimators keep some interesting statistical properties over repeated sampling, namely non-biased and consistency. Finally, combined estimators, or composite estimators, are a combination of direct or direct modified and synthetic estimators.

The methodology comprises two types of inference designated as design-based, linked to direct estimators, and model-based, linked to indirect estimators. The main difference between these two inferences resides in the randomness component, which guarantees the stochastic property of the inference (Särndal et al., 1978).

The objective of a design-based inference is to infer over the descriptive parameters of the finite population, which can be expressed as functions of the vector of values of the variable of interest.

The difference between these estimators lies on their properties. Firstly, the design-based approach considers that the vector of values taken by the variable of interest over the finite population is a constant parameter. Instead, the model-based approach is a realization of random variables whose distribution is specified by a model. Concerning the probabilistic component, this is introduced solely by adopting a sampling plan in the design-based approach. In contrast, in the model-based approach, it is introduced by the same component and an additional one in the form of a model hypothesis. In both inferences, the observation

of the values of the variable of interest is limited to the units of a sample selected from the finite population.

Furthermore, in the design-based approach, the properties of the estimators are evaluated based on the sampling plan. In the model-based approach, the assessment depends on whether the inference was made over the parameters of the superpopulation model or the finite population. In the second case, besides being based on the sampling plan, similarly to the first case, the evaluation is also based on the postulated model.

There are two sub-categories of inferences within the model-based approach: pure model-based and model-assisted. The pure model-based is solely based on the model and does not consider the sampling plan, which the model-assisted approach considers along with its properties evaluation. In other words, the model assists the estimation, but the sampling plan's statistical properties are also considered. There is no assumption that the finite population was generated from the postulated model, only that it can be described approximately by the model. The model-assisted approach is linked to the direct modified estimator. The statistical quality of the inference of both sub-categories of inferences depends on verifying the hypothesis underlying the model. However, in model-assisted inference, it is usual to look for estimators that keep some basic design-based properties that are not dependent on these assumptions. In the model-based approach, the models will be used to predict the error amount, risk of material error, and the prediction precision. In the model-assisted approach, the models integrate the extrapolation process regarding the operations sampling for audit.

3.2. Precision and Metrics

Generally, precision and bias metrics are given by the estimators' variance and expected value. In the design-based approach, these are obtained with repeated sampling and are referred to as design-based expected value and variance. In the alternative approach, the metrics can be obtained based on the model assumptions and are, therefore, referred to as model-based expected value and variance.

3.3. Bootstrap Technique For Variance Estimation

Synthetic estimation is a pure model-based approach where there are usually no assumptions about properties such as unbiasedness and consistency over repeated sampling, as mentioned in the sub-section above. The statistical properties are evaluated according to the sampling plan or exclusively according to the supposed model that generated the finite population. In the particular case of these estimators, when using an algorithm estimation that is not based on a statistical theory, the approximated variance of such estimator must be obtained using a bootstrap approach over the predicted errors.

This method for variance estimation was proposed by Efron (2007)and consists of generating many independent bootstrap samples drawn from the original sample with replacement. For

each bootstrap sample, the variance is calculated, and its values serve as a basis for inference (Lahiri, 2003).

Assume that an arbitrary sampling design without replacement is used to extract a probability sample s from a population U (Särndal et al., 1992). We are interested in estimating the variance of the estimator of the population parameter θ denoted as $V(\hat{\theta})$. The Bootstrap method works as follows:

1. Obtain an artificial population $U*$ from the sample data to simulate the population U.
2. Draw a bootstrap sample from $U*$ by following the same design to the one by which s was drawn from the original population U.
3. Independently repeat the previous step B times. For each bootstrap sample compute an estimate $\hat{\theta}_b^*$ $(b = 1, \ldots, B)$ following the same way $\hat{\theta}$ was computed, yielding estimators $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$.
4. Estimate $V(\hat{\theta})$ by:

$$\hat{V}_{BS} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2$$

where

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

3.4. MACHINE LEARNING ALGORITHMS

It should be noted that the aim of this study is not to compare the performance of multiple ML models but rather, compare different estimation approaches. Therefore, we choose a prediction model that has been linked with state-of-the-art results (Hanif, 2020) – XGBoost – and Classification Tree for the classification part.

3.4.1. Extreme Gradient Boosting (XGBoost)

As mentioned in section 2, Gradient Boosting is a type of ensemble models and, more particularly, the main objective of this algorithm is constructing base-learners to maximize the correlation between the negative gradient of the lost function associated with the set of all the models. Hence, the technique consists of adding new models to the ensemble in a sequential way where, at each iteration, a new weak model (base learner) is trained based on

the error of the set of all the models. This model proposed by T. Chen & Guestrin (2016) is a scalable machine learning system for tree boosting and was developed as an end-to-end system. One of the key features from this proposed system is the scalability which increases the running speed and ability to scale to a substantial number of examples. Scalability was possible due to algorithmic optimization including a weighted quantile sketch[5] with a theoretic foundation, which enables it to handle large datasets efficiently while still providing reasonably accurate estimates of quantiles, the capacity to deal with sparse data and a more efficient use of the memory.

3.4.1.1. Loss Function

The loss function measures quantitively the loss incurred from choosing an action from the set of all allowable actions - action space A -, given that the actual outcome is y with $y \in Y$ and, hence, the lower the loss, the better (Nielsen, 2016). The loss function is given by:

$$L : Y \times A \rightarrow \mathbb{R}_+$$

In predictive modeling, the loss function is used to measure the quality of a prediction by providing an indication of the divergence between the true outcome and our obtained prediction. The loss function we used was the squared error loss, given by:

$$L(y, \alpha) = \frac{1}{2}(y - \alpha)^2$$

3.4.1.2. Hyperparameters and Regularization Parameters

A few hyperparameters used by boosting trees, including XGBoost, are essential to optimize the model. Firstly, the number of iterations representing the basis functions is essential as the more iterations we have, the more complex the model will be since we have greater representational capacity. Another important hyperparameter, because of its contribution to control overfitting, is the learning rate, which shrinks the added basis function at each iteration. To achieve good results and avoid overfitting while keeping an excellent representational ability, one can simultaneously lower the learning rate (as increasing the learning rate is associated with a high variance) and increase the number of trees.

In terms of regularization parameters, the complexity of an additive tree is related to the number of trees as well as the complexity of each individual tree. To constrain its complexity, one can constrain the number of trees or restrict the maximum number of terminal nodes of each tree. It can be translated into two main parameters: maximum number of terminal nodes and minimum sum of observation weights in leaf.

[5] The weighted quantile sketching algorithm creates a quantile summary for each partition, or block, within the training set, permitting more efficient estimation of approximate percentiles of each feature, regardless of the dataset size (Grislain & Gonzalvez, 2021).

When tunning the maximum number of terminal nodes, one should consider the trade-off between variance and bias, as the higher the number of nodes, the fewer observations will be in each region, increasing the variance. Conversely, the lower the number of nodes is, the lower the capacity to capture an adequate order of interactions. As for the second parameter and considering that a leaf weight must be estimated for each region, we are interested in obtaining regions with high observation weights, as they allow for a more reliable estimation of the regional leaf weight.

One can introduce randomness in the learning process to improve the model's generalization performance. In practical terms, this is achieved by applying randomization parameters such as row subsampling, consisting of subsampling at each boosting iteration to fit a random subsample of the data. Another way to achieve this is using column subsampling, which means the tree will randomly sample the predictors.

3.4.2. Classification Tree

A decision tree is a recursive partition of the input space. It comprises a root node with no incoming edges and multiple nodes with one incoming edge, some of which have outgoing edges (“Data Mining and Knowledge Discovery Handbook,” 2010). The terminal nodes assign class labels to regions believed to represent the most suitable output value. Typically, the term node is restricted to the internal nodes, and the terminal nodes are named leaves.

The process of creating a tree involves gradually adding new nodes or leaves, which is referred to as growing. Due to its recursive nature, the decision tree growing algorithm comprises recursive procedures that take the entire training set to formulate the first root nodes and then call themselves for the subsections obtained using splits. The Decision tree Growing Algorithm can be found in (Cichosz, 2015a).

3.4.2.1. Hyperparameters and Regularization Parameters

The complexity also plays a vital role in the performance of the tree, which can be regulated by applying a pruning method and using stop criteria. One of the most used stop criteria is based on the maximum tree depth, which prevents further splitting after forming a sufficiently long path. Other stop criteria can be referred to as "no instances left," which happens when the set of training data assigned to a node is empty, and, therefore, performing splits will not improve the results.

4. RESEARCH METHODOLOGY

The principal objective of this work is to compare classical theories for finite population inference to more advanced techniques that make use of ML algorithms. For this purpose, we use data from three different funds (later aggregated into two funds) and two different auditing years, as well as metrics to evaluate the performance of the previously mentioned methods.

4.1. EXPERIMENTAL DATA

The datasets used to assess the performance of the diverse methods were originally obtained from IGF. The datasets comprise data regarding the operations amount and information for each sample and each population and the error-rate related data for each sample. The years under analysis are 2017/2018 and 2018/2019 for the funds ERDF, ESF and CF. Details of the datasets can be found in Appendix Table 1. The datasets were then merged into 1 dataset for ESF (n=132) and other for ERDF_CF (n =279).

4.1.1. Data Exploration

Data exploration is vital to understand which transformations are needed and how is the distribution of the error rate and features. We can observe that the proportion between operations with and without error for both funds is not evenly balanced. However, we decided not to apply Imbalanced Data techniques as we considered it to be only marginally imbalanced (He & Ma, 2013), as the proportions are 62.1%/37.9% for ESF and 22.2%/ 77.8% for ERDF_CF.

