U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Detecting Outliers in Accounting Data: a Machine Learning Approach

Ana Filipa Vieira dos Santos

Dissertation
Master in Modelling, Data Analysis and Decision Support Systems

Supervised by
Professor João Manuel Portela da Gama

2023

Acknowledgments

This dissertation marks the conclusion of an enriching learning process at the Master's Degree in Modelling, Data Analysis and Decision Support Systems.

I would like to thank my internship supervisors, for expertly introducing me to the problem and to the operations of the institution, as well as providing me with everything I needed to succeed in my work. To my dissertation supervisor, for helping me guide my research in the right direction.

Lastly, to my family and friends, for always supporting and encouraging me.

Abstract

Non-financial corporations with a permanent establishment in Portuguese territory are obligated to submit financial statements to Banco de Portugal every year. Quality control of the information received is a crucial part of the data collecting process, since these reports might contain erroneous values that should be identified and corrected.

The purpose of this work was to implement an unsupervised Machine Learning algorithm for outlier detection, the Isolation Forest, to identify irregularities in the data and point out which companies should be further examined by an analyst. After a thorough selection of accounting items and the separation of the companies according to their characteristics, ensuring that each corporation is only compared against their close peers, the model assigns each observation in the sample an anomaly score.

The companies that are identified as containing irregularities can be sorted according to their attributed score, letting the analyst know which reports are a priority. Furthermore, with the use of Shapley values, it is possible to determine which features contributed the most to the score of an anomalous corporation, highlighting which sections of the report seem to contain an abnormal value.

In complement with some mechanisms that are already in place, these procedures will help to detect irregularities in the data that should be further examined by an analyst, improving the efficiency and efficacy of the quality control process.

Keywords: Outlier Detection, Unsupervised Learning, Quality Control, Accounting Data, Isolation Forest, Shapley Values

Resumo

Sociedades não financeiras com estabelecimento estável em território português estão obrigadas a submeter, todos os anos, declarações financeiras ao Banco de Portugal. A qualidade de controlo da informação recebida é uma etapa importante do processo de recolha de dados, uma vez que esta pode conter valores incorretos que devem ser identificados e corrigidos.

O objetivo desta dissertação é implementar um algoritmo de aprendizagem não supervisionada para a deteção de outliers, o Isolation Forest, para identificar irregularidades nos dados e assinalar as empresas que devem ser examinadas mais detalhadamente por um analista. Após uma seleção minuciosa das rubricas contabilísticas e a separação das empresas de acordo com as suas características, garantindo que cada sociedade é comparada apenas com sociedades similares, o modelo atribui a cada observação da amostra um score de anomalia.

As empresas que são identificadas como contendo irregularidades podem ser ordenadas de acordo com o score atribuído, indicando ao analista quais as declarações que devem ser priorizadas. Com o uso de valores de Shapley, é ainda possível determinar quais as variáveis que mais contribuíram para o score de uma empresa anómala, destacando as secções da declaração que podem conter valores anormais.

Em complemento com alguns mecanismos que já estão em prática, este modelo vai ajudar a detetar irregularidades nos dados que devem depois ser examinadas por um analista, melhorando a eficiência e a eficácia do processo de controlo de qualidade.

Palavras-Chave: Deteção de Outliers, Aprendizagem não Supervisionada, Controlo de Qualidade, Dados Contabilísticos, Isolation Forest, Valores de Shapley

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1. Motivation

The growth of microdata, both in usage and in volume, raises multiple new issues and challenges. One of the most important steps in the data collecting process is ensuring that the data is accurate and representative of the population under study. However, with high volumes of microdata, quality control might become a complex and time-consuming procedure.

Banco de Portugal's Central Balance Sheet Office gathers data regarding the Informação Empresarial Simplificada (IES), an annual report that must be submitted by non-financial corporations with a permanent establishment in Portuguese territory. The mandatory submission of IES started in 2006, with the intention of aggregating, in one single report, all the economic and financial information about companies that previously had to be separately submitted to various different entities. The information collected from these reports is essential for accounting, tax, and statistical purposes, covering nearly one hundred percent of the Portuguese business sector.

Business surveys submitted by companies, like any other type of data, are subject to the presence of incorrect information. The erroneous values might derive from unintentional mistakes (such as writing a wrong digit while entering the data or failing to correctly follow the established accounting guidelines), or from intentional attempts at manipulating financial statements. Therefore, the process of quality control is essential in ensuring that the data collected every year accurately represents each corporation, as well as the business sector as a whole.

At the moment, the institution has a few automatic mechanisms in place for the detection of erroneous values, such as the correction of systematic errors that are known to be made by a specific company every year or the crossing of information with other internal and external databases. These procedures help to significantly improve the quality of the data but are not able to detect every single irregularity. For a more in-depth investigation, the financial statements submitted by corporations need to be subject to manual checks by analysts, who examine the reports looking for values that seem to be inconsistent with the typical behavior of the company or with the behavior of companies with similar

characteristics. Performing manual inspections on financial statements is a time-consuming process and, due to the high number of corporations that are obligated to submit this report every year, it is not possible to examine the entirety of the collected data with the resources available. The selection of the companies to be examined is currently made based on simple rules that take into account some qualitative and quantitative information about the enterprises and their financial reports.

With the inability to manually check every report, it is important to have a robust method for selecting the companies, reducing the probability of significant irregularities going unidentified in the corporations' financial statements. Therefore, it would be beneficial to study the implementation of new data quality control procedures that can replace the current selection rules and increase the efficiency and efficacy of this process.

1.2. Problem Description

Given the issues previously described, the quality control process might benefit from the implementation of an anomaly detection procedure, which can detect irregularities in the financial statements and identify the companies that should be subject to a manual inspection. The first stage of quality control, which consists of automatic corrections, would remain in place, considering that it has presented satisfactory results, with the new procedure being implemented in the second phase, replacing the rules that are currently in place to select companies for further examination.

The aim of this work is to introduce to the quality control stage the implementation of a Machine Learning algorithm. Due to the considerable size of the financial reports, the ideal outlier detection mechanism should be able to handle large amounts of data while maintaining a reasonably low execution time. Moreover, the IES data is not labeled, and introducing labels to the problem would be an unviable and time-consuming procedure, so the use of unsupervised learning algorithms is essential.

The mechanism will detect anomalies in the data, highlighting the financial statements of the companies that seem to display suspicious and concerning behavior. Each corporation should be assigned an outlier score by the algorithm, allowing for the irregular

observations to then be sorted according to their attributed marks, from the most severe irregularity to the least.

The most critical outlying reports, detected by the algorithm, will then earn a follow-up by a specialized analyst, who will manually check the information and further investigate the source of the irregularity, usually by cross-checking information with other sources or by personally contacting the respective corporation. The priorities for the examination should be the companies that the algorithm attributed the most severe outlier scores to, eventually moving to other corporations down the sorted list, as long as there are enough available resources.

Considering the large size of the IES reports, an additional issue that needs to be addressed in order to optimize the examination process is to identify, in the anomalous statements, which accounting items seem to contain irregularities so that the examiner is aware of where the anomalous values are located and can, therefore, focus their inspection on those fields.

It is expected that the practical implementation of this mechanism will help to improve the quality of the information collected and stored by the institution in the Central Balance Sheet Database. The detection of anomalies will allow for the examination, and possible correction (if needed), of irregularities in the companies' financial statements that would not have been detected otherwise.

1.3. Dissertation Structure

Following the introduction to the problem under study, which was done in this chapter, this dissertation contains four more main sections. Chapter 2 presents the main information gathered during the process of literature review, which is essential for the posterior practical implementation of this work: fundamental aspects of the concept of outlier, an overview of the most important details about the process of outlier detection, and the exposition of the Isolation Forest algorithm. Subsequently, in Chapter 3, the methodology used is introduced and explained, covering issues such as the deliberation of which current company selection criteria should be replaced, the initial data extraction and treatment, the selection of features to be included in the model, the approach chosen to deal with the diversity of the companies

in the sample, and the procedures behind the practical application of the work. Chapter 4 contains the results obtained in the implementation and their respective analysis, comprising the application of the anomaly detection model and Shapley values for feature explainability. Lastly, Chapter 5 presents an overview of the main results and some considerations about work to be implemented in the future.

Chapter 2 – Literature Review

The practical implementation of this work requires a coherent theoretical background about the aspects surrounding the concept of outliers and the approach that will be used for outlier detection. This chapter presents the main information collected during the literature review process through the analysis of books and scientific articles. Firstly, some fundamental notions about outliers are presented, followed by an overview of the most important aspects of the outlier detection process. Subsequently, the Isolation Forest algorithm is introduced in detail, along with some of its advantages and proposed improvements.