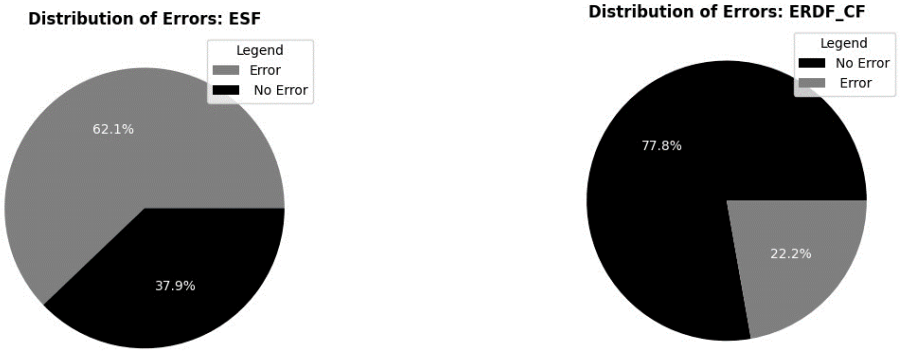


Figure 4.1 - Distribution of Errors

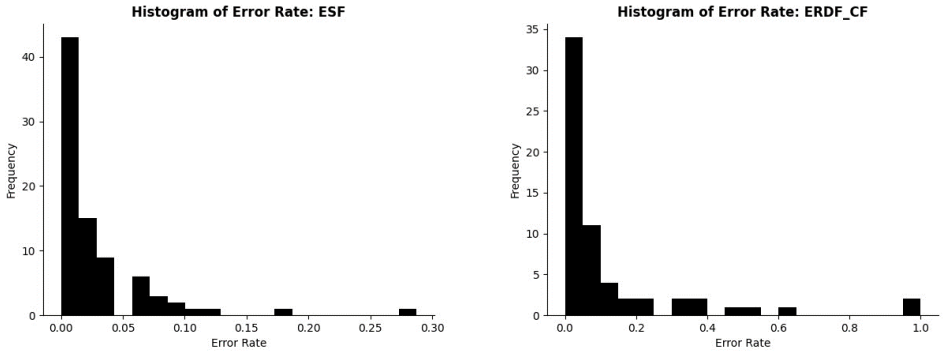


Figure 4.2 - Histogram of Error Rate

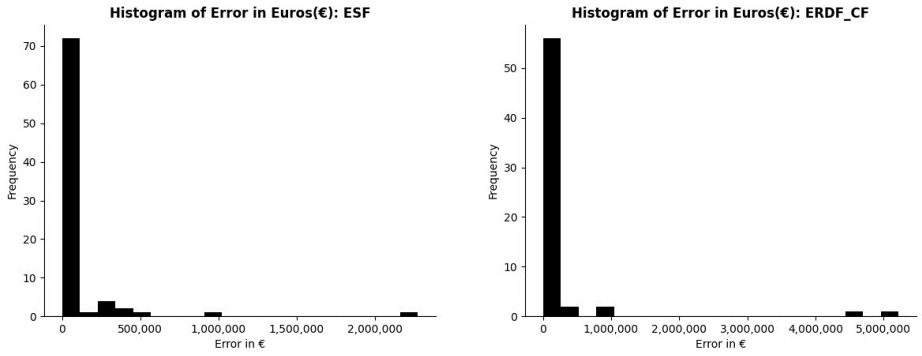


Figure 4.3 - Histogram of Error in Euros

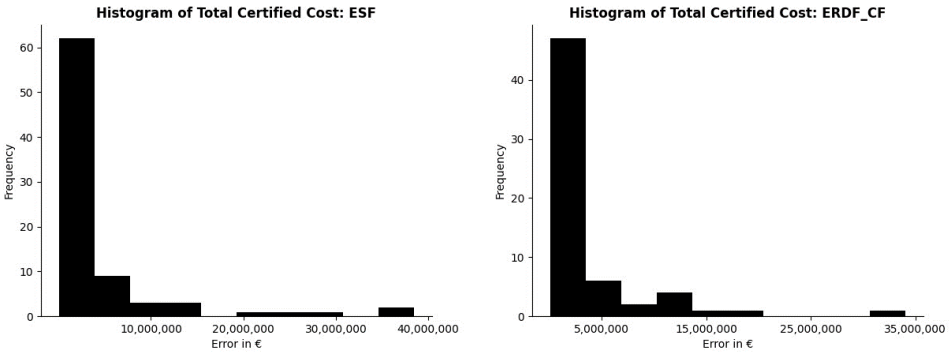


Figure 4.4 - Histogram of Total Certified Cost

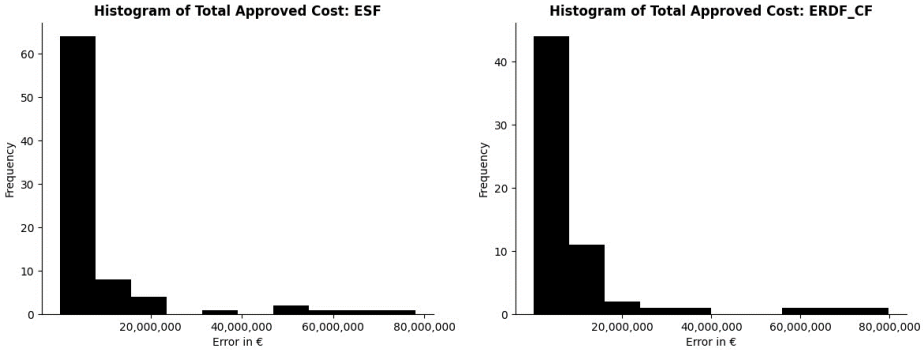


Figure 4.5 - Histogram of Total Approved Cost

By looking at the histograms, it is possible to detect some outliers and, thus, we decided to exclude 1 operation whose Error Rate was above 0.25, 1 operation with an Erro in euros above 2000000€ and 4 operations whose Total Certified Cost was above 35000000€ from ESF, and 2 operations whose Error Rate was above 0.8 and 4 operations whose Total Certified Cost was above 25000000€ from ERDF_CF, leaving ESF with 127 observations and ERDF_CF with 273.

By looking at the distribution of the numeric variables, it is possible to see that the data is right-skewed. To tackle this, one might resort to transformations such as applying the logarithmic transformation, however in this case we decided to use the original distribution of the data as these types of transformations are linked with an increased difficulty in the interpretability of results, which is a fundamental component, if not the most important, of this paper (Lee, 2020).

4.2. DATA CLEANING AND TRANSFORMATION

Data cleaning and transformation were fundamental, and we created new variables from the population dataset to extend the number of independent variables. Both the original and created variables can be found in Appendix Table 2. The main transformation was the aggregation of the funds ERDF and CF into a unique dataset [6] and creation of three variables Sample, Fund and Strata, for operations which have strata, to identify each operation. A new variable was created based on a significant characteristic of the operations, which is the value category they belong to: operations with a high audited expense are classified has “High Value” and take 1 in the variable and the remaining ones as 0. Other variables related to the accumulated amounts were created from the population's datasets. We created three variables related to previously audited operations – in previous periods of the same audit year and previous audit years- and previously audited operations' beneficiaries.

[6] These two funds were already merged in some of the data from the original files ,and therefore, we decided to follow the same logic due to the reduced size of the CF fund dataset (n=9 for 2017/2018 and n=28 for 2018/2019).

Finally, we created a new strata variable based on three variables: Sample, HV_TX (High Value), and Strata – for the samples and populations- and for each composed strata, we computed the Total Certified Cost, Public Expense and Total Approved Cost. Additionally, we aggregated the observations by beneficiaries and performed the exact computations from the ones using the created strata, resulting in new variables.

Another performed transformation was reducing the amount of Operation Typology Codes, which initially totaled 167. We created fourteen new Operation Typology Codes, as shown in the figure below, along with the number of original codes they enclose. This transformation was necessary since not all Operation Typology Codes of the population set were included in the sample.

There was no need to transform the data for the classification model as R can handle categorical features by himself. However, to use the *xgboost* package, we needed to transform the data into a matrix, as we will explain in further detail in section 4.7. Before transforming the data into a matrix, we applied one hot encoding to the categorical variables. One hot encoding represents categorical values as binary, where 1 is for presence and 0 for absence (Ashenden, 2021). We excluded one of the created columns for each category to avoid multicollinearity between variables.

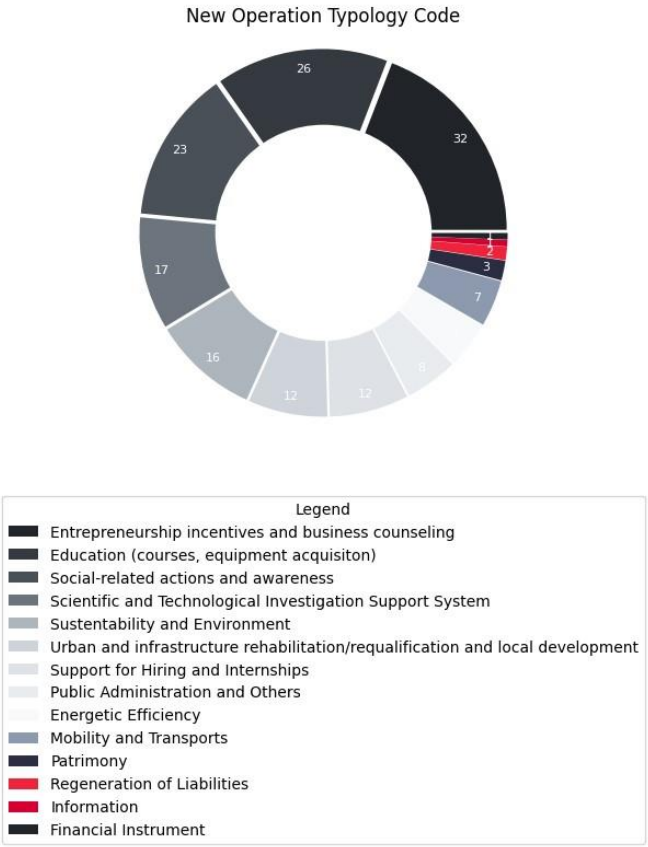


Figure 4.6 - Distribution and amount of original Operation Typology Codes aggregated into the new Operation Typology Code

4.3. EVALUATION METRICS

For the estimation, the level of materiality[7] defined by the EU is 2%. This indicator serves as a benchmark used to guarantee that the population does not contain any material error, and, therefore, the results are considered inconclusive if the precision overcomes this threshold.

Standard error estimation gives the accuracy level, but an estimate of the precision can also be obtained by multiplying the standard error by an appropriate quantile of the normal distribution and use this to construct a Confidence Interval. Given that the maximum level of confidence used in auditing for European funds is 90%, this coefficient is 1.645. It should be noted that we only compute the precision for the operations with a low value as the operations with a high value are entirely observed and, therefore, have no variability and should not be accounted for the design-based precision.

In the model-based approach, two models were used to test the quality of the adjustments of the ML models. The first method is the holdout method, which partitions the data randomly into a train and a test set with an 80% and 20% partition, respectively. The second method was k-fold cross-validation, which divides the dataset into k groups randomly. The process consists of using the test set for scoring after the model has been trained using the training set, which is formed by k-1 folds. This process is repeated for reduced variance until all folds have been employed for evaluating the model, and the final evaluation is determined by averaging the results across all test folds (Cichosz, 2015b). We used repeated cross-validation with ten folds and three repetitions for the classification and cross-validation with five folds and 100 repetitions for the prediction.

Regarding the model performance metrics for prediction, the three metrics we used, R^2, RMSE, and MAE, are summarized in the Appendix Table 13. RMSE and MAE were computed using the predicted and real values in a rate form (hence the results in decimals) while R^2 was obtained using the error expressed in euros to provide more realistic results of the explained variance. For classification we considered a few metrics that derived from the confusion matrix, which is constructed based on cases that were classified as positives (TP if they are indeed positives and FP otherwise), and cases that were classified as negatives (TN if they are actual negatives and FN otherwise). These metrics are also described in the appendix (Appendix Tables 14 and 15).

[7] In financial and compliance audit, materiality is a crucial concept that determines the degree of deviation that the auditor deems capable of influencing the decisions of the stakeholders. Errors or deviations are considered material if they can reasonably affect the underlying audit conclusions or the decisions of the audit report recipients, either individually or when combined with other errors (*Materiality*, 2021).

4.4. EXPERIMENTAL PROCEDURE

In this work, the design-based approach will serve as a benchmark for the approaches that use of a ML model.

For the model-related approaches, we will use two different prediction methods: 1-stage and 2-stage prediction. The first stage is more straightforward, training and testing with data containing errors and non-errors and simply using a prediction model (XGBoost). The 2-stage prediction will introduce an additional stage with a classification model. Since there is a considerable number of operations with no error rate, we decided first to classify whether the observations have errors and, only after, predict the error rate on the subset of data that was classified as having errors. For this approach, we created a training and test pipeline that initially trains a classification model with all the data from the sample and outputs the results with the values 0 and 1 for the observations classified as not having or having errors, respectively. For the second part of the pipeline, the training data was limited to the observations from the sample that had errors. Including the observations with no error in the training process could bias the outcome since it represents noise. The testing and metrics were obtained based on the prediction results of the observations classified as having errors in the previous stage.

In the subsections below we will cover how we will estimate the error for each inference type.

4.4.1. Design-based inference

The operations from ESF were initially divided into strata depending on their nature: public or non-public. We created a new stratum based on this original stratification, on the fund, on the value of the operations (high and low value), and the audited sample period. Due to the random selection of the samples, the estimates produce different values according to the population elements that were selected for the sample. Since the data follows a stratified sampling based on the MUS method, we applied the Horvitz-Thompson (HT) Estimator, in which the inclusion probability is proportional to the book value of the operations. It is noteworthy that in MUS, the population comprises a set of accounts representing a group of currency units, and the book value of each account corresponds to the number of euros, in this case, of that account (Thompson, 1997). In this context, the book value corresponds to the declared expense in euros of each operation, which the variable Total Certified Cost gives.

As stated in Theme: Weighting and Estimation-Main Module Contents (2014), let n be the sample size, BV the book value of operation i, and BV the total declared expense, the inclusion probability is given by:

$$\pi_i = n \cdot \frac{BV_i}{BV}$$

And let y_i be the value of audited error of the *i* operation, the error extrapolation based of the HT estimation is given by:

$$\hat{\tau} = \sum_{i \in s} w_i y_i ,$$

where $w_i = \frac{1}{\pi_i}$, that is, the inverse of inclusion probability.

The projected error can be obtained in monetary units or rate, equal to the projected error in euros divided by the Book Value of the population. The MUS estimator's precision is a function its expected value and variance.

4.4.2. Model-based

We will use the model XGBoost to obtain the model-based estimates in opposition to the previous approach, which does not use any model. All the sample datasets are used to train the model, and since the model can output negative values, we truncate the results to zero. To predict the error of the operations in the universe, we use the estimated model with the training data and the bootstrap technique to estimate the variance, as explained in section 3.3, with $B = 1000$. Finally, the projected error is predicted using the estimated model on the data from the universe.

4.4.3. Model-assisted

This approach combines the errors from the sample and the population to compute the total population error.

Let y_i be the value of the audited error of the *i* operation, the error extrapolation is given by:

$$\hat{\tau} = \sum_{i \in s} \hat{y}_i + \sum_{i \in s} w_i \, (y_{i-} \hat{y}_i)$$

Firstly, we compute the difference between the audited error and the model-based estimate, project the model's prediction error to the population using the weights, and then sum the estimates, also referred to as the model-assisted estimate. Finally, we compute the prediction error variance following the design-based methodology.

4.5. FEATURE SELECTION

The goal of feature selection is to reduce the number of independent variables by selecting the most prominent features that will be input to the algorithms. Even though we are not

dealing with high-dimensional data, feature selection is still important as it decreases computing costs, minimizes overfitting, and enhances learning performance by taking out features that contain a significant amount of random or irrelevant information, otherwise known as noisy features, or redundant features from the dataset (Li et al., 2017).

In this research, we used different methods for feature selection. In the 1-stage prediction approach, we used a wrapper method called Recursive Feature Selection proposed by Guyon *et al*.(2002) which consists of training the classifier or, in our case, regression model, computing the ranking criterion for all the response variables, and removing the ones with the minor ranking criterion. We also used the extreme gradient boosting model to estimate the feature importance. In the 2-stage prediction approach, we used a combination of xgboost importance, random forest importance, and a wrapper method using Boruta for the classification stage. We re-used the features selected in the first approach for the prediction stage.