2.1. The Concept of Outliers

2.1.1. Definition of Outlier

Over the years, the presence of outliers has been the subject of thorough studies across various fields, such as statistics and data mining, with many characterizations being proposed for this type of observations. One of the most cited definitions for outlier was suggested by Grubbs (1969), who stated that "an outlying observation … is one that appears to deviate markedly from other members of the sample in which it occurs".

Other important aspects of the concept of outlier were pointed out by different authors, such as the fact that these observations have distinct characteristics from their neighbors (Papadimitriou et al., 2003), seem to be inconsistent with the rest of the sample (Barnett & Lewis, 1994), and might have had a different process of origin (Hawkins, 1980).

2.1.2. Types of Outliers

The broad definition of outlier aggregates several different types of anomalies, which can be further divided according to their characteristics. In literature, various possible classifications are proposed, with each concentrating on different properties of these observations.

There are multiple different causes that lead to the appearance of anomalous observations, thus these instances can be separated according to their origin. Anscombe (1960) presented the following three sources for the generation of outliers:

- inherent variability – the natural volatility of the data, which is intrinsic to the population under study and cannot be adjusted since the recorded values are representative of the individuals;
- measurement error – deviations resulting from the inaccuracy of the instruments or techniques that are used to collect the data;
- execution error – misjudgments made during the data collection phase (for example, including in the sample individuals that do not belong to the population under study).

A different anomaly classification method was suggested by Chandola et al. (2009), considering different aspects of the concept of outlier, such as its nature and context. The proposed three groups are the following:

- point outliers – individual data points that express a different behavior from the rest of the sample (the fundamental and most researched type of anomalies, which can appear in any set of data);
- contextual outliers – instances that appear to be abnormal when a particular context is considered (can only be detected when the set of data contains contextual attributes, in addition to the usual behavioral attributes);
- collective outliers – a group of related data points that exhibit deviating behavior when considered as whole (as individuals, the instances do not necessarily have to be anomalies).

2.1.3. Outliers in a Financial Context

The data collected from business surveys is not immune to the presence of outliers. An important step in survey processing is to find the values that fall outside of the expected range and further examine them, investigating their origin and exploring the best way to deal with them (Chambers et al., 2004).

Due to the normal fluctuation of the companies' activities, it is not easy to define what values are dubious enough to deserve a further investigation from an analyst. Bay et al.

(2006), in their research of anomalies in accounting data, defined suspicious company behaviors as those that deviate from the behavior that company has displayed in the past, diverge from the behavior of the companies that operate in the same sector, or resemble the behavior of companies that, in the past, were found to be abnormal. These authors defined three types of irregularities that can be found in financial statements: unintentional mistakes (such as typing a wrong digit while entering the data), intentional misreporting of the financial statement, or normal value deviations due to the course of business.

2.2. The Outlier Detection Process

Outlier detection is the process through which irregular data points or abnormal clusters of instances are identified (Duan et al., 2009). An important aspect to ensure an effective detection of anomalies is the selection of the adequate method. However, there is no consensus in literature as to what technique will provide the best results, with each having its own advantages and disadvantages. Generally, this choice depends on various factors related to the problem under study, such as the field, the data and the goal that one intends to achieve (Cousineau & Chartier, 2010).

There are three main types of anomaly detection methods, which can be divided according to whether they do or do not require instances with data labels, that is, instances that are previously labeled as being normal or abnormal data points (Chandola et al., 2009; Gao et al., 2006; Goldstein & Uchida, 2016; Görnitz et al., 2013; Su & Tsai, 2011):

- supervised outlier detection – approaches that require the training and test data to be labeled; classifier algorithms are usually used, and the focus is on discriminating the two classes; it is inviable in most cases, seeing as the availability of correctly labeled data is scarce;
- semi-supervised outlier detection – methods that use a labeled training dataset, consisting only of non-irregular data points; the models focus on learning the normal instances, with the intention of then being able to identify the observations that do not conform with the normal behavior of the data; compared to supervised learning, it has the advantage of requiring less data labeling;

- unsupervised outlier detection – techniques that do not require any type of data labeling; the focus is on data characterization and the output is computed based on the data's intrinsic attributes; the algorithms generally assume that abnormal data points are a minority in the population.

The output obtained from the outlier detection process also varies depending on the method chosen. Some algorithms return an outlier score for each data point, representing the extent to which that instance deviates from the rest of the sample. In these cases, it is necessary to define a threshold, which separates the points that should be considered as anomalies and the ones that should not. Other models will return a binary label, indicating explicitly which data points the algorithm considers to be anomalies. With the defined threshold, it is also possible to convert the outlier scores into binary labels (Aggarwal, 2017).

2.3. Isolation Forest

2.3.1. The Isolation Forest Algorithm

The Isolation Forest algorithm was first proposed by Liu, Ting, and Zhou (2008), being introduced as an innovative procedure for unsupervised outlier detection which diverges from the existing approaches. While most of the previously explored model-based procedures, such as classification and clustering methods, detect outliers by identifying the data points that do not satisfy the characteristics of a normal instance, this new technique explores the concept of anomaly isolation. The algorithm is constructed based on two main principles, which are compatible with the general definition of outlier: abnormal data points are a minority in the population, and the attribute-values of the anomalous instances are significantly different from those of the remaining population.

The algorithm uses proper binary trees, which are a class of tree data structures characterized by the fact that each of the decision nodes has exactly two children nodes (Goodrich & Tamassia, 2001). A binary isolation tree is constructed through successive partitions of the sample: at each non-terminal node, an attribute and a split value are selected at random, partitioning the instances into two separate sets, with each set forming a child node. This process is executed until either each instance is isolated in a terminal node, or the tree height limit is reached. Given their characteristics, the anomalous data points are more

prone to isolation and need less partitions to form a terminal node than the remaining observations. Therefore, in an isolation tree, outliers are the instances that have shorter paths from the root node to their respective terminal nodes (Liu et al., 2008).

The process of data partitioning can be visually observed in Figure 1, which portrays the isolation of two points, x_o and x_i.

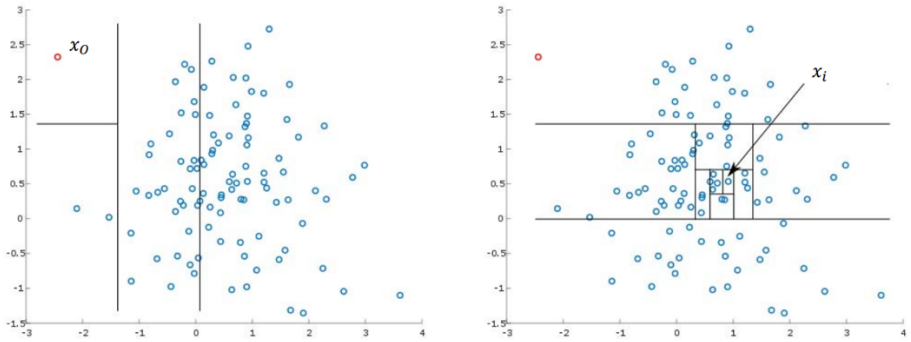

Figure 1 - Isolating an outlier and an inlier through data partitions (Source: Susto et al., 2017).

The point x_o, an anomalous instance, was isolated after just three partitions of the sample, while the point x_i, a normal instance, needed nine partitions. In a tree structure, the outlier point x_o would have a shorter path length from the root to its terminal node than the inlier point x_i (Susto et al., 2017).

Seeing as each tree is created through randomly generated partitions of the instances, the training stage of the algorithm involves the construction of several isolation trees, using a sub-sample of the data, which leads to the formation of an isolation forest. Subsequently, in the evaluation phase, the test instances go through each of the previously built trees, receiving outlier scores which measure their degree of deviation from the remaining sample. The selection of an adequate ensemble size allows for the determination of the instances' expected path lengths, with the highest outlier scores being attributed to the data points that have the lowest average path lengths in the collection of trees (Liu et al., 2008).

Due to its structure, the isolation trees share similarities with Binary Search Trees (BST), with the formation of external nodes in the Isolation Forest algorithm being equivalent to unsuccessful searches in BST. Thus, the average path length in an isolation tree can be estimated from the average path length of unsuccessful searches in BST (Liu et al., 2012), which is given by equation (2.1), where *n* represents the number of instances in the sample and *H(i)* designates the harmonic number estimated by (2.2) (Preiss, 1999).

$$c(n) = \begin{cases} 2H(n-1) - 2(n-1)/2 & for\ n > 2 \\ 1 & for\ n = 2 \\ 0 & otherwise \end{cases} \quad (2.1)$$

$$H(i) = ln(i) + 0.5772156649 \quad (2.2)$$

Then, taking the average value of an instance's path lengths in the trees from the forest, designated by *E(h(x))*, and the value computed in (2.1), the anomaly score *s* returned by the algorithm for an instance *x* in a sample with *n* instances is given by the expression (2.3) (Liu et al., 2012).

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2.3)$$

The outlier scores attributed by the model range from 0 to 1. Instances with low average path lengths will have scores close to 1, being identified as abnormal observations, while instances with scores significantly below 0.5 will be considered normal observations. If all the data points exhibit outlier scores around 0.5, then it can be considered that the sample does not contain any remarkable anomalies (Liu et al., 2008).

2.3.2. Advantages and Further Improvements of the Algorithm

The introduction of the Isolation Forest model brought many benefits to the processes of outlier detection. Its most important advantage is the fact that it can deal with complex problems, being able to handle large volumes of data while maintaining satisfactory time, computational and memory requirements. This is mainly because, unlike other commonly used models, this algorithm does not need to compute measures of distance or density. In addition, in most cases, isolation trees do not have to be grown in full, since the section of the tree that isolates normal instances is irrelevant, allowing the model to reach great results with small sample sizes (Liu et al., 2008).

Since the initial presentation of this model, several improvements have been suggested in literature. Hariri et al. (2019) proposed an extension of this approach, which

allows the data to be partitioned using hyperplanes with any slope, improving the detected bias that the original algorithm possesses.

Xu et al. (2022) pointed out the algorithm's difficulty in detecting outliers in data spaces that are non-linear-separable, an issue that had not yet been addressed in any research. These authors introduced a new version of the model, named Deep Isolation Forest, which uses deep neural networks to create new data spaces, with the data being then separated through axis-parallel partitions. This model allows for a better outlier score attribution in high-dimensional problems than the original method, where the data partitions are exclusively linear.

A new and improved approach, named Majority Voting IForest, was proposed by Chabchoub et al. (2022). Instead of computing the output based on the global decision of the forest, this proposal takes the results of each individual isolation tree and applies the majority vote rules. This method provides similar accuracies to the original approach, while requiring less computational power and time.

Chapter 3 - Methodology

This Chapter introduces and explains the main methodology of this work. Firstly, an overview of the current company selection criteria is made, along with the deliberation of which rules will be replaced by the anomaly detection mechanism. Then, the process of data extraction and its preliminary treatment is presented, followed by the election of features to be included in the model. Afterwards, the diversity of the companies in the sample is addressed. Lastly, the approaches for the practical implementation of the algorithm and for the analysis of the results using Shapley values are described.

3.1. The Current Company Selection Criteria

The IES data quality control process is divided into two main phases: the first one consists of automatic corrections, while the second one requires a manual in-depth investigation by an analyst. The rules behind the automatic corrections were crafted based on past experiences (such as detecting that a specific company makes a systematic reporting mistake every year) and provide a great first approach to improve the quality of the data. For the second step of this process, the manual examination, it is necessary to make a narrow selection of companies, due to the limited resources available.

At the moment, there are 17 criteria in place for the identification of the corporations that must be subject to manual inspection. Through the years, it has been observed that these criteria do not provide the most satisfactory selection, seeing as some reports that are being examined turn out to be error-free (thereby wasting resources), while other companies with significant irregularities are being left out. Therefore, there is a desire to replace the current selection criteria with an anomaly detection mechanism that can provide a better identification of the reports that need to be manually examined.

The work began with a brief analysis of the current selection process, with the intention of deciding which rules can be accurately replaced by outlier detection procedures. Due to the limited time available for the implementation, it was decided to prioritize two main criteria: Thresholds and Funding.

The Thresholds criterion intends to capture the most relevant variations in the companies' reports. There is a preestablished set of 22 indicators (which are shown in Figure 2) and the rules state that a company that presents, for any of the indicators, a value higher than 100 million euros, in the current year or the homologous one, along with a relative variation above 30%, must be selected for further examination. This criterion was chosen for this work because it covers a significant part of the IES data and enables the characterization of the companies from various economic and financial angles, thus aggregating a large and diverse amount of information.

1. Total Assets
2. Fixed Assets, Inventories and Biological Assets
3. Financial Investments
4. Clients
5. Cash and Bank Deposits
6. Other Assets
7. Equity
8. Total Liabilities
9. Obtained Funding
10. Suppliers
11. Other Liabilities
12. Total Income
13. Sales and Services provided
14. Total Expenses
15. Cost of Goods sold and Material consumed
16. External Supplies and Services
17. Payroll Expenses
18. Depreciation and Amortization
19. Interest Expenses
20. Income Tax Expense
21. Earnings before Interest, Taxes, Depreciation, and Amortization
22. Net Income

Figure 2 - Set of indicators used in the Threshold criterion.

In addition to these accounting indicators, it was decided to include in this work the capturing of some differences between the values reported in IES and the values recorded in other databases. The Funding criterion is the rule, currently in place, that catches most of these differences, highlighting companies that present deviations greater than 5 million euros in Bank Loans or 50 thousand euros in Securities when the IES values are compared to information from other sources.

Despite the focus being on the replacement of the two mentioned criteria, other rules will also be indirectly incorporated into the outlier detection model. For instance, when a company registers a remarkable event, such as a merger or an acquisition, the current rules dictate that it needs to be subject to manual inspection. Seeing as these events lead to significant changes in the financial statements, it is expected that most of the companies in this situation will be picked up by the algorithm.

3.2. Construction of Preliminary Dataset

Having determined which company selection criteria are intended to be replaced by the anomaly detection mechanism, the following step of the work was to extract the data needed for the implementation of the algorithm, which was stored across multiple tables in the Central Balance Sheet Database and other internal databases. This subsection describes the process, carried out using Structured Query Language (SQL), that led to the construction of a preliminary dataset.

3.2.1. Data from the IES Reports

The quantitative information collected from the IES reports is stored using a data panel structure, that is, the data corresponding to a specific report submitted by a company spans across multiple observations, one for each of its accounting items. The main attributes that characterize the observations in this table are:

- CompanyID – the identification code that was internally assigned to the company;
- PeriodID – the identification of the period (year) the report refers to;
- Regime – the identification of the accounting regime adopted by the company (out of four possible regimes);
- TagXML - the identification code of the accounting item (the field in the report);
- ReportedValue – the value reported by the company in the field;
- FinalValue – the final value of the field, after automatic and manual corrections (when applicable).

For this work, it was necessary to extract information regarding the reports that were submitted during two different years: 2021, the period that the algorithm will be tested on, and its homologous period, 2020, which will be used to compute some of the variables for the model. Only companies that submitted reports on both periods can be considered for this implementation, which led to the exclusion of 27 665 enterprises that did not have any data for 2021 and 38 278 that did not make a submission in 2020. Thus, the information gathered refers to a total of 468 202 companies, with each being characterized by over 4000 accounting items.

In an accounting setting, a null value reported in a financial statement usually means that the company does not have anything to declare in that field. As long as the main accounting principles are met, it can be considered that an enterprise has a value of zero for the fields that were left blank, therefore, the null values in the collected data were converted to zero. In the case that a company mistakenly reports a blank field, it is expected that the algorithm will be able to detect the value as an anomaly if the error is of significant dimension.

As previously mentioned, the first phase of the IES data quality control procedure, which consists of automatic corrections, will remain in place. In order to draw valid conclusions from this work, it is essential to accurately replicate the conditions in which the selection of companies is made in reality, that is, the data used to train and test the model needs to include the automatic corrections but not the manual ones. The information stored in the columns ReportedValue and FinalValue could not be used without modifications, considering that the first one contains the raw data submitted by companies, before any corrections, while the second one already includes both types of corrections.

Consequently, it was necessary to resort to a different table from the database, which contains information about all the corrections that have been applied to the IES data. Selecting only the corrections that had an automatic source, the information from this table was organized based on combinations of CompanyID/PeriodID/TagXML and then sorted by timestamp. From each combination of these three variables, the observation with the latest timestamp was selected, which contains, in a column named "EstimatedValue", the value that a specific field from a company's report took after the last automatic correction was applied to it. The automatic corrections information was then merged with the IES data that was previously extracted, by executing a left join that matched the records with the same CompanyID/PeriodID/TagXML.

Afterwards, a new column was created and given the name of AdjustedReportedData. The values in this column were set up to be equal to the EstimatedValue in the cases where the financial statement fields suffered an automatic correction, and equal to the ReportedValue in all other instances. The data in this column contains the information reported by companies, adjusted by the automatic corrections (in case they exist), therefore, it replicates the circumstances in which the selection of companies is made and can be used for the implementation of this work.