The latter stated method is a wrapper method using the Boruta package from R that follows the same idea behind RF of adding randomness and using an ensemble of samples drawn at random to produce a more reliable result indifferent to the presence of correlations or meaningless oscillations. In the iterative process, "shadow" attributes (generated by rearranging the values of the initial attribute among different objects) are added to the information system, and their relevance is determined by comparing them to a maximum Z score calculated from the data. Attributes surpassing the score are considered necessary, while those falling significantly below are labeled 'unimportant' and removed. The process continues until all attributes are assigned importance or a preset limit of iterations is reached to optimize feature selection for the dataset.

For the 1-stage prediction, we used RFE and obtained the optimal number of features to be 5 for both ERDF_CF and ESF. For ERDF_CF, the selected features were Total Certified Cost, Total Approved Cost, Operation Typology Code 1, Previous Beneficiary False, Program “Norte” and Year 2017/18, and for ESF were Total Approved Cost, Total Certified Cost, Program “Centro”, Program “POCI” and Operation Typology Code 1.

By looking at the importance given by the XGBoost model (Appendix B), it is clear that the numeric variables have a greater importance, with the most significant gain coming from the Total Certified Cost for the first set of data and the Total Approved Cost for the second.

Since the selected features from both methods differed slightly, we created four sets of variables, two for each fund, ran the model for each set, and chose the best one based on metrics as we will explain in section 5.1.1. The resultant set of variables can be found in table 1 below.

Table 1 - Final Selected Variables for 1-stage prediction

Var	ERDF_CF		ESF	
	Set 1	Set 2	Set 1	Set 2
Operation Typology Code 1	x	x		
Operation Typology Code 4			x	
Program “Centro”		x		
Program “POCI”		x		
Program “POCH”			x	x
Program “Norte”	x			
Total Approved Cost	x	x	x	x
Total Certified Cost	x	x	x	x
Previous Beneficiary False	x			
Year 2017/18	x			
Previously period Audited False			x	
Previously Audited False			x	x
Strata NPP				x
Total	6	5	6	5

For the 2-stage prediction, we only used feature selection methods for the first stage of the process (classification). Firstly, RFE selected an optimal number of 17 variables for ERDF_CF and 6 for ESF: Program "Lisboa", Total Certified Cost, Total Approved Cost, Program "POCI", Previously Audited False, Operation Typology Code 5, the Year 2017/18, Previous Beneficiary False, Program "Norte", Previously period Audited False, Operation Typology Code 1, Program "Centro", Program "M1420", Program "Açores", Operation Typology Code 6, Operation Typology Code 2 and Program "POSEUR" for the first population, and Program "M1420", Total Approved Cost, Operation Typology Code 3, Program "Lisboa", the Year 2017/18, Program "POCH" for the second. Contrary to the classification, there is confirmation that Total Certified Cost is irrelevant in all populations.

The random forest importance produced two different metrics Mean Decrease Accuracy and Mean Decrease Gini, as it can be observed in the Appendix B. Finally, Boruta identified the variables that were considered important, unimportant, or tentative attributes that were left, which – also represented in the Appendix.

In the 2-stage approach, the selected variables across the different methods were still not consistent, and hence, we created two sets of variables for each fund as well, which can be found in table 2 below.

Table 2 - Final Selected Variables for 2-stage prediction for classification stage

Var	ERDF_CF		ESF	
	Set 1	Set 2	Set 1	Set 2
Program “M1420"	x		x	x
Program “POCH"				
Program “Lisboa"	x	x	x	
Program “POCI"	x	x		
Program “POCH”			x	x
Program “Norte”	x			
Program “Centro”	x			
Program “Açores”	x			
Program “POSEUR”	x			
Strata NPP				x
Total Approved Cost	x	x	x	x
Total Certified Cost	x	x		
Year 2017/18	x		x	x
Previously Audited False	x	x		
Previous Beneficiary False	x			
Previously period Audited False	x			
Operation Typology Code 1	x			
Operation Typology Code 2	x			
Operation Typology Code 3			x	x
Operation Typology Code 5	x			
Operation Typology Code 6	x			
Total	17	5	6	6

4.6. HYPERPARAMETER AND PARAMETER TUNNING

The parameter configuration of an algorithm plays a crucial role on its performance. Hence, we used Grid Search with cross-validation to optimize the parameters and hyperparameters, whose values cannot be predicted from the data. This method tries all combinations of parameters and hyperparameters to identify the best ones to achieve the highest accuracy (Liashchynskyi & Liashchynskyi, 2019). As referred to in section 3.4.1.2, we used “Eta” which represents the learning rate hyperparameter, two randomization parameters, them being row subsampling and column subsampling here represented by “Subsample” and “Colsample by

tree", respectively, and "MaxDepth" and "Min child weight" which are the complexity parameters. The results of the grid search show that the optimal values were consistent across sets of data for the "Eta" as all of the sets registered a learning rate of 0.1 (which was the highest one between the possible parameter values) and "Min child weight" to 1, which means the algorithm will allow each leaf node in the tree to have at least one instance – making it less conservative in the way it learns from the data. From the table, we can also observe that better results were obtained with a tree with a low maximum depth and that the values of the randomization parameters differed from set to set.

Using the parameters defined in the grid search, we obtained the optimal number of iterations based on the metric results from the repeated cross-validation – we used a weighted metric with 60% for the RMSE and 40% for MAE – as we wanted to give more importance to the former. With this said, the optimal number of iterations obtained were 52 and 59 for the final selected dataset of ERDF_CF and ESF, respectively, in the 1-stage prediction and 40 and 61 for the homologs for the 2-stage prediction.

Table 3 - Chosen Parameters for 1-stage prediction

(Hyper) Parameter	**ERDF_CF**		**ESF**	
	Set 1	**Set 2**	**Set 1**	**Set 2**
Eta	0.1	0.1	0.1	0.1
Max depth	3	3	3	4
Subsample	1	1	0.6	0.6
Colsample by tree	0.6	0.5	0.5	0.5
Min child weight	1	1	1	1

For the classification part, we found the optimal values of three stop criteria parameters: "Max depth", which sets the maximum depth of each node of the final tree, and two parameters to verify the "no instances left" criteria mentioned in section 3.4.2.1. , them being the "Min split" and "Min bucket". The former controls the minimum number of instances required for a split, while the latter defines the minimum number of observations in any terminal node. Here, on the other end, the best results were obtained with higher values for the maximum depth of the tree, a behavior also seen for the "Min split" parameter.

Table 4 - Chosen Parameters for 2-stage prediction for classification stage

(Hyper) Parameter	ERDF_CF		ESF	
	Set 1	**Set 2**	**Set 1**	**Set 2**
Max depth	8	15	15	8
Min split	16	15	15	15
Min bucket	4	4	2	2

The tuning process for the second stage of the 2-stage prediction consisted of tuning the same hyperparameters and parameters for the prediction mentioned above, only now with a different training set. For the second stage of this approach, the obtained optimal parameters were similar to the previous approach, only differing in the "MaxDepth" parameter, which had a value of 6.

Table 5 - Chosen parameters for 2-stage prediction for prediction stage

Var	ERDF_CF	ESF
Eta	0.1	0.1
Max depth	6	6
Subsample	0.5	0.5
Colsample by tree	0.5	0.6
Min child weight	1	1

4.7. SOFTWARE IMPLEMENTATION

The data transformation process and implementation of the experimental procedure were based on R programming language, using the *dplyr* library for data manipulation. To deal with categorical variables, we use an appropriate data structure in R which stores this type of data by calling the function factor and to feed these data into the ML model we used a class called *dummyVars* to create a full set of dummy variables based on our categorical variables.

We created a function for the model-based approach that implements a single-stage modeling approach for irregularity prediction. The function takes several input parameters, including training and test data, model parameters, and the number of boosting rounds (*nrounds*). The function involves a single stage of model building and prediction using the XGBoost algorithm, given by the *XGBoost* package for R. In each loop iteration (repeated 1000 times), a random sample is drawn from the training data. The predictor and response variables are then defined

in the training set. A data matrix and a matrix for the test data are created using the XGBoost-specific data structure (*xgb.DMatrix*). The XGBoost model is then trained using the specified parameters and the training data. After training the model, predictions are made on the test data. The irregularity values are computed using the XGBoost predictions. If the prediction is less than or equal to zero, the irregularity value is set to zero; otherwise, it is calculated by multiplying the XGBoost prediction by the corresponding Total Certified Cost value from the original scale of the test data. Finally, the function returns the vector of total irregularity values.

The function of the other prediction approach is similar but adapted to the two-stage modeling approach, using the *rpart* package for modelling Regression or, in this case, Classification trees. The package follows the principles established by Breiman (1996). Essentially, the package creates a tree, which is then split into different branches by features, and the outcome derives from predicting the most frequent outcomes on each split. It differs from a Random Forest algorithm because we only build one single tree. In contrast, in a Random Forest algorithm, the outcome comes from combining of the results of multiple trees.

Different functions and packages were used for feature selection, namely *rfe* (an algorithm for backwards feature selection), *xgb.importance* and *Boruta* package.

5. RESULTS AND DISCUSSION

5.1. 1-STAGE PREDICTION

5.1.1. Metrics

Table 6 – Test Evaluation Metric Results

Parameters	R squared	RMSE	MAE
ERDF_CF set 1	0.356	0.052	0.031
ERDF_CF set 2	0.274	0.054	0.029
ESF set 1	0.801	0.013	0.011
ESF set 2	0.569	0.023	0.016

The selection of which set to move forward was solely based on the evaluation metrics referred in section 4.3. We computed the coefficients of determination (R^2 values) to understand the proportion of variance in the dependent variable that the independent variables can explain as well as RMSE and MAE. Firstly, looking at the RMSE and MAE and comparing them with the train and test results – which were lower than 0.05 for both metrics in every set but for the training results of ERDF_CF - it is possible to see that we obtained good results. The train and test results are presented visually in the Appendix C. In terms of R^2, the independent variables explain more than 70% of the variance for set 1 of ESF, but the ERDF_CF registered lower determination coefficients, with the best set of features with a determination coefficient of 27.40%. Based on these metrics, we chose set 1 for ERDF_CF and set 1 for ESF (both with six variables) as the final set of selected features as they could predict more variance of the independent variable.

5.1.2. Inference Results

Table 7 - Design-based estimation results

	ERDF_CF	ESF
Book Value	4396804087	2602688095
Projected Error	93081110	35582146
Projected Error Rate %	2.117	1.367
SD	15401417	5384896
SD %	0.350	0.207

In the benchmark approach, we obtained a level of materiality - given by the projected error rate- below the established 2% for the set of data of ESF but exceeded the threshold by 5.68% (difference of 0.117) for the set of data of ERDF_CF. The results indicate that the projected error was quite precise since the standard error for both data sets is below 0.4%.

It is worth mentioning that the benchmark approach's results are equal across prediction approaches since we are not using an ML model here.

Table 8 - Model-based estimation results for 1-stage prediction

	ERDF_CF	**ESF**
Projected Error	135733651	17863932
Projected Error Rate %	3.087	0.686
SD	29434859	3337554
SD %	0.669	0.128

Table 9 - Model-assisted estimation results for 1-stage prediction

	ERDF_CF	**ESF**
Projected Error	121979576	12981118
Projected Error Rate %	2.774	0.499
SD	14739107	2581782
SD %	0.335	0.099

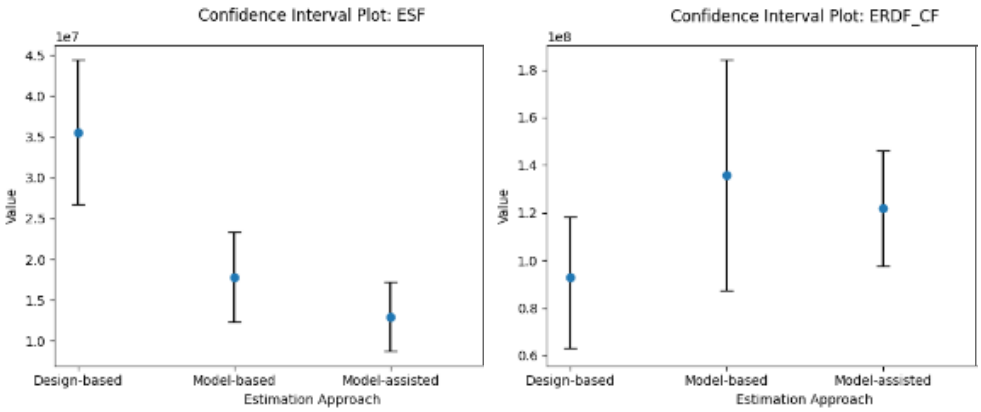
The projected error rate of each respective fund was more homogenous across the remaining approaches, with an average value of 2.93% for the ERDF_CF set and an average of 0.59% for the ESF, which is a considerable rate decrease compared to the design-based approach one. The highest projected error rate was registered in the 1-stage prediction model-based approach, which surpassed a value of 3.00% by 2.86% (difference of 0.087). The model-assisted estimation registered the lowest projected error rate for ESF, with 0.499%, but the benchmark approach produced the lowest value for ERDF_CF. The Confidence Intervals of the ESF fund sustain these findings as the upper limits are all within the threshold (none of the populations registered an upper IC limit above 1.7%, which is an indication of inexistence of material error). However, in the ERDF_CF fund, while the upper limits of the CI surpass the 2% threshold, for the DB and MB approaches, the lower limit is inferior to the threshold, subsisting a doubt concerning the result's conclusiveness.

Looking at the precision levels, given by the standard error of the prediction, a similar behavior is detectable, with better precision in the approaches that introduce a model to the estimation in complement to the sampling method. The model-assisted estimation produced the results with the best precision, with both funds having a standard error below 0.34%. The model-based approach registered the highest standard error for the ERDF_CF among all the estimation methods (0.669%). The design-based approach had the second-highest value for ERDF_CF (0.350%) and the highest value for ESF (0.207%), even though the difference from the other estimations is not substantive.

Table 10 - Confidence Intervals for 1-stage prediction

		Mid	Lower	Upper	% Lower	% upper
ERDF_CF	**DB**	93081110	63125353	118416442	1.436	2.693
	MB	135733651	87313308	184153993	1.986	4.188
	MA	121979576	97733746	146225406	2.223	3.326
ESF	**DB**	35582146	26723991	44440300	1.027	1.707
	MB	17863932	12373655	23354208	0.475	0.897
	MA	12981118	8734086	17228150	0.336	0.662

Figure 5.1 - Confidence Intervals Plots for 1-stage prediction



5.2. 2-STAGE PREDICTION

5.2.1. Metrics

Table 11 - Train and Test Classification Model Performance Metrics

Parameters	Accuracy Train	Accuracy Test
ERDF_CF set 1	0.772	0.764
ERDF_CF set 2	0.776	0.855
ESF set 1	0.552	0.750
ESF set 2	0.637	0.583

Table 12 - Confusion Matrix Results for Classification model ERDF_CF

ERDF_CF set 1		
	Reference	
Prediction	0	1
0	39	9
1	4	3

ERDF_CF set 2		
	Reference	
Prediction	0	1
0	40	5
1	3	7

Table 13 - Confusion Matrix Results for Classification model ESF

ESF set 1		
	Reference	
Prediction	0	1
0	4	1
1	5	14

ESF set 2		
	Reference	
Prediction	0	1
0	3	4
1	6	11

Table 14 - Prediction Model Performance Metrics

Parameters	R squared	RMSE	MAE
ERDF_CF	0.328	0.192	0.159
ESF	0.610	0.020	0.015

The 2-stage prediction approach produced good metrics for the classification stage but slightly worse determination coefficients for the prediction stage (around 32.8% and 61.0% for the two data sets). Here, the selection of the final set of variables classification-wise was also achieved by analyzing the metrics, which led to the selection of the second set of variables for ERDF_CF and set 1 for ESF (with 5 and 6 variables, respectively).

5.2.2. Inference Results

Table 15 - Model-based estimation results for 2-stage prediction

	ERDF_CF	**ESF**
Projected Error	122225156	14597498
Projected Error Rate %	2.780	0.561
SD	313717711	3080391
SD %	7.135	0.118

Table 16 - Model-assisted estimation results for 2-stage prediction

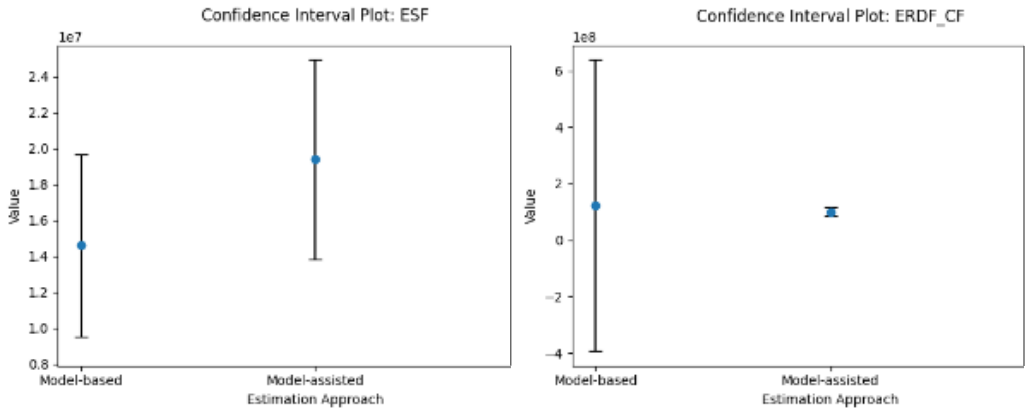
	ERDF_CF	**ESF**
Projected Error	100594511	19395831
Projected Error Rate %	2.288	0.745
SD	9365544	3370079
SD %	0.213	0.129

The projected error rate for the model-based and model-assisted for the ERDF_CF decreased, surpassing the threshold only by 8.15%. The rate of the ESF was already significantly below the threshold and remained so with an average of 0.653% over both approaches. In terms of precision, this prediction approach resulted in better overall results, with reduced standard errors for the ESF of both estimations for the ERDF_CF of the model-assisted approach. However, the model-based approach resulted in an extremely high value for the ERDF_CF, making us question the reliability of the generated error. Accordingly, the IC results above, the IC upper limits from the ESF are also enclosed in the 2.00% threshold, a behavior that cannot be found in any of ERDF_CF upper IC limits (14.52% and 2.64% for the MB and MA), reinforcing once again the uncertainty linked to the obtained results for this population.

Table 17 - Confidence Intervals for 2-stage prediction

		Mid	Lower	Upper	% Lower	% upper
ERDF_CF	**MB**	122225156	-393840479	638290790	-8.957	14.517
	MA	100594511	85188191	116000831	1.938	2.638
ESF	**MB**	14597498	9530256	19664741	0.366	0.756
	MA	19395831	13852050	24939612	0.532	0.958

Figure 5.2 - Confidence Intervals Plots for 2-stage prediction



5.3. DISCUSSION

Given the difficulties involved in the traditional methods of estimation, which require auditing, the main goal of this study was to explore data science-related methods for estimation by using ML models in replacement or complementary to the design-based approach.

From a results' conclusiveness view, there is an indication that, regarding the 1 stage prediction approach, the two estimation methods under comparison can meaningfully contribute as they produce more conclusive results for ESF, even though no estimation approach was able to improve the conclusiveness of the set of data that surpassed the threshold of materiality in the benchmark approach (ERDF_CF). The results of the projected error rates from these new approaches were also better overall compared to the benchmark approach, especially for the model-assisted.

By using the 2-stage prediction approach, there is no evidence that the results had significant improvement in conclusiveness as the number of conclusive results is the same across prediction approaches (only ESF produces conclusive results). However, the ERDF_CF set

became closer to the 2% threshold in this prediction approach, meaning that the results are more reliable even if inconclusive.

Regarding the precision of the prediction, the model-assisted had better precision levels for the ERDF_CF in this approach. On the other hand, the model-based estimation had a gain of precision only in the set of data related to ESF but a substantial worsening for ERDF_CF.

By comparing the approaches regarding the method of prediction, we can see that the gain obtained by the 2-stage prediction is not significant since the increase in the reliability of results in the ERDF_CF is not sustained by the obtained level of precision, particularly in the ERDF_CF. The obtained level of precision allied to the cost of implementation, and taking into consideration that by using this approach, the training data for the prediction stage is minor (we use only the observations that have an actual error), we can conclude that it is not viable nor beneficial to opt for this approach.

However, even if the 2-stage prediction pipeline was not beneficial, we must recognize the positive metrics from the classification stage. Especially the metrics of the ERDF_CF set, which, contrary to the prediction stage, have even better metrics than the ESF fund. This is a clear advantage of classification against prediction as it does not require many operations with error, which is a typical pattern in the sample sets of the operations under the scope of this work. It is an indicator that the traditional estimation could be complemented differently. Once we draw a sample, a classification model can run on the sample data, and only the operations classified with a positive class (meaning that we are upon the presence of error) could be considered in the auditing process, which would also lead to a reduction of the efforts. Essentially, a model would be introduced but not for a prediction task.

From the observation of the different sets of data, we can conclude that, regardless of the prediction approach, the ESF fund produced better results, both in terms of conclusiveness of results and precision levels. This finding proves that the data is imperative for the estimation's success. It leads to an important conclusion: introducing a model to the estimation can improve the results' conclusiveness and precision if there are a significant number of errors from which the model can learn. This could also explain why the model could have been more effective in the other data set, as the error presence in the training set is not predominant.

Overall, the results confirm that the opportunity to use alternative approaches to the design-based one should be more deeply explored, especially considering the high sample sizes that have been used for auditing. Moreover, the results show that despite the efforts linked with the design-based approach, it can still lead to inconclusive results, which means that this approach alone is not presently more beneficial than the other methods.

In fact, by comparing the pure model-based estimation with the model-assisted one, there is a clear indication that the hybrid approach leads to better results, meaning that data science can contribute significantly to the error estimation of operations. However, it still benefits from the consideration of the sampling plan.

6. CONCLUSIONS

This study aimed to study the viability of using inference approaches that use ML models, integrally or partially. Hence, two estimation methods were applied to compare with the base design-based approach: (pure) model-based, and model-assisted. The results of this comparison provide valuable insights that regulators and the EU can use to understand the value of integrating data science in the current method of auditing operations.

The second research goal was to provide a comparison between the methods. One crucial finding was that the approach that uses both model and sampling methods performed better than the (pure) model-based approach, stating the importance of complementing the traditional auditing methods with ML models but not resorting exclusively to this type of estimation. In particular, these alternative approach under examination provided more conclusiveness results than the benchmark one. The findings related to using a classification model in the model-based approach before the error prediction proved to significantly improve the reliability of results but needed to produce precise results, which weakens this approach. Therefore, we concluded that no value was added by adding complexity to the prediction pipeline, which also increased the estimation costs.

We must highlight some limitations in our research. Firstly, the number of variables available was limited and restricted to the financial state of the operation. Given the nature of the operations, it is essential to consider another type of variables, namely those related to the characteristics of the beneficiary. Secondly, since we are using a sample to train a model, which will be tested on the population data, it is natural that the proportion we used for training and testing was contrary to most of the suggested proportions. Future research may include data from an extended number of periods, which would benefit the model's training because more data means the model would be trained with more examples, particularly more observations with an actual error rate. The data used for this work contained predominantly observations with no error rate (specially for ERDF_CF), which might have harmed the results. Moreover, it would also ensure the diversity of the code typologies in the sample, and, in this way, there will not be a need to aggregate the operations, which could also be a factor that biased the importance of the variable Operation Typology Code.

Even considering the limitations regarding the constraint of available data, we could estimate models with a significant prediction capability. This demonstrates the potential of using alternative methods to the design-based approach, and with improved data quantity and quality, feature coverage can only improve.

To conclude, data science techniques can be extended into the important process of operations' error estimation domain, which has been using outdated techniques with significant costs for the stakeholders. With this said, the contribution of this work may represent a future step in how EU conducts error estimation, using modern techniques to relieve auditing-related efforts and improve fund allocation.

7. BIBLIOGRAPHIC REFERENCES

Ali, A. Al, Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences*, *13*(2272).

Ashenden, S. K. (2021). The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry. In *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry* (pp. 15–26). https://doi.org/10.1016/B978-0-12-820045-2.09990-6

Ashtiani, M. N., & Raahemi, B. (2022). Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review. In *IEEE Access* (Vol. 10). https://doi.org/10.1109/ACCESS.2021.3096799

Baker, R. L., & Copeland, R. M. (1979). Evaluation of the Stratified Regression Estimator for Auditing Accounting Populations. *Journal of Accounting Research*, *17*(2). https://doi.org/10.2307/2490521

Beck, P. J. (1980). A Critical Analysis of the Regression Estimator in Audit Sampling. *Journal of Accounting Research*, *18*(1), 16–37.

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, *26*(2). https://doi.org/10.1007/s11142-020-09563-8

Bowley, A. L. (1906). Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science, York, 1906. *Journal of the Royal Statistical Society*, *69*(3). https://doi.org/10.2307/2339344

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2). https://doi.org/10.1007/bf00058655

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, *63*(3). https://doi.org/10.1093/biomet/63.3.615

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

Chen, Y., Wu, Z., & Yan, H. (2022). A Full Population Auditing Method Based on Machine Learning. *Sustainability*.

Cichosz, P. (2015a). Data Mining Algorithms: Explained Using R. In *Journal of Statistical Software* (Vol. 66, Issue August).

Cichosz, P. (2015b). Data Mining Algorithms: Explained Using R. In *Journal of Statistical Software* (Vol. 66, Issue August, pp. 71–77).

Cochran, W. (1977). *Sampling Techniques* (3rd ed. , pp. 8–373).