Following this, using the adjusted reported values that were just computed, the table was transposed so that each report could form an observation, with each unique TagXML comprising a column. The data transposition is displayed in Figure 3, where ARV_{ij_2020} represents the adjusted reported value of the company with $CompanyID_i$ for the item with $TagXML_j$ in 2020, and ARV_{ij_2021} represents the corresponding value in 2021. Due to restrictions in the maximum number of columns SQL allows in a table, and in an effort to shorten the execution time, a preliminary selection of the accounting items was made, reducing the number of fields kept during the transposition phase from over 4000 to only 605.

CompanyID	PeriodID	Regime	TagXML	Adjusted Reported Value	Reported Value	Final Value
$CompanyID_1$	2020	$Regime_{CompanyID1}$	$TagXML_1$	ARV_{11_2020}	….	…
$CompanyID_1$	2021	$Regime_{CompanyID1}$	$TagXML_1$	ARV_{11_2021}	…	…
…	…	…	…	…	…	…
$CompanyID_1$	2020	$Regime_{CompanyID1}$	$TagXML_j$	ARV_{1j_2020}	…	…
$CompanyID_1$	2021	$Regime_{CompanyID1}$	$TagXML_j$	ARV_{1j_2021}	…	…
…	…	…	…	…	…	…
$CompanyID_i$	2020	$Regime_{CompanyIDi}$	$TagXML_j$	ARV_{ij_2020}	…	…
$CompanyID_i$	2021	$Regime_{CompanyIDi}$	$TagXML_j$	ARV_{ij_2021}	…	…

CompanyID	PeriodID	Regime	$TagXML_1$	…	$TagXML_j$
$CompanyID_1$	2020	$Regime_{CompanyID1}$	ARV_{11_2020}	…	ARV_{1j_2020}
$CompanyID_1$	2021	$Regime_{CompanyID1}$	ARV_{11_2021}	…	ARV_{1j_2021}
…	…	…	…	…	…
$CompanyID_i$	2020	$Regime_{CompanyIDi}$	ARV_{i1_2020}	…	ARV_{ij_2020}
$CompanyID_i$	2021	$Regime_{CompanyIDi}$	ARV_{i1_2021}	…	ARV_{ij_2021}

Figure 3 - Representation of the data transposition.

In this new data disposition, each company occupies two observations: one containing the data related to the 2021 financial statement and the other the data from the homologous period, the 2020 report.

3.2.2. Data from Other Sources

With the intention of including in this implementation some of the differences between the data reported in IES and the data recorded in other sources, as mentioned in subsection 3.1, it was necessary to gather additional information from other internal databases, namely regarding Bank Loans and Securities.

Firstly, the data concerning Bank Loans granted in Portuguese territory was extracted from the Central Credit Register (Central de Responsabilidades de Crédito - CRC) database, while the information regarding external Bank Loans was collected from the Balance of Payments (BOP) database. For the Securities aspect, the sum of four types of financial instruments was selected, namely, commercial paper, non-convertible bonds, convertible bonds, and equity securities. The mentioned data was extracted from the Securities Statistics Integrated System (Sistema Integrado de Estatísticas de Títulos - SIET).

The information recorded in these other sources has different frequencies (such as monthly or quarterly), so the data that was collected refers only to the final position of each company at the end of the year (on the 31st of December), in order for it to be directly comparable to the data in IES, which is yearly. The information gathered was then attached to the table created in the previous subsection.

3.2.3. Characterization of Companies

Furthermore, the implementation of this work required some characteristics of the companies to be taken into account, more specifically the institutional sector, the activity sector, and the company size.

The institutional sector of each corporation was identified with the intention of selecting only those that are non-financial institutions, given that companies belonging to other intuitional sectors require a different treatment. At this point, a total of 8999 companies were excluded from the sample due to not being non-financial intuitions.

Afterwards, information regarding the activity sector and the company size of each enterprise was gathered. The activity sectors are codified according to the sections from the Third Revision of the Portuguese Economic Activity Classification (see Annex A - Classification of Activity Sectors for full decoding). Four activity sectors will not be included

in the work since these corporations require a different type of analysis. As a result, 47 companies were excluded from the sample: 35 belonging to sector K, 9 to sector O, 2 to sector T and 1 to sector U. The company size is organized into four categories, codified from 1 to 4, with the following translation: 1 for micro-sized, 2 for small-sized, 3 for medium-sized and 4 for large-sized corporations. The companies that presented missing data for either the activity sector or the business size, which were 15, had to be excluded from the sample since the use of this information will be indispensable in the implementation of the outlier detection mechanism.

The institutional sector did not need to be used past this point, unlike the activity sector and the company size, which will be essential, therefore being appended to the table constructed across the previous steps. The exclusion of companies carried out during this phase left the dataset with 459 141 enterprises.

3.2.4. Overview of Preliminary Dataset

This subsection exposed the main procedures that were performed on the first approach to the data. The preliminary dataset constructed contains the identification of the observations (given by the CompanyID and the PeriodID), the accounting regime the corporations follow, 605 accounting items from the reports (using the values reported by the companies adjusted by the automatic corrections), 3 fields from other sources (national Bank Loans from CRC, external Bank Loans from BOP, and Securities from SIET), and 2 company characterizing variables (the activity sector and the company size).

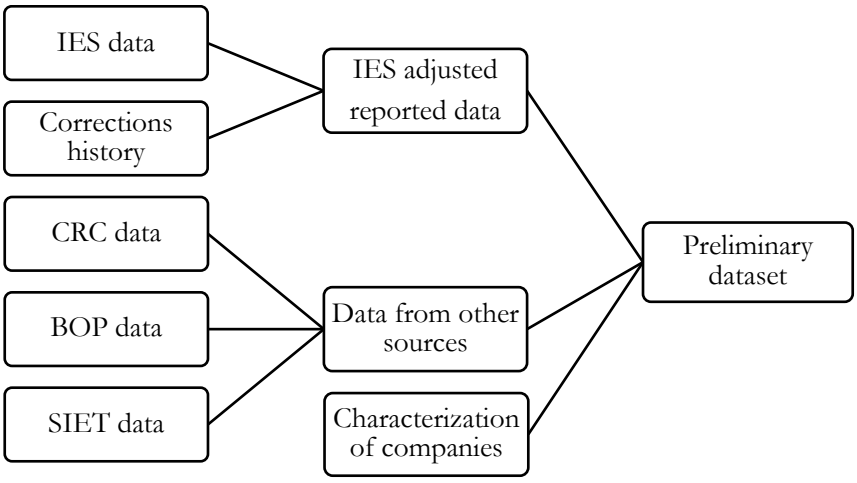

Figure 4 - Overview of the construction of the preliminary dataset.

The procedures that were used to treat the gathered preliminary data, with the objective of forming a final dataset that is suitable for the implementation of the anomaly detection mechanism, are presented across the upcoming subsections.

3.3. Selection of Features

The IES reports comprise large amounts of companies' economic and financial information, spread across thousands of accounting fields. In this sense, an important part of the work was to examine and decide what information must be included in the model and how it should be organized. It is essential that the features used are able to accurately portray the corporations on their various business angles, while bearing in mind that a higher number of variables might compromise the execution time of the algorithm.

The indicators from the Thresholds criterion, previously presented in Figure 2, were used as a starting point for this selection. This initial set of variables was then manipulated, with the intention of studying how the anomaly detection mechanism reacted when features were removed, added, or replaced by some of their constituent parts. The main point noted during this process was that it was advantageous for the variables to have similar levels of detail, seeing as the ones that aggregated large amounts of information (such as the Total Assets or the Total Liabilities) were increasing the volatility of the data. In summary, during this phase, the indicators:

- Financial Investments, Clients, Cash and Bank Deposits, Other Assets, Other Liabilities, Suppliers, Interest paid, Sales and Services provided, Cost of Goods sold and Material consumed, External Supplies and Services, Payroll Expenses, Depreciation and Amortization, Income Tax Expense, Earnings before Interest, Taxes, Depreciation, and Amortization, and Net Income were kept;
- Total Assets was removed and replaced by Tangible Fixed Assets, Investment Properties, Goodwill, Intangible and Biological Non-current Assets, Inventories and Biological Current Assets, Non-current Assets held for Sale, Shareholders, Other Accounts Receivable, and Deferrals;
- Fixed Assets, Inventories and Biological Assets was eliminated (all of its constituents parts had already been added);

- Total Liabilities was cut off and replaced by Other Accounts Payable (non-current), Other Accounts Payable (current), and Provisions;
- Obtained Funding was eliminated and substituted by Bank Loans, Securities, Other Financiers, Capital Participation, and Subsidiaries;
- Total Income was removed and replaced by Other Income (its main constituent part, Sales and Services provided, was already included);
- Total Expenses was cut off and substituted by Other Expenses (all of its other constituent parts were already part of the indicators set);
- Equity was removed and replaced by Equity excluding Net Income, Capital Accounts and Reserves, and retained Earnings.