Data Mining and Knowledge Discovery Handbook. (2010). In *Data Mining and Knowledge Discovery Handbook* (pp. 149–174). https://doi.org/10.1007/978-0-387-09823-4

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12). https://doi.org/10.1145/2500499

Efron, B. (2007). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1). https://doi.org/10.1214/aos/1176344552

European Comission. (2015, December 14). *2014-2020 European structural and investment funds*.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604). https://doi.org/10.1098/rsta.1922.0009

Godambe, V. P. (1955). A Unified Theory of Sampling from Finite Populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(2). https://doi.org/10.1111/j.2517-6161.1955.tb00203.x

Godambe, V. P. (1965). A Review of the Contributions Towards a Unified Theory of Sampling from Finite Populations. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *33*(2). https://doi.org/10.2307/1402030

Grislain, N., & Gonzalvez, J. (2021). *DP-XGBoost: Private Machine Learning at Scale*. http://arxiv.org/abs/2110.12770

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3). https://doi.org/10.1023/A:1012487302797

Hanif, I. (2020). *Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction*. https://doi.org/10.4108/eai.2-8-2019.2290338

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*(384). https://doi.org/10.1080/01621459.1983.10477018

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. In *Journal of Educational and Behavioral Statistics* (Vol. 44, Issue 3). https://doi.org/10.3102/1076998619832248

He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, algorithms, and applications. In *Imbalanced Learning: Foundations, Algorithms, and Applications* (pp. 13–41). https://doi.org/10.1002/9781118646106

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, *47*(260). https://doi.org/10.1080/01621459.1952.10483446

Huang, F., No, W. G., Vasarhelyi, M., & Yan, Z. (2022). Audit Data Analytics, Machine Learning, and Full Population Testing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4033165

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. In *Science* (Vol. 349, Issue 6245). https://doi.org/10.1126/science.aaa8415

Kaplan, R. S. (1975). Sample Size Computations for Dollar-Unit Sampling. *Journal of Accounting Research*, *13*. https://doi.org/10.2307/2490487

Kotu, V., & Deshpande, B. (2015). Predictive Analytics and Data Mining. In *Predictive Analytics and Data Mining*. Elsevier. https://doi.org/10.1016/c2014-0-00329-2

Lahiri, P. (2003). On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation. In *Statistical Science* (Vol. 18, Issue 2). https://doi.org/10.1214/ss/1063994975

Lee, D. K. (2020). Data transformation: A focus on the interpretation. *Korean Journal of Anesthesiology*, *73*(6). https://doi.org/10.4097/kja.20137

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Issue 6). https://doi.org/10.1145/3136625

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. http://arxiv.org/abs/1912.06059

Lohr, S. L. (2019). Sampling: Design and Analysis: Design And Analysis. In *CRC Press (Taylor & Francis Group)*.

Mabelane, K., Mongwe, W. T., Mbuvha, R., & Marwala, T. (2023). An Analysis of Local Government Financial Statement Audit Outcomes in a Developing Economy Using Machine Learning. *Sustainability*, *15*(12).

Materiality. (2021). Https://Methodology.Eca.Europa.Eu/Aware/GAP/Pages/CA-FA/Planning/Materiality.Aspx.

Mitchell, T. M. (1997). *Machine Learning: A multistrategy approach* .

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of Machine Learning (Adaptive Computation and Machine Learning series). In *The MIT Press* (Vol. 17, Issue 4).

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning, second edition - Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar - Google Books. In *The MIT Press: Vol. №3* (Issue 1).

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2). https://doi.org/10.1257/jep.31.2.87

Neter, J., & Loebbecke, J. K. (1975). *Behavior of Major Statistical Estimators in Sampling*. American Institute of Certified Public Accountants Keywords.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, *97*(4). https://doi.org/10.2307/2342192

Nielsen, D. (2016). Tree Boosting With XGBoost Why Does XGBoost Win "Every" Machine Learning Competition? *Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?*, *December*.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing*, *30*(2). https://doi.org/10.2308/ajpt-50009

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. *Accounting Review*, *92*(2). https://doi.org/10.2308/accr-51562

POAT. (2021). *Carta de Missão e Valores do Programa Operacional de Assistência Técnica*.

Roberts, D. M. (1978). *Statistical auditing*. https://egrove.olemiss.edu/aicpa_guides

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer New York, NY.

Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D. V., Barndorff-Nielsen, O., & Dalenius, T. (1978). Design-Based and Model-Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, *5*(1), 27–52.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-STS330

Smith, T. M. F. (1976). The Foundations of Survey Sampling: A Review. *Journal of the Royal Statistical Society. Series A (General)*, *139*(2). https://doi.org/10.2307/2345174

Song, X. P., Hu, Z. H., Du, J. G., & Sheng, Z. H. (2014). Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. *Journal of Forecasting*, *33*(8). https://doi.org/10.1002/for.2294

Theme: Weighting and Estimation-Main Module Contents. (2014).

Thompson, M. E. (1997). *Theory of Sample Surveys* (1st ed.). Chapman and Hall/CRC.

Zhang, C. (Abigail), Cho, S., & Vasarhelyi, M. (2021). Explainable Artificial Intelligence (XAI) in Auditing: A Framework and Research Needs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3981918

Zhang, C. (Abigail), Vasarhelyi, M., & Cho, S. (2021). Identifying Informative Audit Quality Indicators (IAQI) Using Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3981622

APPENDIX A. METADATA

Appendix Table 1 - Metadata of the Original Data

Dataset	Number of observations	Number of attributes	Year	Fund
Sample_1o_per	81	14	2017/2018	ERDF **and CF**
Sample_2o_per	78	14	2017/2018	ERDF **and CF**
Mapa_tx_erro	159	17	2017/2018	ERDF **and CF**
Universe_1o_per	5188	14	2017/2018	ERDF **and CF**
Universe_2o_per	4646	14	2017/2018	ERDF **and CF**
Sample_pilot	35	22	2017/2018	ESF
Sample_2o_per	33	14	2017/2018	ESF
Map_tx_error	73	20	2017/2018	ESF
Universe_Samp_Pilot	894	16	2017/2018	ESF
Universe_2o_per	1981	14	2017/2018	ESF
Sample_1o_per	38	22	2018/2019	ERDF **and CF**
Sample_2o_per	82	23	2018/2019	ERDF **and CF**
Map_tx_error	120	20	2018/2019	ERDF **and CF**
Universe_1o_per	7487	16	2018/2019	ERDF **and CF**
Universe_2o_per	4976	14	2018/2019	ERDF **and CF**
Sample_1o_per	36	14	2018/2019	ESF
Sample_2o_per	21	12	2018/2019	ESF
Sample_adic_2o_per	2	21	2018/2019	ESF
Map_tx_error	59	19	2018/2019	ESF
Universe_1o_per	3628	15	2018/2019	ESF
Universe_2o_per	4371	15	2018/2019	ESF

Universe_adic_2o_per	55	13	2018/2019	ESF

Appendix Table 2 - Description of the Variables

Variable	Description
Strata	Strata – PP (Public) or NPP (Not Public)
HV_TX	Identifies operations classified as "High Value" as 1 - operations found in the second table in the Error Rate Map file
Sample	Sample – 1st, 2nd, and 3rd
NIF Beneficiary	Beneficiary's NIF
Fund	Fund - ERDF, CF or ESF
Operation Code	Operation Code
Program	Program
Operation Designation	Designation of Operation
Operation Typology Code	Operation Typology Code
Designation of Operation Typology	Designation of Operation Typology
Designation of Beneficiary	Designation of Beneficiary
Intermediate Body	Intermediate Body
Total Approved Cost	Total Approved Cost
Total Certified Delta Cost Mom	Total Certified Delta Cost for the Period
Total Certified Delta Cost without Reinstatements	Total Certified Delta Cost without Reinstatements
Certified Public Expenditure Delta Mom	Certified Public Expenditure Delta for the Period
Delta Certified Fund	Delta Certified Fund

Accumulated Total Certified Cost	Total Certified Cost Accumulated Delta for the Periods
Accumulated Total Certified Cost Delta Mom	Certified Public Debt Accumulated Delta for the Periods
Amounts certified to the EC - Total Cost	Amounts certified to the EC - Total Cost
Audited Expenditure	Audited Expenditure
Total CT Irregular	Total CT Irregular (if the sub-sample is simple random or there is no sub-sample)
Irregular Fund	Irregular background (if the sub-sample is simple random or there is no sub-sample)
Error	Error (if the subsample is simple random or there is no subsample)
Error above Interval	Error above Interval (if subsample is by intervals)
Sub-sample interval value	Sub-sample interval value (if sub-sampling by intervals)
Sum of the error of the values	Sum of the error of the values (if the subsample is by intervals)
Projected error of the subsample	Projected error of the subsample (if the subsample is by intervals)
Total Error	Total Error (if the subsample is by intervals)
Error Rate	Error Rate
Error expressed in euros	Error expressed in euros - product of Error Rate and Total Certified Cost Delta MOM
Audit error	Audit error (at random)
Audit error	Audit error
Projected error	Projected error
Previously audited	Previously audited - TRUE if the beneficiary has already been audited in previous years

Audited in previous periods	Audited in previous periods - TRUE if the beneficiary has already been audited, in the same year, in previous periods
Previously Beneficiary	True if the beneficiary has already benefited from funds in previous years
Subsample	"Yes" if there was under-sampling, i.e. the audited expenditure is less than the total delta certified cost for the period
Sum Strata Pop TCD Cost	Sum of Total Certified Cost Delta (mom) per stratum in the population; A stratum is the set of variables SAMPLE, STRATEGUE and HV_TX
Sum Strata Pop PCD Expenditure	Sum of Delta Certified Public Expenditure (mom) by stratum in the population; A stratum is the set of variables SAMPLE, STRATUM and HV_TX
Sum Strata Pop CTA	Sum of the Total Approved Cost per stratum in the population; A stratum is the set of variables SAMPLE, STRATUM and HV_TX
Nº obs Strata Pop	Number of observations per stratum in the population; A stratum is the set of variables SAMPLE, STRATUM and HV_TX
Sum TCD Cost NIF	Sum of Total Delta Certified Cost (mom) per Beneficiary NIF (in the population)
Sum PCD Expenditure NIF	Sum of Delta Certified Public Expenditure (mom) by Beneficiary NIF (in population)
Sum CTA NIF	Sum of Total Approved Cost per Beneficiary NIF (in the population)
Nº obs NIF Pop	Number of observations by Beneficiary's NIF (in the population)
Sum Strata Sample TCD Cost	Sum of Total Certified Cost Delta (mom) per stratum in the sample; A stratum is the set of variables SAMPLE, STRATUM and HV_TX

Sum Strata Sample PCD Expenditure	Sum of Delta Certified Public Expenditure (mom) by stratum in the sample; A stratum is the set of variables SAMPLE, STRATEGUE and HV_TX
Sum Strata Sample CTA	Sum of Total Approved Cost per stratum in the sample; A stratum is the set of variables SAMPLE, STRATUM and HV_TX
Nº obs Strata Sample	Number of observations per stratum in the sample; A stratum is the set of variables SAMPLE, STRATUM and HV_TX

Appendix Table 3 - Original and Created Operation Typology Codes

Original Operation Typology Code ID	Operation Typology Code Description	New Operation Typology Code ID	New Operation Typology Code Description
1	Production and distribution of renewable energy sources	12	Sustentability and Environment
3	Energy efficiency in Central PA	3	Energetic Efficiency
5	Energy efficiency in Central AP - Awareness campaigns	3	Energetic Efficiency
6	Energy efficiency in regional and local government	3	Energetic Efficiency
8	Energy efficiency in homes - social housing	3	Energetic Efficiency
9	Energy efficiency in homes - housing (private individuals) - Financial Instrument	3	Energetic Efficiency
10	Energy efficiency in homes - housing (private individuals) - Awareness campaigns	3	Energetic Efficiency
12	Urban mobility plans - planning	11	Mobility and Transports
13	Urban mobility plans - investments (infrastructure and equipment)	11	Mobility and Transports
14	Electric mobility	11	Mobility and Transports
15	Energy efficiency in public transport	3	Energetic Efficiency

16	Adapting to climate change	12	Sustentability and Environment
17	Material coastal protection actions in risk zones	12	Sustentability and Environment
19	Forest fires	12	Sustentability and Environment
20	Floods	12	Sustentability and Environment
21	Emergency means and structural actions in the face of accidents and disasters	12	Sustentability and Environment
22	Planning, monitoring and communication	12	Sustentability and Environment
23	Innovative actions for risk prevention and management	12	Sustentability and Environment
24	Waste	12	Sustentability and Environment
25	Water supply	12	Sustentability and Environment
27	Wastewater Sanitation	12	Sustentability and Environment
29	Water resource management	12	Sustentability and Environment
30	Natural Heritage	4	Patrimony
31	Cultural Heritage	4	Patrimony
32	Tourism promotion	4	Patrimony
33	Nature Conservation	12	Sustentability and Environment
34	Management and planning of protected and classified areas	12	Sustentability and Environment
35	Information	13	Information
36	Urban regeneration - financial instrument	6	Urban and infrastructure rehabilitation/requalification and local development
37	Comprehensive Building Rehabilitation	6	Urban and infrastructure rehabilitation/requalification and local development
39	Rehabilitation of public spaces	6	Urban and infrastructure rehabilitation/requalification and local development
41	Studies and actions associated with quality improvement, noise reduction and quality of life in urban environments	12	Sustentability and Environment
42	Regeneration of environmental liabilities	5	Regeneration of Liabilities
43	Regeneration of mining liabilities	5	Regeneration of Liabilities