This process led to the election of 37 accounting items. This set of features gathers the most important information from the financial statements and allows for the corporations to be characterized from various business angles: Non-current Assets and Liabilities, Liquidity, Funding, Activity, and Own Funds.

The values for these variables were generated, for the 459 141 companies in the sample, using the 605 fields that were previously selected and stored in the preliminary dataset. Some features had a direct correspondence to one of those fields, while others had to be computed through the use of predetermined formulas, requiring the employment of multiple fields. In this last case, the accounting regimes adopted by the companies had to be taken into account, considering that the formulas used vary slightly among them.

With the utilization of these features, the outlier detection mechanism will be able to detect if the value that a company's accounting item takes in a specific period is anomalous. However, it was essential to analyze not only the values of these items but also how much they changed when compared to the previous year. Indeed, a corporation might present a normal value for an element but have registered an abnormal change since the preceding period – in this case, the occurrence could be worth analyzing. Therefore, it was decided to include in the implementation the absolute change of each of the 37 features, computed according to the formula (3.1), where $feature\ value_n$ represents the value of the variable in the current year and $feature\ value_{n-1}$ the value in the previous year.

$$feature's\ absolute\ change_n = feature\ value_n - feature\ value_{n-1} \quad (3.1)$$

As an alternative to the absolute change, the use of the features' relative change was considered. However, when the value of a company's accounting item changes from zero, in

the previous year, to a value different from zero, in the current year, it is not possible to attribute a correct finite value to the relative change, which ultimately led to the discarding of this option.

Furthermore, two additional features were included in the model to capture the possible differences between the IES data and the information recorded in other databases. The first one represents the differences in Bank Loans against the data from CRC and BOP and was computed, for each company, by subtracting from the Bank Loans variable the sum of the values from the two other sources, which had previously been collected and stored in the preliminary dataset. The second feature catches the differences in Securities against the data from SIET and was obtained by subtracting from the Securities variable the value of the sum of the four types of financial instruments extracted from the SIET database.

In Table 1, the list of accounting items to be included in the implementation is presented, along with the economic and financial angle covered and the variable code assigned to each element and to the feature containing the respective absolute change. All 76 variables are quantitative and continuous, and the values are represented in euros.

Table 1 - Variables and their respective codes.

Variable name	Variable code	Variable code (absolute change)	Economic and financial angle
Tangible Fixed Assets	V01	V01_chan	Non-current Assets and Liabilities
Investment Properties	V02	V02_chan	
Goodwill	V03	V03_chan	
Intangible and Biological Non-current Assets	V04	V04_chan	
Financial Investments	V05	V05_chan	
Non-current Assets held for Sale	V06	V06_chan	
Provisions	V07	V07_chan	
Other Accounts Payable (non-current)	V08	V08_chan	
Inventories and Biological Current Assets	V09	V09_chan	Liquidity
Clients	V10	V10_chan	
Cash and Bank Deposits	V11	V11_chan	
Other Assets	V12	V12_chan	

Table 1 – (continued).

Variable name	Variable code	Variable code (absolute change)	Economic and financial angle
Other Liabilities	V13	V13_chan	Liquidity
Other Accounts Receivable	V14	V14_chan	
Other Accounts Payable (current)	V15	V15_chan	
Deferrals	V16	V16_chan	
Suppliers	V17	V17_chan	
Bank Loans	V18	V18_chan	Funding
Securities	V19	V19_chan	
Other Financiers	V20	V20_chan	
Capital Participation	V21	V21_chan	
Subsidiaries	V22	V22_chan	
Shareholders	V23	V23_chan	
Interest paid	V24	V24_chan	
Sales and Services provided	V25	V25_chan	Activity
Other Income	V26	V26_chan	
Cost of Goods sold and Material consumed	V27	V27_chan	
External Supplies and Services	V28	V28_chan	
Payroll Expenses	V29	V29_chan	
Depreciation and Amortization	V30	V30_chan	
Other Expenses	V31	V31_chan	
Income Tax Expense	V32	V32_chan	
Earnings before Interest, Taxes, Depreciation, and Amortization	V33	V33_chan	
Net Income	V34	V34_chan	Own Funds
Equity excluding Net Income	V35	V35_chan	
Capital Accounts	V36	V36_chan	
Reserves and retained Earnings	V37	V37_chan	
Bank Loans (differences against CRC and BOP)	V38	-	Differences against other sources
Securities (differences against SIET)	V39	-	

3.4. Partition of Companies into Subgroups

The Portuguese business sector is extremely diverse and, as a result, the population under study is heterogeneous. Implementing an outlier detection mechanism on the sample as a whole is not a viable option since the algorithm would be attempting to compare enterprises with very distinct business characteristics and, therefore, would not be able to correctly identify what constitutes an anomalous value for a specific company. In order to solve the issue at hand, it was necessary to define a partition criterion to form homogeneous subgroups of companies and, after several experiments, it was decided that the best option would be to separate the sample based on business size and on activity sector.

On the one hand, the use of the business size for the partition prevents the model from becoming biased towards identifying large-sized companies as outliers. These enterprises tend to have higher values for most of the accounting items in their financial statements and would be instantly identified as anomalies if the outlier detection procedure was performed on a mixture of companies of different sizes, even though their values might not be abnormal given their business characteristics.

On the other hand, the partition of the companies based on their activity sector allows the model to take into account the differences in their business structures that occur due to the characteristics of the activities carried out, while also assimilating the normal fluctuations that happen over time within each sector. This last issue is particularly evident in the data that is being used in this work to train and test the algorithm: the years 2020 and 2021 were marked by the economic impact of the COVID-19 pandemic, with some sectors being more severely affected than others. It is essential that companies are identified as anomalies only if their behavior deviates from their close peers - for instance, a corporation that sees a large decrease in sales during a specific period should not be considered an irregularity if that same decrease occurred sector-wide.

In this regard, the crossing between the company size and the activity sector was deemed to be the most satisfactory partition criterion, providing an adequate level of homogeneity within each subgroup. The algorithm will be contrasting each company only with those that, simultaneously, have the same size and operate in the same activity sector.

However, this partition left some subgroups with very few observations, especially those of dimension 4, which are a small minority in the sample. The scarcity of information

for these subgroups might severely deteriorate the results of the model, considering that there is not enough data for the algorithm to accurately learn what should be considered an irregularity for these companies. As a result, it was decided to aggregate the subgroups that had less than 20 observations with others of similar characteristics, thus forming subgroups with more information without severely compromising their homogeneity. Within the companies of size 4, sector B (4 observations) was merged with A, sector L (5 observations) was coupled with M, and sectors P, R and S (11, 11 and 3 observations, respectively) were combined with Q.

Table 2 - Number of companies that form each subgroup after the partition.

		Business dimension				Total
		1	2	3	4	
Activity sector	A	17 093	1 424	172	(21) 25	18 710
	B	609	178	27	(4)	818
	C	28 663	9 337	2 389	385	40 774
	D	962	167	63	25	1 217
	E	693	205	94	32	1 024
	F	41 900	6 060	692	65	48 717
	G	92 858	9 839	1 348	234	104 279
	H	20 584	1 751	385	83	22 803
	I	40 959	4 984	488	45	46 476
	J	12 449	1 107	307	90	13 953
	L	42 574	990	93	(5) 59	43 662
	M	46 897	2 499	315	(54)	49 765
	N	14 571	1 492	425	(11) 178	16 666
	P	5 072	689	125	(33)	5 897
	Q	23 873	1 581	188	(11) 58	25 675
	R	8 620	371	67	(3)	9 069
	S	9 280	322	31		9 636
Total		407 657	42 996	7 209	1 279	459 141

This approach led to the formation of 63 subgroups of companies, including the 3 aggregated ones. Table 2 above portrays this partition, presenting the number of observations that form each subdivision.

3.5. Model Implementation

The final dataset, constructed throughout the past subsections, is divided into 63 parts, one for each subgroup of companies. Each fraction is composed of 76 variables (37 accounting items from IES, their respective 37 absolute changes in relation to the homologous period, and 2 variables comprising the differences between the IES data and the information from other sources). The algorithm will run separately inside each subgroup, assuring that each company is only compared with their close peers.

The Isolation Forest model was chosen as the preferred outlier detection mechanism due to its ability to deal with high-dimensional data, while managing to maintain a relatively low execution time and requiring few input parameters. Other Machine Learning unsupervised learning algorithms were considered in earlier stages but were not carried on with due to their higher computational demands, especially when dealing with large quantities of data, and sensitivity to input parameters, which would have needed to be finely calibrated according to the characteristics of each subgroup of companies. The choice of the algorithm was supported by the expertise obtained by the members of the institution during similar past internal projects, which culminated with the implementation of this mechanism due to its excellent results.

The practical execution of this model was performed in Python, taking advantage of its Scikit-learn module. The outlier score returned by the software, for a company x in a subgroup with n companies, is computed according to the formula (3.2), where $s(x,n)$ represents the anomaly score obtained by the formula (2.3), that is, the original score proposed by Liu, Ting, and Zhou (2008). While the authors' scores range from 0 to 1, the ones outputted by this library fall in the interval between -0.5 and 0.5, where instances with values below 0 are considered to be outliers, with the most severe anomalies presenting scores closer to -0.5.

$$final\ score(x,n) = 0.5 - s(x,n) \quad (3.2)$$

The evaluation of the model was a large obstacle throughout the work, given that the data available for training and testing is not labeled. Introducing a classification system to the problem was not a viable option because there was not enough information about the data to properly identify which companies actually do or do not contain irregularities. The possibility of constructing a validation subset of data, using the companies that were manually examined, was initially considered but eventually discarded. These reports were only investigated in the fractions related to the criterion they were selected in so, even if a report was manually examined and did not suffer any automatic corrections, it cannot be concluded with complete certainty that it does not contain irregularities.

As a result, the various formulations of the algorithm, executed throughout this work until reaching the final model, were assessed based on three metrics: the total number of companies with outlier-like scores, the ratio between the sum of the Total Assets of these companies and the sum of the Total Assets of the whole sample, and the ratio between the sum of the Sales and Services provided of these corporations and the sum of the Sales and Services provided of the whole sample. Due to the volatility of these metrics, they were merely used as guidance and not as a means of accuracy.

The number of base estimators of Isolation Forest controls the number of trees to be built in the ensemble. Among the various values tested for this parameter, such as 50, 100, 250, and 500, it was concluded that, all else equal, the higher number of estimators (500) provided moderately more satisfactory results, without compromising the low execution time of the algorithm. During the construction of each tree, the number of observations drawn was the minimum value between 256, which was the default setting suggested by Liu, Ting, and Zhou (2008), and the total number of observations in the subgroup. The contamination parameter represents the proportion of anomalies that are believed to exist in the data, which, in turn, controls the number of observations identified as outliers by the algorithm. Although the modification of this parameter was assessed, and the number of anomalies spotted was used as a guide for the evaluation of the model, it will not be as important in future practical applications of the work developed, since the focus will be on the most severe outliers from the sorted list. The final choice for this parameter was the "auto" option provided by the library, which places the offset value used for the decision function at the middle point between the lowest and the highest possible raw scores.

The standardization of the feature values was attempted but did not provide any significant changes to the results, therefore, it was not applied in the final formulation of the model. This decision was supported by the fact that tree-based algorithms are not sensitive to feature scaling.

Seeing as Isolation Forest is an unsupervised learning model, it was trained and tested using the same dataset, that is, in the first run, the algorithm learns the patterns of the data and, in the second run, it attributes scores to the observations according to what it just learned. The IES data used corresponds to the reports of the year 2021 (along with support from the reports of the homologous period, 2020, for the incorporation of annual changes), which, at the point this work started, was the most recent period with fully available data.

After the implementation of the algorithm, the companies that were identified as probable outliers in each partition of the dataset are sorted according to the anomaly score they were assigned, from the most severe to the least severe outlier, with the intention of providing the analyst who is going to manually examine the reports with a priority list for each subgroup. The scores obtained across the different partitions are not directly comparable and it is not possible to construct an accurate priority list with the companies from all the subgroups, so the examiner must decide the best way to allocate the available resources across the subdivisions.

3.6. Use of Shapley Values

Lloyd Shapley proposed an approach for evaluating the outcome that a player in a cooperative game can expect to obtain from their participation, a solution concept that has been named Shapley value (Dubey, 1975; Hart, 1989; Roth, 1988). This notion has since been expanded to various aspects of Machine Learning, with a particular mention to model explainability, where this concept is applied to determine the contribution of each feature to the outcome (Rozemberczki et al., 2022; Merrick & Taly, 2020).

In the context of this work, it is essential to identify, for the anomalous companies, which sections of the report contain the most critical irregularities so that the examiner can focus their investigation on those fields, thereby economizing the resources that would have

been poorly spent if the whole report had to be analyzed. This proposal can be achieved with the use of Shapley values for model explainability.

Taking advantage of Python's Shap module, the contribution of each feature to the score that was assigned by Isolation Forest to each observation was computed. The features with the lowest Shapley values are the ones that contributed the most to decreasing the score of the model, that is, to push the company towards being identified as an anomaly. The results of this implementation can be visualized graphically, by means of a bar plot or a force plot. Alternatively, the full information about the feature contributions can be extracted, with the intention of easily highlighting the variables that seem to contain abnormal values for each anomalous report and providing the analyst with these findings.

Chapter 4 - Results

This Chapter contains the main results obtained during the practical implementation of this work, along with their discussion. Firstly, the output of the Isolation Forest model is analyzed, followed by the presentation of some interesting details found with the application of Shapley values.

4.1. Model Results

The results obtained from the implementation of the Isolation Forest model are presented in Table 3, disclosing, for each of the 63 subgroups, the number of companies the algorithm assigned an outlier-like score to, the total number of corporations, the ratio between the two previous values, and the score of the most severe outlier.

In total, the model attributed an anomaly score lower than zero to 7 111 companies, which represents 1.55% of the sample under study. This set of companies holds 43.89% of the Total Assets and 18.05% of the Sales and Services provided when compared to the totals of the whole sample.

The proportion of selected corporations was consistent among the subgroups with more observations. Considering only the ones with more than 100 companies, the highest proportion was identified in 2 / I (2.29%), while the lowest one was in 1 / L (1.20%). On the opposite side, the percentages in subgroups with fewer companies varied greatly: while 2 / G and 2 / L presented percentages of only 0.97% and 1.01%, respectively, 4 / A, B and 4 / D had 16.00% each. This aspect reveals how the subgroups with fewer observations, especially those of dimension 4, might suffer from a deterioration of the model due to the fact that the model has less information to learn from.

As for the scores of the most severe outliers, the subgroups 4 / C (-0.336), 4 / H (-0.314), and 4 / J (-0.305) seem to have the most critical irregularities, that is, at least one of the companies in each of these subdivisions deviates greatly from its close peers. On the other end of the spectrum, 3 / R (-0.050), 3 / L (-0.056), and 3 / S (-0.059) are the subgroups where even the most anomalous instances have scores close to zero, that is, no corporations present substantial irregularities.

Table 3 - Results of the Isolation Forest model.

Subgroup (Business dimension / Activity sector)	Number of companies identified	Total number of companies	Percentage of companies identified	Score of the most severe outlier
1 / A	329	17 093	1.92	-0.249
1 / B	8	609	1.31	-0.118
1 / C	485	28 663	1.69	-0.274
1 / D	16	962	1.66	-0.121
1 / E	12	693	1.73	-0.214
1 / F	625	41 900	1.49	-0.267
1 / G	1 314	92 858	1.42	-0.257
1 / H	369	20 584	1.79	-0.289
1 / I	618	40 959	1.51	-0.294
1 / J	183	12 449	1.47	-0.265
1 / L	510	42 574	1.20	-0.230
1 / M	715	46 897	1.52	-0.281
1 / N	228	14 571	1.56	-0.226
1 / P	77	5 072	1.52	-0.253
1 / Q	353	23 873	1.48	-0.257
1 / R	138	8 620	1.60	-0.227
1 / S	135	9 280	1.45	-0.195
2 / A	29	1 424	2.04	-0.202
2 / B	4	178	2.25	-0.168
2 / C	211	9 337	2.26	-0.199
2 / D	7	167	4.19	-0.227
2 / E	3	205	1.46	-0.267
2 / F	134	6 060	2.21	-0.244
2 / G	95	9 839	0.97	-0.145
2 / H	22	1 751	1.26	-0.250
2 / I	114	4 984	2.29	-0.248
2 / J	16	1 107	1.45	-0.159
2 / L	10	990	1.01	-0.186
2 / M	47	2 499	1.88	-0.278
2 / N	27	1 492	1.81	-0.288
2 / P	11	689	1.60	-0.119
2 / Q	33	1 581	2.09	-0.187

Table 3 – (continued).

Subgroup (Business dimension / Activity sector)	Number of companies identified	Total number of companies	Percentage of companies identified	Score of the most severe outlier
2 / R	5	371	1.35	-0.188
2 / S	7	322	2.17	-0.156
3 / A	6	172	3.49	-0.190
3 / B	4	27	14.81	-0.165
3 / C	38	2 389	1.59	-0.159
3 / D	4	63	6.35	-0.118
3 / E	3	94	3.19	-0.099
3 / F	17	692	2.46	-0.257
3 / G	15	1 348	1.11	-0.153
3 / H	9	385	2.34	-0.170
3 / I	8	488	1.64	-0.168
3 / J	7	307	2.28	-0.118
3 / L	3	93	3.23	-0.056
3 / M	9	315	2.86	-0.158
3 / N	10	425	2.35	-0.207
3 / P	5	125	4.00	-0.125
3 / Q	5	188	2.66	-0.183
3 / R	7	67	10.45	-0.050
3 / S	4	31	12.90	-0.059
4 / A, B	4	25	16.00	-0.226
4 / C	8	385	2.08	-0.336
4 / D	4	25	16.00	-0.246
4 / E	4	32	12.50	-0.087
4 / F	7	65	10.77	-0.260
4 / G	7	234	2.99	-0.244
4 / H	4	83	4.82	-0.314
4 / I	4	45	8.89	-0.095
4 / J	5	90	5.56	-0.305
4 / L, M	6	59	10.17	-0.285
4 / N	8	178	4.49	-0.229
4 / P, Q, R, S	6	58	10.34	-0.160
Total	7 111	459 141	1.55	

4.2. Analysis Using Shapley Values

Having applied the Isolation Forest algorithm to the data and obtained the list of anomalous companies, the Shapley values were computed for each of these corporations, with the intention of understanding which variables contributed the most for the observations to be considered outliers. An interesting analysis performed was to highlight the top 5 features that presented the lowest Shapley values, for each company, and analyze which variables appeared within that top 5 for the greatest and the smallest number of anomalous enterprises. The results are presented in Tables 4 and 5.

Table 4 - Features that appeared the most within the anomalous companies' top 5 variables with the lowest Shapley values.

Variable code	Variable designation	Number of companies
V30	Depreciation and Amortization	1726
V18	Bank Loans	1675
V24	Interest paid	1635
V05	Financial Investments	1622
V26	Other Income	1590

Table 5 - Features that appeared the least within the anomalous companies' top 5 variables with the lowest Shapley values.

Variable code	Variable designation	Number of companies
V39	Securities (differences against SIET)	4
V19_chan	Securities (absolute change)	6
V06_chan	Non-current Assets held for Sale (absolute change)	12
V22_chan	Subsidiaries (absolute change)	13
V06	Non-current Assets held for Sale	16

The accounting items Depreciation and Amortization, Bank Loans, Interest paid, Financial Investments, and Other Income seem to be the ones where most companies present severe irregularities, with each having heavily contributed for over 1500 corporations to be identified as anomalies. On the opposite side, the differences in Securities against SIET and the absolute changes in Securities were pointed out as some of the variables that contributed the most to the outlier-like score for only 4 and 6 companies, respectively.

Subsequently, some individual cases were analyzed and compared to the outcome of the 2021 data quality control procedure, which had been carried out previous to the start of this work. Three of these cases will now be presented and, due to data protection reasons, the companies will be designated by X, Y and Z.

Company X, which belongs to the subgroup 2 / I, had been selected for manual examination during the quality control phase and, upon being analyzed, received manual corrections in Bank Loans (V18), Capital Participation (V21), Other Financiers (V20), Tangible Fixed Assets (V01), Capital Accounts (V36) and Equity excluding Net Income (V35).

The Isolation Forest algorithm identified this company as an outlier and assigned a score of -0.214 to it, one of the most severe anomaly scores within its subgroup. Figure 5 below presents the bar plot outputted by the Shap module, with the features sorted from lowest to highest contribution. Out of the 8 variables that contributed the most for this company to be considered an outlier, 5 had received corrections during the examination, thereby supporting the results of the model.

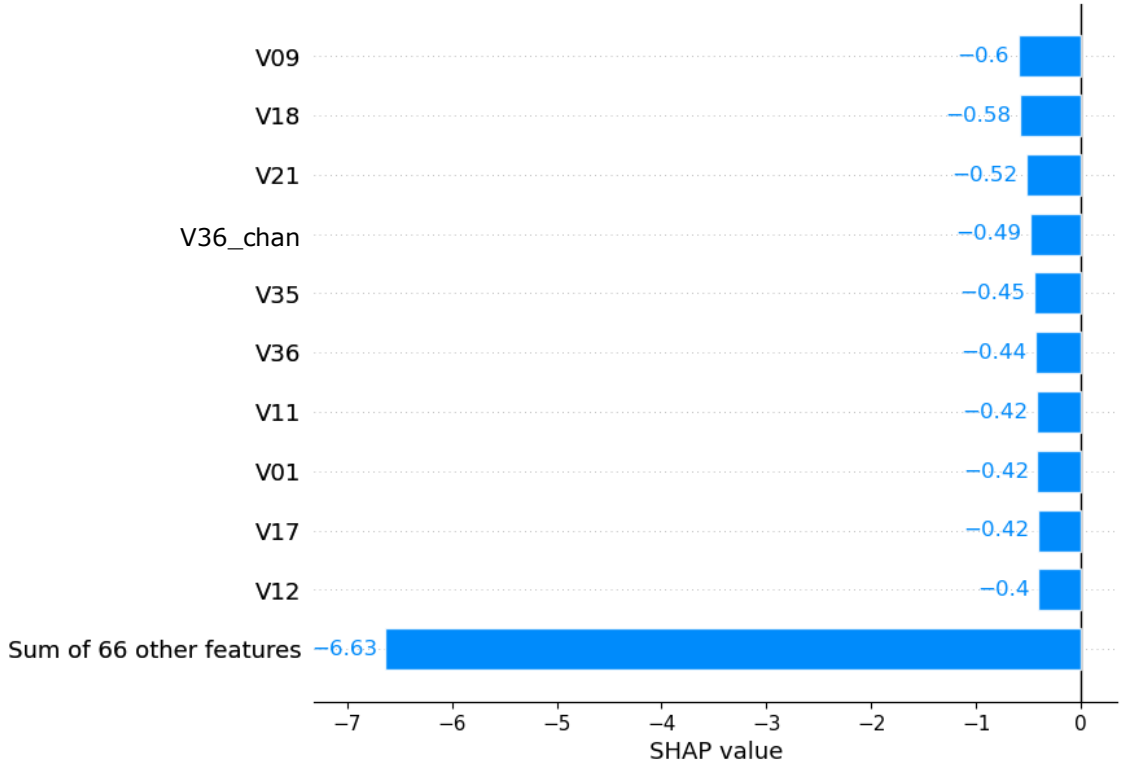

Figure 5 - Shap bar plot for Company X.

Company Y, from the subgroup 4 / C, was selected for manual investigation but no irregularities were discovered by the analyst, so it did not sustain any manual corrections, leading to a waste of resources that could have been spent examining other reports.

The algorithm, in turn, identified this company as an inlier. The Shap force plot for this observation is shown in Figure 6, where the red arrows represent the features that are pushing the company towards being considered an outlier and the blue arrows the ones that are pushing towards an outlier status. With the employment of the model, no resources would have been misused on this company, seeing as almost all of its feature values support the inlier status.

Figure 6 - Shap force plot for Company Y.

Lastly, Company Z, belonging to subgroup 1 / H, was not selected during the data quality control phase and, therefore, was not subject to a thorough manual examination.

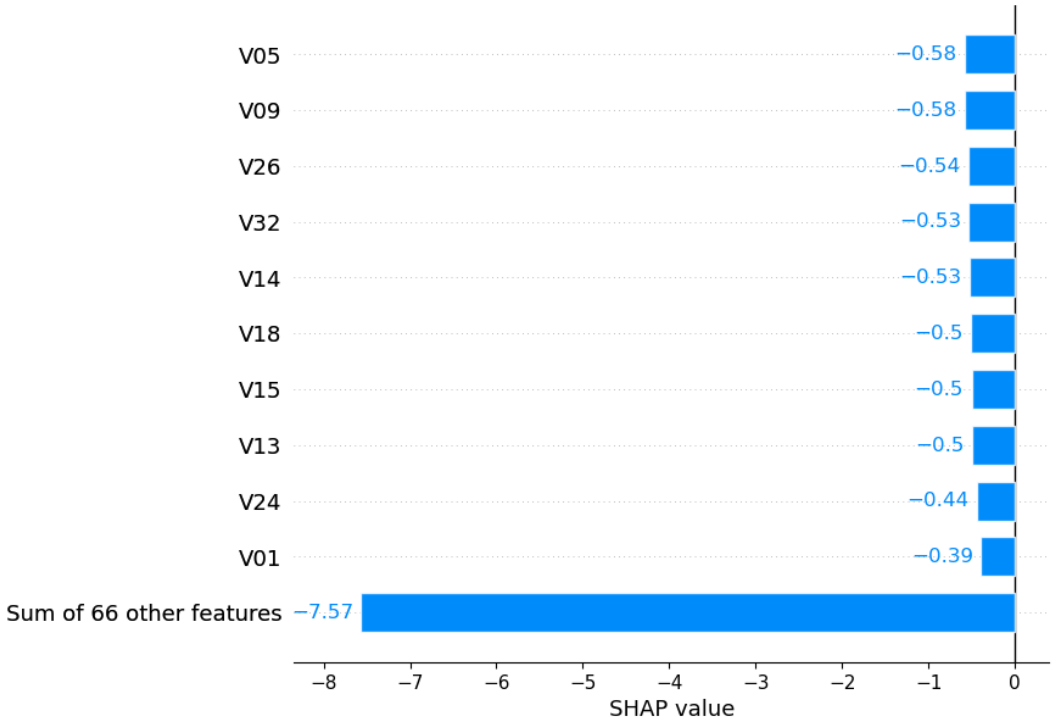

Figure 7 - Shap bar plot for Company Z.

However, the model identified this corporation as an anomaly and attributed a score of -0.274 to it, being one of the most severe outliers in its subgroup. According to the feature contributions, which are shown above in the Shap bar plot of Figure 7, this company might have irregularities mainly in Financial Investments (V05), Inventories and Biological Current Assets (V09), and Other Income (V26). With the implementation of Isolation Forest, this company would have been selected for manual investigation and its irregularities would not have gone unnoticed.

Chapter 5 – Conclusions and Future Work

This work, carried out during an internship at Banco de Portugal, had the main objective of supplying the institution's data quality control process with a new compelling approach. The IES reports, annual financial statements submitted by companies, comprise large amounts of data, which need to be subject to a rigorous controlling procedure in order to ensure that the information collected accurately represents the Portuguese business sector. After the application of automatic corrections, which provide initial treatment to the irregularities in the reports, the companies go through a selection process in order to determine which ones should be subject to further manual examination. Seeing as the resources for this type of inspection are limited, the selection of companies should be as efficient and effective as possible, which led to the desire to replace some of the current selection criteria with an anomaly detection mechanism.

The IES reports contain thousands of fields so the election of the accounting items to be included in the model was a fundamental matter. Using a set of indicators that was part of an in-place criterion as a starting point, various alterations to the collection of features were tested, culminating in a final group of 37 variables, which characterize the companies from their different economic and business angles. In order to account for the magnitude of the changes in the accounting items when compared to the homologous period, the absolute changes of each of the features were added as new variables. Furthermore, 2 additional features that capture the differences between the IES data and the information recorded in other sources were incorporated into the model.

Due to the heterogeneity of the sample, the companies had to be divided into subgroups, ensuring that the algorithm was going to compare each corporation only against their close peers. This partition was implemented using the activity sector and the business size of the companies as the criterion, which, after merging the subdivisions that had few observations, led to the formation of 63 subgroups.

The Isolation Forest was chosen as the preferred algorithm due to its ideal characteristics, namely, the ability to deal with large amounts of data, the low execution time and the small number of parameters required as input. It is an unsupervised learning model, suitable for the issue under study (with unlabeled data) that attributes an anomaly score to

each observation. The evaluation of the algorithm was, however, a limitation throughout this work, given that it was not possible to resort to the usual evaluation metrics.

The final formulation of the model identified 1.55% of the companies as anomalies. This proportion was consistent between the subgroups with more observations but varied a lot among the ones with fewer. The anomalous corporations were, subsequently, sorted according to their scores, with the objective of informing the analyst which reports should be prioritized in the manual examination. Feature explainability, through the use of Shapley values, was then performed, leading to the identification of the features that contributed the most for each observation to be considered an outlier, a convenient approach to advise the examiner on which fields of the report the investigation should focus on.

Overall, although the performance of the algorithm could not be asserted according to evaluation metrics, the results were satisfactory, a fact that was supported by a brief individual case analysis.

In the future, the construction of a labeled dataset, which identifies, for a group of companies, the ones that contain irregularities and the ones that do not, would be useful for the evaluation and possible improvement of the algorithm. The model should also be expanded to include some of the remaining company selection criteria currently in place. Although not all the rules are suitable to be replaced by anomaly detection, some could still be incorporated, which was not done during this work due to the limited time available. Furthermore, Machine Learning algorithms could be implemented in the automatic corrections phase of the quality control procedure, with the intention of improving the prediction of the alterations that need to be made and, therefore, reducing the necessity for a thorough manual examination.

References

Aggarwal, C. C. (2017). *An introduction to outlier analysis* (pp. 1-34). Springer International Publishing.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics, 2*(2), 123-146.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3, No. 1). New York: Wiley.

Bay, S., Kumaraswamy, K., Anderle, M. G., Kumar, R., & Steier, D. M. (2006). Large scale detection of irregularities in accounting data. In *Sixth International Conference on Data Mining (ICDM'06)* (pp. 75-86). IEEE.

Chabchoub, Y., Togbe, M. U., Boly, A., & Chiky, R. (2022). An in-depth study and improvement of Isolation Forest. *IEEE Access, 10*, 10219-10237.

Chambers, R., Hentges, A., & Zhao, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 167*(2), 323-339.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3), 1-58.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3*(1), 58-67.

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research, 168*, 151-168.

Dubey, P. (1975). On the uniqueness of the Shapley value. *International Journal of Game Theory, 4,* 131-139.

Gao, J., Cheng, H., & Tan, P. N. (2006). Semi-supervised outlier detection. In *Proceedings of the 2006 ACM symposium on Applied computing* (pp. 635-636).

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one, 11*(4), e0152173.

Goodrich, M. T., & Tamassia, R. (2001). *Algorithm design: foundations, analysis, and internet examples*. John Wiley & Sons.

Görnitz, N., Kloft, M., Rieck, K., & Brefeld, U. (2013). Toward supervised anomaly detection. *Journal of Artificial Intelligence Research, 46*, 235-262.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21.

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering, 33*(4), 1479-1489.

Hart, S. (1989). Shapley value. In *Game theory* (pp. 210-216). London: Palgrave Macmillan UK.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD), 6*(1), 1-39.

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4* (pp. 17-38). Springer International Publishing.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003). Loci: Fast outlier detection using the local correlation integral. In *Proceedings 19th international conference on data engineering (Cat. No. 03CH37405)* (pp. 315-326). IEEE.

Preiss, B. R. (1999). *Data structures and algorithms*. John Wiley & Sons, Inc..

Roth, A. E. (Ed.). (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H. T., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*.

Su, X., & Tsai, C. L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(3), 261-268.

Susto, G. A., Beghi, A., & McLoone, S. (2017). Anomaly detection through on-line isolation forest: An application to plasma etching. In *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)* (pp. 89-94). IEEE.

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2022). Deep Isolation Forest for Anomaly Detection. *arXiv preprint arXiv:2206.06602*.

Annex A – Classification of Activity Sectors

Table 6 - Classification of Activity Sectors according to the Third Revision of the Portuguese Economic Activity Classification

Code of Activity Sector	Designation of Activity Sector	Included in the implementation
A	Agriculture, farming, hunting, forestry and fishing	Yes
B	Extractive industries	Yes
C	Manufacturing industries	Yes
D	Electricity, gas, steam, cold and hot water and cold air	Yes
E	Water collection, treatment and distribution; sewerage, waste management and environmental remediation	Yes
F	Construction	Yes
G	Wholesale and retail trade; repair of automobile vehicles and motorcycles	Yes
H	Transportation and storage	Yes
I	Accommodation, food services and similar activities	Yes
J	Information and communication activities	Yes
K	Financial and insurance activities	No
L	Real estate activities	Yes
M	Consultancy, scientific, technical and similar activities	Yes
N	Administrative and support service activities	Yes
O	Public Administration and Defense; compulsory Social Security	No
P	Education	Yes
Q	Human health and social work activities	Yes
R	Arts, entertainment, sports and recreation activities	Yes
S	Other service activities	Yes
T	Activities of households as employers of domestic workers and producing activities of households for own use	No
U	Activities of international bodies and other extraterritorial institutions	No