44	Support for hiring	10	Support for Hiring and Internships
47	Traineeships for Adults	10	Support for Hiring and Internships
55	Integration of young people into the labor market	10	Support for Hiring and Internships
58	Empreende Já - Business Perception and Management Network	1	Entrepreneurship incentives and business counseling
60	Estagiar T (Azores)	10	Support for Hiring and Internships
62	IEJ Internships	10	Support for Hiring and Internships
63	Internships for young people	10	Support for Hiring and Internships
66	INOV Contacto	10	Support for Hiring and Internships
67	INTEGRA	10	Support for Hiring and Internships
68	PEPAC Missions (Professional Internship Program in Central Administration)	10	Support for Hiring and Internships
69	PEPAL (Program for Professional Internships in Local Administration)	10	Support for Hiring and Internships
70	Programa de Incentivo à Inserção do Estagiar L e T - PIIE (Azores)	10	Support for Hiring and Internships
76	Actions to support entrepreneurship (Ideia Jovem Invest)	1	Entrepreneurship incentives and business counseling
80	Support for entrepreneurship	1	Entrepreneurship incentives and business counseling
85	Local entrepreneurship projects - Support for employment and investment	1	Entrepreneurship incentives and business counseling
90	Maternity protection program and promotion of female employability	1	Entrepreneurship incentives and business counseling
94	Modular Training	1	Entrepreneurship incentives and business counseling

96	Modular training aimed essentially at workers in micro and small businesses	1	Entrepreneurship incentives and business counseling
97	Modular Training for the Employed and Unemployed	1	Entrepreneurship incentives and business counseling
108	Strengthening the Institutional Capacity of the Social Partners with a seat on the Social Dialogue Standing Committee	8	Social-related actions and awareness
109	Promoting micro-entrepreneurship	1	Entrepreneurship incentives and business counseling
110	Valuing endogenous resources in specific territories	12	Sustentability and Environment
111	Support for the employment of people with disabilities (Internships for people with disabilities)	8	Social-related actions and awareness
113	Qualification of people with disabilities	8	Social-related actions and awareness
116	Training for Inclusion	8	Social-related actions and awareness
120	Occupational programs	8	Social-related actions and awareness
125	Modular training for the long-term unemployed (DLD)	8	Social-related actions and awareness
128	CLDS (Local Social Development Contracts)	8	Social-related actions and awareness
131	Socio-professional integration of the Roma community	8	Social-related actions and awareness
132	Program "Escolhas"	8	Social-related actions and awareness
140	Financial and technical support for non-profit civil society organizations	8	Social-related actions and awareness
141	Awareness-raising actions and campaigns in the field of gender equality, prevention and combating domestic violence	8	Social-related actions and awareness
144	Specific Instruments for Protecting Victims and	8	Social-related actions and awareness

	Accompanying Aggressors in Domestic Violence		
155	National Immigrant Support Centers	8	Social-related actions and awareness
156	National Immigrant Support Centers (Lisbon and Algarve)	8	Social-related actions and awareness
159	Training for teachers, technicians and other CPCJ professionals	8	Social-related actions and awareness
162	Idade +	8	Social-related actions and awareness
170	RLIS (Local Social Intervention Network)	8	Social-related actions and awareness
175	Institutional capacity building for social economy organizations that are members of the national council for the social economy	8	Social-related actions and awareness
177	Social Innovation Funds	8	Social-related actions and awareness
178	Partnerships for impact	8	Social-related actions and awareness
180	Social Impact Bonds	8	Social-related actions and awareness
181	Supporting entrepreneurship - promoting networking	1	Entrepreneurship incentives and business counseling
182	Support for local entrepreneurship	1	Entrepreneurship incentives and business counseling
183	Local socio-economic development	6	Urban and infrastructure rehabilitation/requalification and local development
184	Social infrastructure and facilities	6	Urban and infrastructure rehabilitation/requalification and local development
185	Health infrastructure and equipment	6	Urban and infrastructure rehabilitation/requalification and local development
186	Urban regeneration - disadvantaged communities	6	Urban and infrastructure rehabilitation/requalification and local development

188	Scientific and Technological Research Support System - IC&DT Projects	2	Scientific and Technological Investigation Support System
190	Scientific and Technological Research Support System - Joint Activity Projects (PAC)	2	Scientific and Technological Investigation Support System
191	Scientific and Technological Research Support System - Integrated IC&DT Programs	2	Scientific and Technological Investigation Support System
193	Scientific and Technological Research Support System - Protection of intellectual property rights	2	Scientific and Technological Investigation Support System
194	Scientific and Technological Research Support System - Projects for the development and implementation of research infrastructures	2	Scientific and Technological Investigation Support System
195	Scientific and Technological Research Support System - R&D internationalization projects	2	Scientific and Technological Investigation Support System
196	SI Incentive System for Research and Technological Development - Protection of intellectual and industrial property	2	Scientific and Technological Investigation Support System
197	SIAC Support System for Collective Actions - Transfer of scientific and technological knowledge	2	Scientific and Technological Investigation Support System
198	SI Incentive System for Research and Technological Development - R&DT Projects Companies	2	Scientific and Technological Investigation Support System
199	SI Incentive System for Research and Technological Development - Demonstrator projects	2	Scientific and Technological Investigation Support System
200	SI Incentive System for Research and Technological Development - Mobilizing Programs	2	Scientific and Technological Investigation Support System

201	SI Incentive System for Research and Technological Development - R&TD Centers	2	Scientific and Technological Investigation Support System
202	SI Incentive System for Research and Technological Development - Internationalization of R&D	2	Scientific and Technological Investigation Support System
203	SI Incentive System for Research and Technological Development - Vale I&D	2	Scientific and Technological Investigation Support System
204	SI Incentive System for Research and Technological Development - contractual regime	2	Scientific and Technological Investigation Support System
205	SIAC Support System for Collective Actions - Networks and other forms of partnership and cooperation	1	Entrepreneurship incentives and business counseling
206	SI Business Innovation and Entrepreneurship - Productive Innovation Non-SMEs	1	Entrepreneurship incentives and business counseling
207	SI Business Innovation and Entrepreneurship - Productive Innovation for Non-SMEs - contractual schem	1	Entrepreneurship incentives and business counseling
208	System to support the modernization and empowerment of Public Administration - Promoting a networked administration	14	Public Administration and Others
208	Modernization of the Public Administration through ICT (ERDF)	14	Public Administration and Others
210	System to support the modernization and training of the Public Administration - Implementation of new integrated decentralized service models in the Public Administration (citizens' stores, citizens' spaces and mobile services)	14	Public Administration and Others
212	SI Business innovation and entrepreneurship - Qualified	1	Entrepreneurship incentives and business counseling

	and creative entrepreneurship - Individual project		
213	SI Business Innovation and Entrepreneurship - Qualified and creative entrepreneurship - contractual regime	1	Entrepreneurship incentives and business counseling
214	SI Business Innovation and Entrepreneurship - Vale Empreendedorismo	1	Entrepreneurship incentives and business counseling
215	SIAC Support System for Collective Actions - Promoting entrepreneurship	1	Entrepreneurship incentives and business counseling
216	Financial Instrument (FI)	15	Financial Instrument
217	SI Qualification and internationalization of SMEs - Individual project	1	Entrepreneurship incentives and business counseling
218	SI qualification and internationalization of SMEs - Vale Internacionalização	1	Entrepreneurship incentives and business counseling
219	SI qualification and internationalization of SMEs - Joint project for the internationalization of SMEs (except training-action)	1	Entrepreneurship incentives and business counseling
220	SIAC Support System for Collective Actions - Internationalization	1	Entrepreneurship incentives and business counseling
221	SI Business Innovation and Entrepreneurship - SME Productive Innovation	1	Entrepreneurship incentives and business counseling
222	SI Business Innovation and Entrepreneurship - SME Productive Innovation - contractual regime	1	Entrepreneurship incentives and business counseling
223	SI qualification and internationalization of SMEs - Innovation Valley	1	Entrepreneurship incentives and business counseling
224	SI SME qualification and internationalization - Joint SME qualification project (except training-action)	1	Entrepreneurship incentives and business counseling
225	SIAC Collective Action Support System - Qualification	1	Entrepreneurship incentives and business counseling

226	SI Business Investment (Azores)	1	Entrepreneurship incentives and business counseling
227	Railroad	11	Mobility and Transports
228	Ports	11	Mobility and Transports
231	Road mobility in the Ors	11	Mobility and Transports
234	SI Qualification and internationalization of SMEs - Hiring highly qualified human resources (SMEs)	1	Entrepreneurship incentives and business counseling
241	SI Incentive System - Training-action for SMEs	1	Entrepreneurship incentives and business counseling
249	Training for business innovation	1	Entrepreneurship incentives and business counseling
254	Capacity Building for Public Administration (ESF)	14	Public Administration and Others
258	Strategic training actions for efficient management in public administration	14	Public Administration and Others
260	Institutional training in territorial partnerships	9	Education (courses, equipment acquisiton)
271	Information, monitoring and evaluation actions relating to measures and devices to prevent school dropout and promote students' educational success	9	Education (courses, equipment acquisiton)
273	Education and Training Courses (CEF)	9	Education (courses, equipment acquisiton)
275	Vocational courses	9	Education (courses, equipment acquisiton)
277	Integrated and innovative plans to combat school failure	9	Education (courses, equipment acquisiton)
278	TEIP, PIEF, Mais Sucesso	9	Education (courses, equipment acquisiton)
279	Youth training and insertion program courses (PROFIJ)	9	Education (courses, equipment acquisiton)
284	Continuous training for teachers and other staff	9	Education (courses, equipment acquisiton)
286	Continuous training for pre-school, primary and secondary school teachers and trainers	9	Education (courses, equipment acquisiton)

288	Quality and efficiency of the education and training system to promote school success	9	Education (courses, equipment acquisiton)
290	SPO (Psychology and Guidance Services) - Network	8	Social-related actions and awareness
294	#NOME?	9	Education (courses, equipment acquisiton)
296	Higher education grants for needy students	9	Education (courses, equipment acquisiton)
297	Technical Higher Professional Courses (TESP Courses)	9	Education (courses, equipment acquisiton)
299	Individual doctoral and post-doctoral scholarships	9	Education (courses, equipment acquisiton)
300	Doctoral Programs and Postdoctoral Fellowships	9	Education (courses, equipment acquisiton)
301	Actions to retrain people with higher education qualifications in areas with career opportunities	9	Education (courses, equipment acquisiton)
302	CQEP (Centers for Qualification and Vocational Education) and Centros Qualifica	9	Education (courses, equipment acquisiton)
303	Apprenticeships	9	Education (courses, equipment acquisiton)
304	Basic skills acquisition courses	9	Education (courses, equipment acquisiton)
305	Adult Education and Training Courses (EFA)	9	Education (courses, equipment acquisiton)
310	Modular dual certification training courses, at basic or secondary level, school or professional certification, included in the CNQ	9	Education (courses, equipment acquisiton)
311	CET (Technological Specialization Courses)	9	Education (courses, equipment acquisiton)
312	Professional Courses	9	Education (courses, equipment acquisiton)
316	Interventions to comply with Assembly of the Republic Resolution no. 24/2003, of April 2, and Law no. 2/2011, of February 9, for the removal of	6	Urban and infrastructure rehabilitation/requalification and local development

	asbestos cement and give the building greater thermal comfort and watertight conditions		
318	Acquisition of new information and communication technology (ICT) equipment when related to the introduction of new courses or methods and when this investment fits in with pedagogical and educational objectives associated with new courses and new methodologies	9	Education (courses, equipment acquisiton)
319	Improvements (infrastructures) to schools in the 2nd and 3rd cycles of basic education and secondary education	6	Urban and infrastructure rehabilitation/requalification and local development
320	Acquisition of higher education equipment for new TeSP/ISCED level short courses 5	9	Education (courses, equipment acquisiton)
321	Acquisition of equipment for new TeSPs or the creation of new higher education programs to meet the needs of the job market	9	Education (courses, equipment acquisiton)
322	Upgrading and modernization of vocational training infrastructures.	6	Urban and infrastructure rehabilitation/requalification and local development
323	RUP	11	Mobility and Transports
324	Studies and Evaluations	14	Public Administration and Others
329	Management and monitoring	14	Public Administration and Others
330	Technical Assistance (not applicable PO AT)	14	Public Administration and Others
444	Socially necessary work - CEI and CEI+	8	Social-related actions and awareness
445	Technological infrastructure	2	Scientific and Technological Investigation Support System

446	SI Support for restoring productive capacity (fire June 2017)	6	Urban and infrastructure rehabilitation/requalification and local development
447	Internships	10	Support for Hiring and Internships
448	Recovering the basic municipal infrastructures affected by the fires (June 2017 fire)	6	Urban and infrastructure rehabilitation/requalification and local development
503	Business Hosting (including ALE and Incubators)	1	Entrepreneurship incentives and business counseling
504	Technological infrastructure	2	Scientific and Technological Investigation Support System

APPENDIX B. FEATURE SELECTION

Appendix Table 4- Feature Selection Results for ERDF_CF: RFE top 8

Var	RMSE	R squared	MAE
1	0.059	0.007	0.026
2	0.046	0.127	0.022
3	0.042	0.144	0.021
4	0.041	0.140	0.021
5	0.041	0.150	0.021
6	0.043	0.093	0.030
7	0.043	0.094	0.031

Appendix Table 5 - Feature Selection Results for ESF: RFE top 8

Var	RMSE	R squared	MAE
1	0.029	0.156	0.020
2	0.028	0.119	0.018
3	0.027	0.097	0.018
4	0.027	0.130	0.018
5	0.027	0.144	0.018
6	0.027	0.072	0.018
7	0.027	0.103	0.018
8	0.027	0.097	0.018

Appendix Table 6 – Feature Selection Results for ERDF_CF: XGBoost importance top 8

Var	Gain	Cover	Frequency
Total Certified Cost	0.445	0.399	0.355
Total Approved Cost	0.365	0.382	0.271
Operation Typology Code 1	0.058	0.065	0.065
Previous Beneficiary False	0.038	0.030	0.037
Program “Norte”	0.032	0.025	0.047
Year 2017/18	0.031	0.010	0.075
Program “POCI”	0.126	0.017	0.028
Program “Lisboa”	0.008	0.015	0.037

Appendix Table 7 - Feature Selection Results for ESF: XGBoost importance

Var	Gain	Cover	Frequency
Total Approved Cost	0.452	0.342	0.347
Total Certified Cost	0.201	0.390	0.327
Program “POCH”	0.107	0.034	0.051
Previously Audited False	0.107	0.034	0.041
Previously period Audited False	0.054	0.042	0.020
Operation Typology Code 4	0.030	0.013	0.031
Strata NPP	0.015	0.037	0.020
Program “POISE”	0.014	0.028	0.020

Appendix Table 8 – Feature Selection Results for ERDF_CF: RFE

Var	Accuracy	Kappa
1	0.749	0.089
2	0.785	0.062
3	0.776	-0.008
4	0.762	0.049
5	0.767	0.053
6	0.771	0.045
7	0.789	0.124
8	0.780	0.104
9	0.781	0.136
10	0.781	0.142
11	0.785	0.122
12	0.780	0.111
13	0.776	0.087
14	0.780	0.094
15	0.789	0.132
16	0.789	0.163
17	0.793	0.153
18	0.793	0.170
19	0.793	0.153

Appendix Table 9 - Feature selection for ERDF_CF: Random Forest Importance

Var	0	1	Mean Decrease Accuracy	Mean Decrease Gini
Year 2017/18	0.006	0.027	0.010	3.960
Program “Açores”	-0.003	0.005	0.001	1.452
Program “ALT20”	-0.001	-0.000	0.000	0.344
Program “Centro”	0.004	-0.002	0.003	1.774
Program “Lisboa”	0.009	0.020	0.011	3.648
Program “M1420”	0.001	0.002	0.001	1.220
Program “Norte”	-0.000	0.001	0.000	1.137
Program “POCI”	0.041	-0.005	0.031	3.830
Program “POSEUR”	0.001	-0.001	0.001	0.717
Previous Beneficiary False	0.002	0.003	0.002	2.914
Previously Audited False	0.007	0.021	0.010	3.045
Previously period Audited False	0.001	0.004	0.001	0.610
Oper.Typology Code 1	0.008	-0.002	0.006	2.503
Oper.Typology Code 2	0.001	0.001	0.001	1.836
Oper.Typology Code 4	0.000	-0.001	0.000	1.300
Oper. Typology Code 5	0.001	0.012	0.003	1.224
Oper. Typology Code 6	0.001	-0.002	0.000	0.756
Total Approved Cost	0.049	0.006	0.039	29.934
Total Certified Cost	0.058	0.022	0.050	32.638

Appendix Table 10 - Feature Selection Results for ESF: RFE

Var	Accuracy	Kappa
4	0.625	0.125
5	0.644	0.176
6	0.665	0.218
7	0.665	0.224
8	0.644	0.183
9	0.632	0.151
18	0.654	0.188

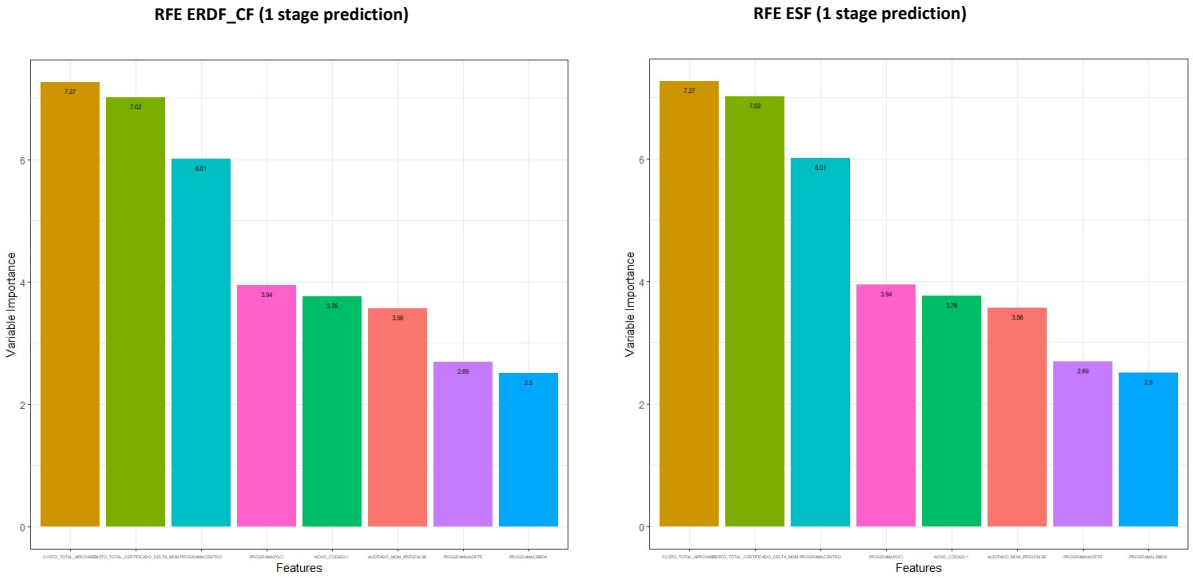
Appendix Table 11- Feature selection for ESF: Random Forest Importance

Var	0	1	Mean Decrease Accuracy	Mean Decrease Gini
Year 2017/18	0.024	0.016	0.019	3.428
Program “Açores”	-0.004	0.005	0.002	1.354
Program “ALG_Centro”	-0.001	0.000	-0.001	0.429
Program “ALT20”	0.000	0.005	0.003	0.890
Program “Lisboa”	-0.003	0.002	0.000	0.784
Program “M1420”	0.015	0.004	0.009	1.681
Program “Norte”	-0.001	-0.002	-0.002	0.781
Program “POCI”	-0.003	0.032	0.019	2.225
Program “POISE”	-0.006	0.011	0.005	1.035
Strata NPP	0.032	0.012	0.019	2.555
Previous Beneficiary False	0.000	0.000	0.000	0.940
Previously Audited False	0.020	-0.009	0.001	1.472
Previously period Audited False	-0.007	0.004	0.000	0.985
Oper.Typology Code 1	-0.012	0.008	0.000	1.299
Oper.Typology Code 3	0.000	0.033	0.021	1.771
Oper. Typology Code 4	-0.006	0.029	0.016	1.522
Total Approved Cost	0.035	0.036	0.034	14.766
Total Certified Cost	-0.023	0.017	0.001	13.047

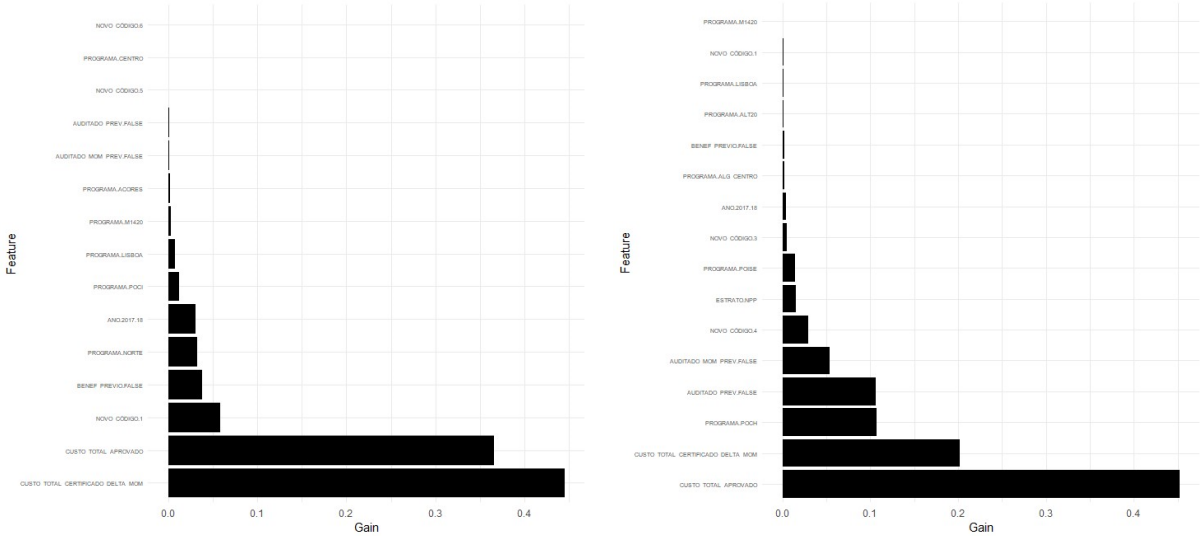
Appendix Table 12 - Feature Selection: Boruta

Parameters	ERDF_CF	ESF
Important	Total Certified Cost	Total Approved Cost
	Total Approved Cost	Oper.Typology Code 3
	Program “Lisboa”	Program “M1420”
	Program “POCI”	Program “POCH”
Unimportant	Year 2017/18	Previously period Audited False
	Previously period Audited False	Previously Audited False
	Previous Beneficiary False	Previous Beneficiary False

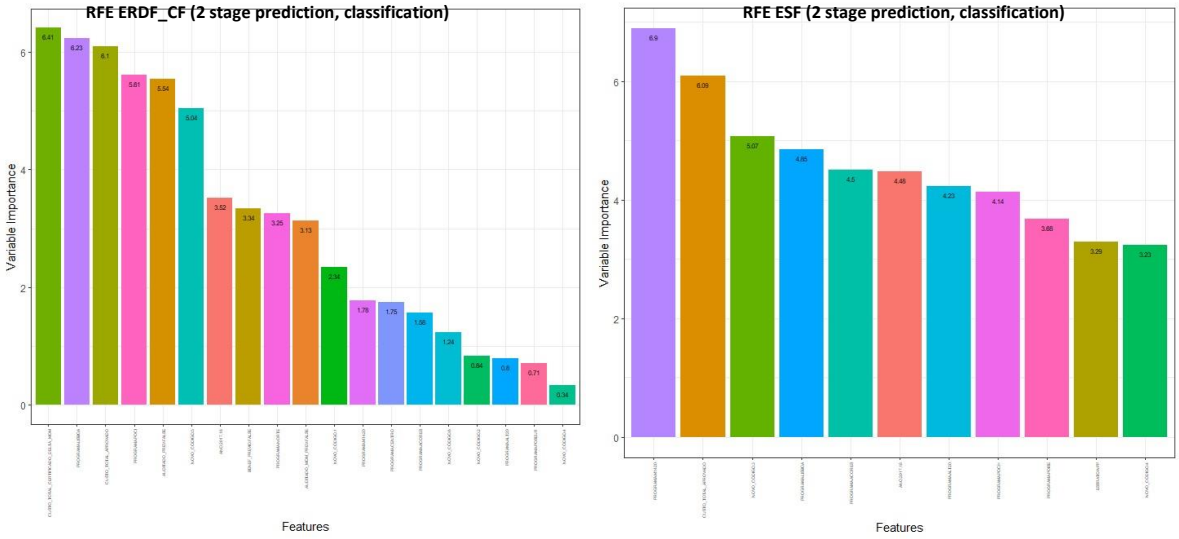
	Oper.Typology Code 1	Total Certified Cost
	Oper.Typology Code 2	Strata NPP
	+ 9	+5
Tentative	Previously Audited False	Year 2017/18 Program “Açores” Program “Lisboa” Program “POISE”



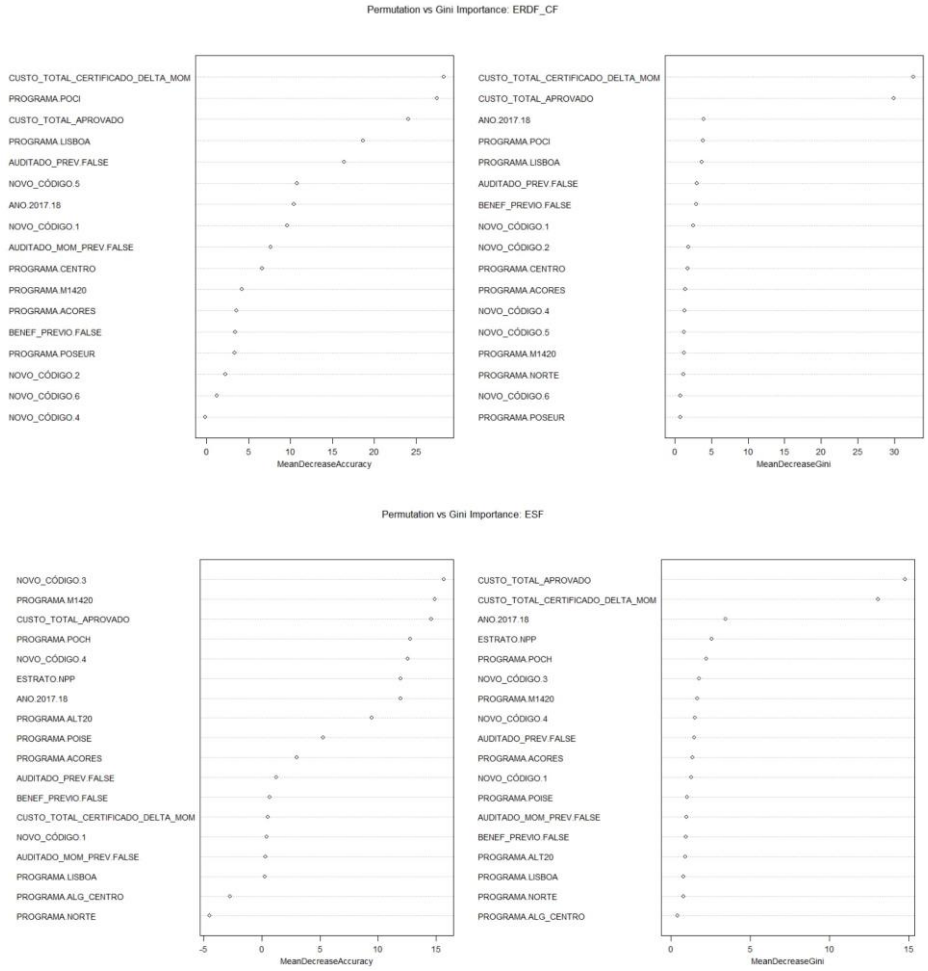
Appendix Plot 1 - RFE Plots for ERDF_CF and ESF (1-stage prediction)



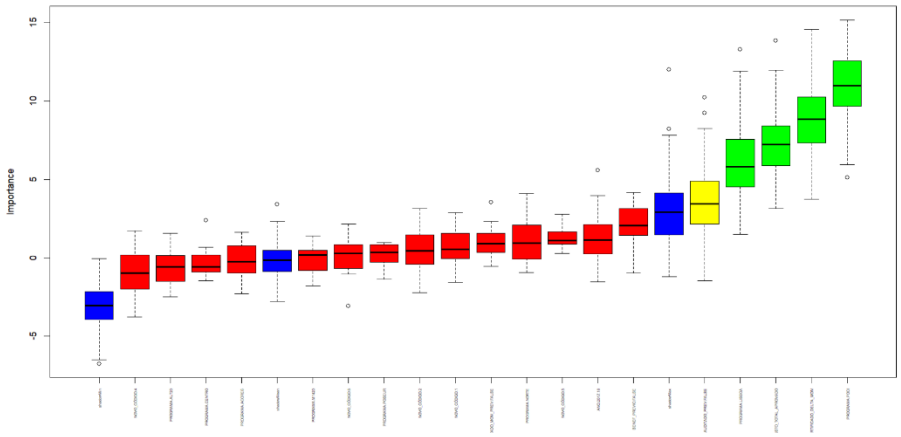
Appendix Plot 2 – XGBoost Importance for ERDF_CF and ESF (1-stage prediction)



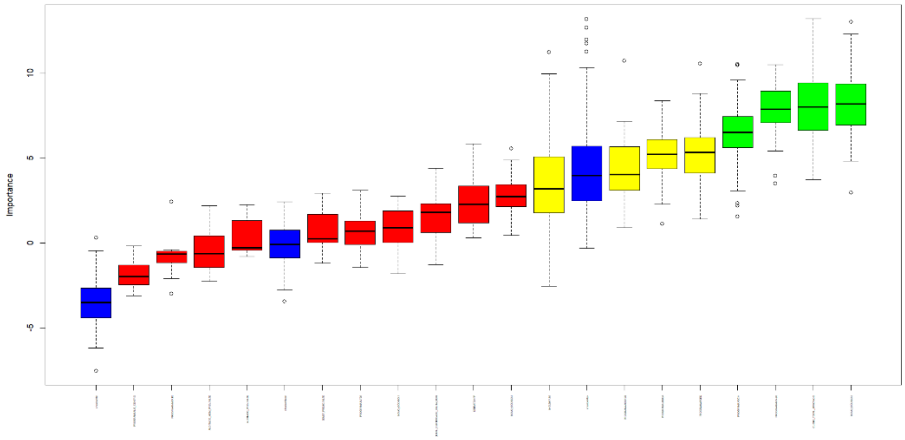
Appendix Plot 3 - RFE Plots for ERDF_CF and ESF (2-stage prediction)



Appendix Plot 4 - Permutation vs Gini Importance Results

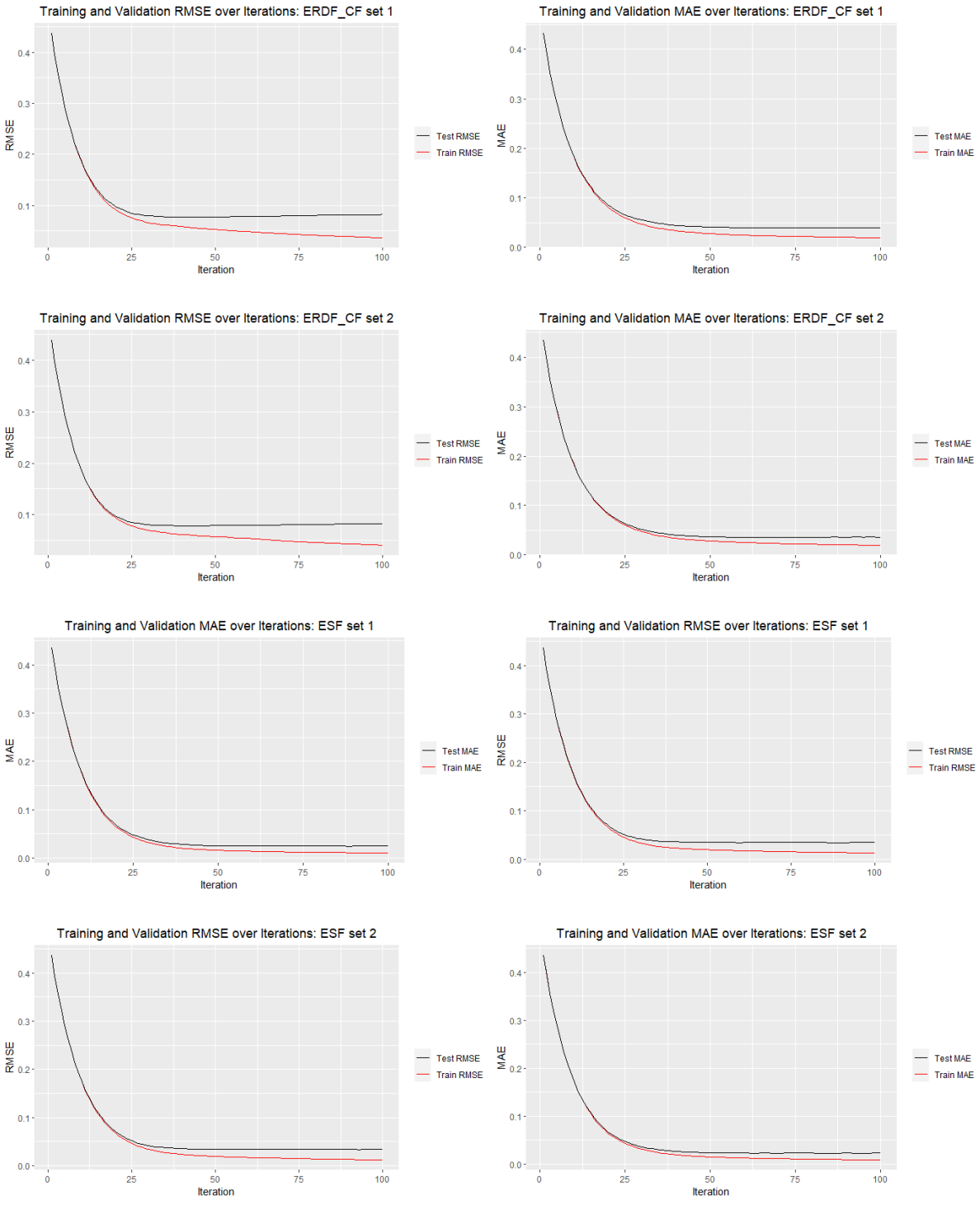


Appendix Plot 5 - Boruta Results for ERDF_CF



Appendix Plot 6 - Boruta Results for ESF

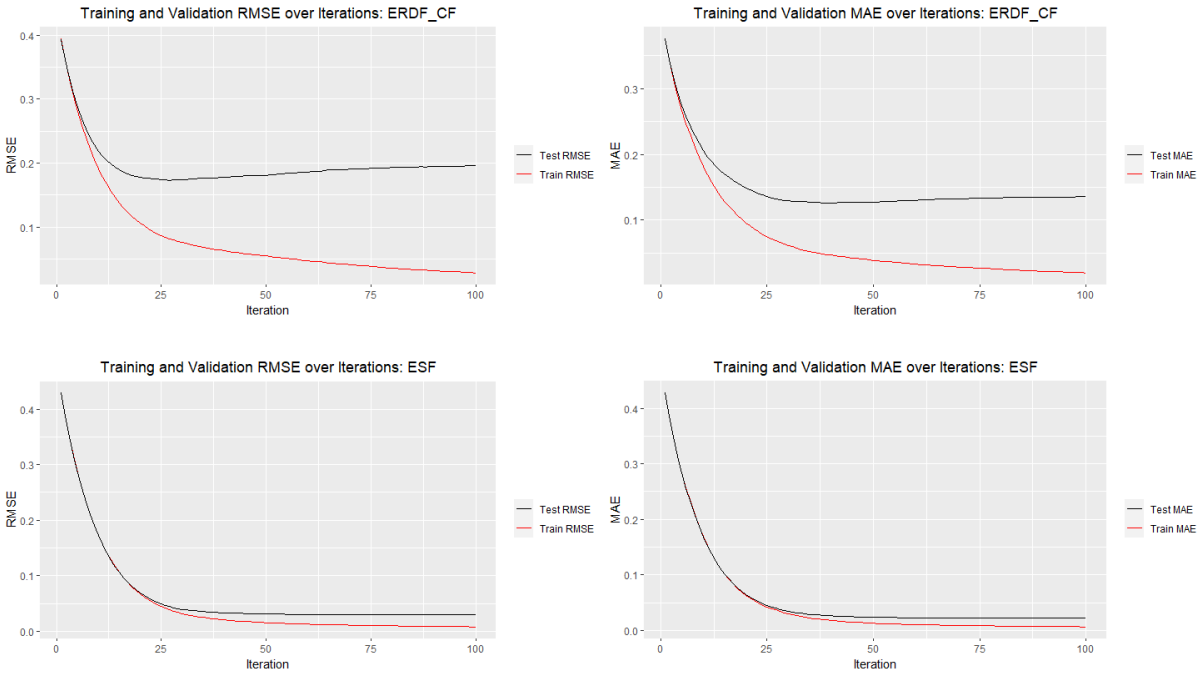
APPENDIX C. 1-STAGE PREDICTION TRAIN AND VALIDATION METRICS AND RESULTS



Appendix Table 13 - Summary of prediction model performance metrics

Metric	Description	Formula
R-squared (R^2)	Proportion of variation in the outcome that is explained by the predictor variables.	$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}$
Root Mean Squared Error (RMSE)	Measures the average error performed by the model in predicting the outcome for an observation.	$\sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$
Mean Absolute Error (MAE)	Measures the average absolute difference between the observed and predicted outcomes.	$\frac{1}{n} \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$

APPENDIX D. 2-STAGE PREDICTION TRAIN AND VALIDATION METRICS AND RESULTS



Appendix Table 14 - Confusion Matrix

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True positives (TP)	False negatives (FN)
	Negative	False positives (FN)	True negatives (TN)

Appendix Table 15 - Summary of classification model performance metrics

Metric	**Description**	**Formula**
Accuracy	Proportion of the total number of correct predictions that were correct	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Proportion of positive cases that were correctly identified	$\frac{TP + TN}{(TP + FP)}$

Sensitivity	Proportion of actual positive cases which are correctly identified	$\frac{TP}{(TP + FN)}$
Specificity	Proportion of actual negative cases which are correctly identified	$\frac{TN}{(TN + FP)}$

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa