
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union
EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL


# Enhancing data quality controls on money market transactional data

## A comparative study of anomaly detection techniques

**Gianluca Boscariol**[*], João Oliveira Ferreira[*], Matteo Accornero[*]

[*]European Central Bank, Frankfurt am Main (Germany)

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

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union

# Overview

1. Introduction to MMSR data and the Data Quality Management process
2. Anomaly detection pipeline
3. Results and comparison of techniques

# 1

# Introduction to MMSR data and the Data Quality Management process

# ECB's Money Market Statistical Reporting (MMSR)

## What is MMSR?

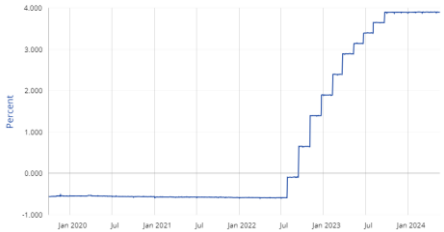
- A daily granular data collection
- Transaction-by-transaction data on the bank's activity in the euro money market
- Data is reported by 47 banks from 10 different euro area countries

## What does MMSR cover?

- Euro-denominated transactions
- Transactions with financial corporations*, general government, and wholesale non-financial corporations
- Transactions with short maturity (up to and including one year)

## What is the data used for?

- monitoring monetary policy transmission and market conditions
- determining the euro short-term rate (€STR)




* Excluding transactions with central banks not for investment purposes

# ECB's Money Market Statistical Reporting (MMSR)

The money market activity can be divided in four segments:

Secured, Unsecured, Foreign Exchange (FX) Swap, Overnight Index Swap (OIS)

**Secured**

- Lending or borrowing in euros in exchange of a security (govt. bonds, …) as collateral
- The biggest segment in terms of transactional volume

**Unsecured**

- Simple lending or borrowing in euros without security collateralisation
- Deposits, call accounts, short-term debt securities, which bear interests at maturity

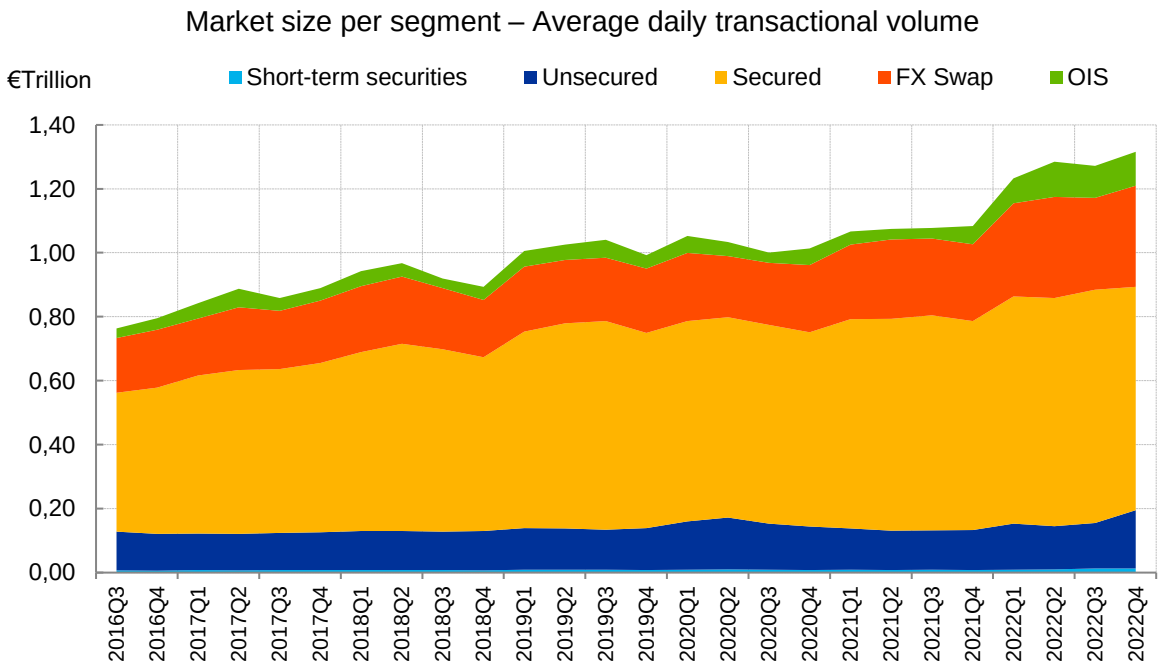

**FX Swap**

- Buying or selling euros against a foreign currency
- The transaction is reversed at a future agreed date
- Main foreign currencies: USD, GBP, JPY, CHF

**OIS**

- Contracts with exchange of daily payments
- One party pays a pre-agreed fixed daily interest
- The other party pays the €STR, the floating rate used as reference

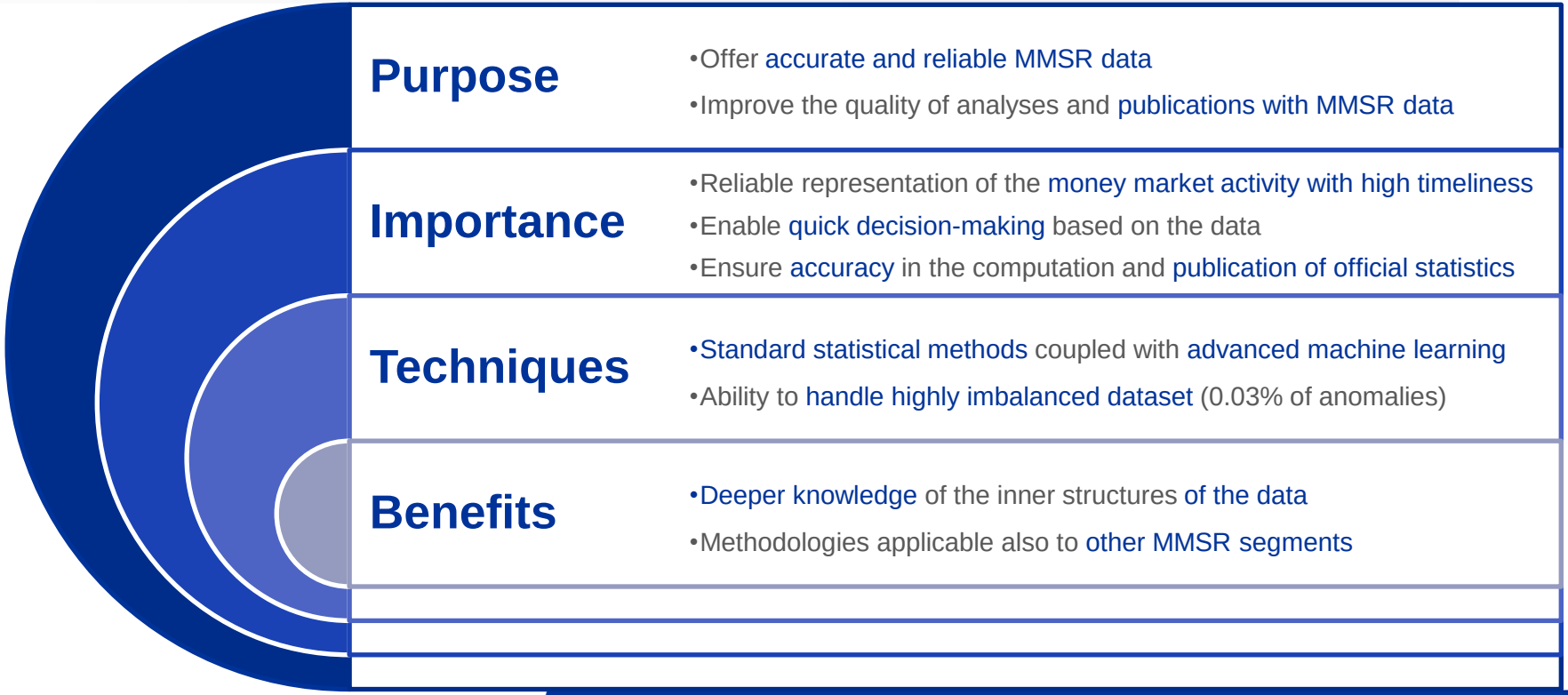
# ECB's Money Market Statistical Reporting (MMSR)




- More than 80,000 daily total transactional records:
  - ~50,000 Secured,
  - ~20,000 Unsecured,
  - ~10,000 FX Swap and OIS
- Money market activity has steadily increased since 2016
- Quarterly aggregate volume has almost reached €1.5 trillion

From the 2022 Euro Money Market Study - ECB
Sources: CSDB for the STS segment and MMSR for the other four segments.

# MMSR Data Quality Management process

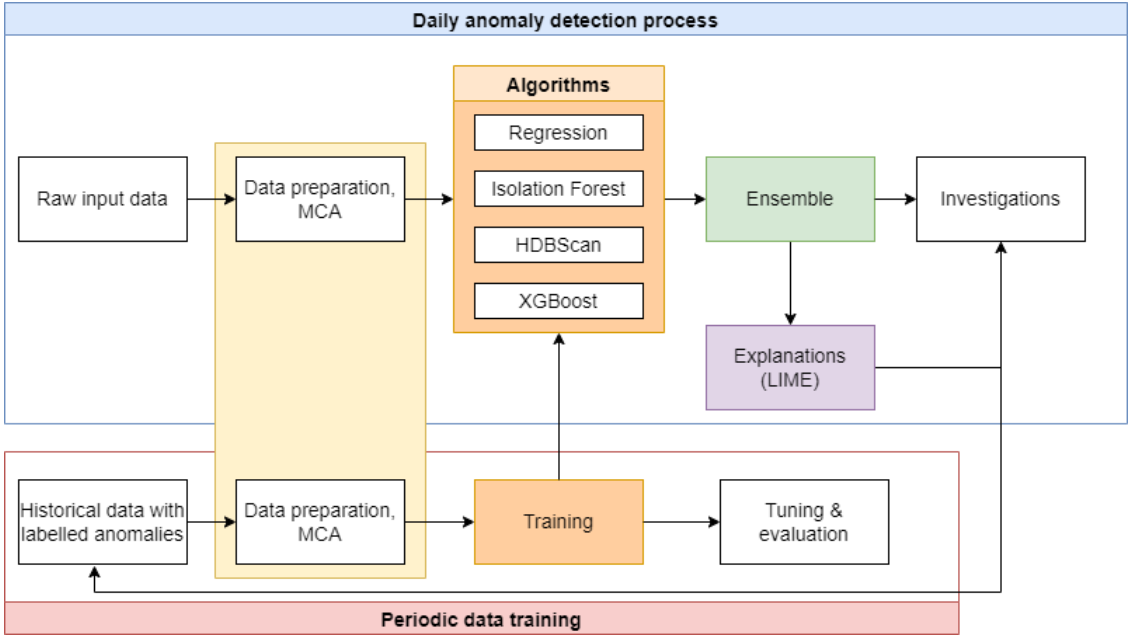


# 2

# Anomaly detection pipeline

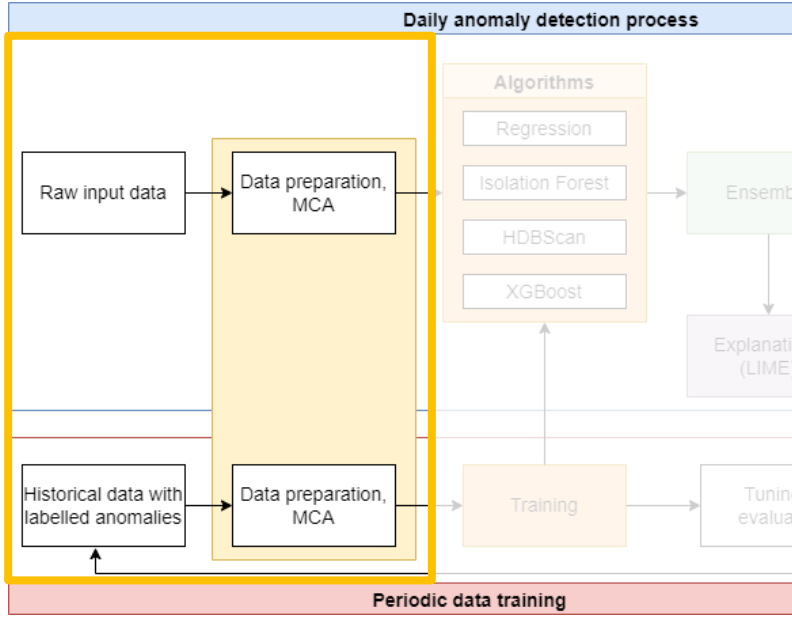
# A pipeline for daily anomaly detection

- **Data preparation**
- **Application of algorithms**
- **Ensemble**
- **Explanation**
- **Feedback loop**
- **Model training**



# A pipeline for daily anomaly detection: data preparation

- **Input data**: the MMSR data received in the early morning of the same day is the first input of the pipeline
- **Training**: labelled anomalies data from MMSR is used for periodic supervised model training
- **Enrichment**: reference rates, counterparty information, …
- **Preprocessing**: categorical variables are transformed into numerical variables using MCA (multiple correspondence analysis)



# A pipeline for daily anomaly detection: regression analysis

- **Regression**: anomalies are defined as transactions far-off the prediction of a model

Spread

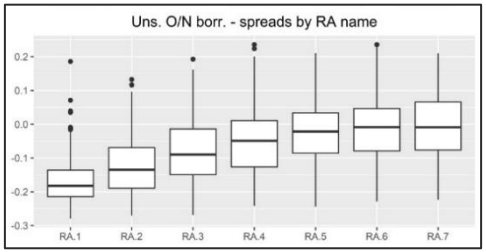
Transactional nominal amount (volume)

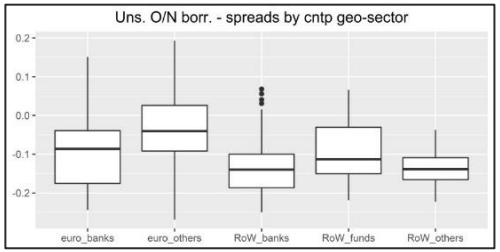
$$s_i = \alpha + \boldsymbol{g}_i'\boldsymbol{\beta} + \boldsymbol{m}_i'\boldsymbol{\gamma} + \delta \log(vol_i) + \varepsilon_i$$

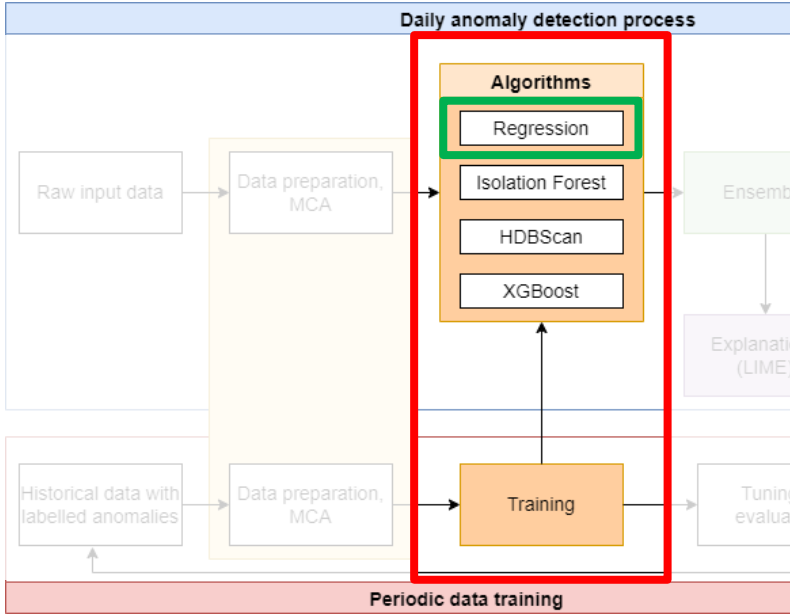
Geographical sector

Contract maturity

Residuals



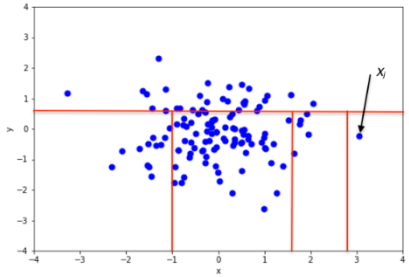



# A pipeline for daily anomaly detection: unsupervised

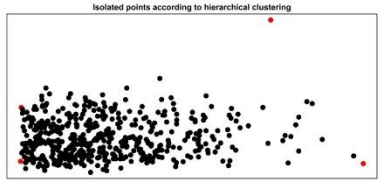
## Isolation Forest

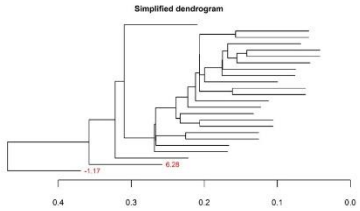
- Repeated random partitioning of the data to isolate observations
- An observation that can be isolated with few partitions is more likely to be an anomaly

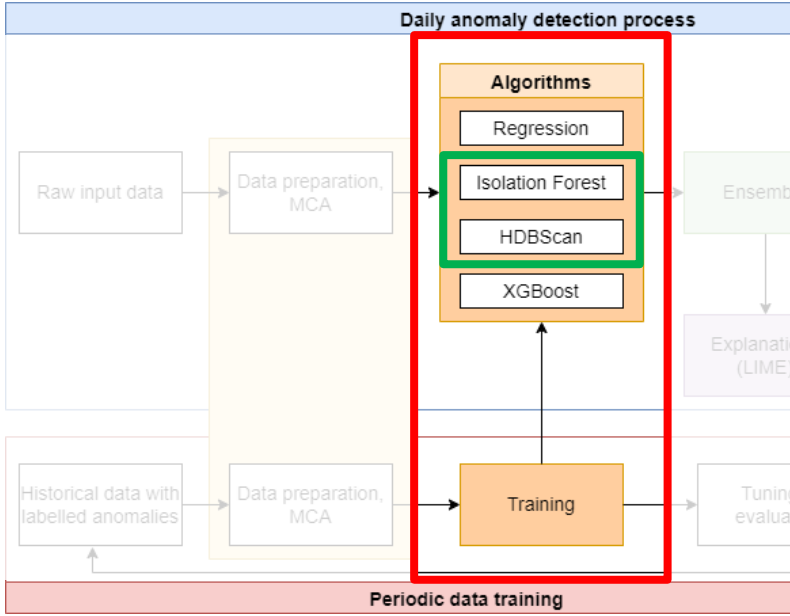



## Hierarchical Clustering

- HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise
- Identifies data points isolated and poorly connected to other data points



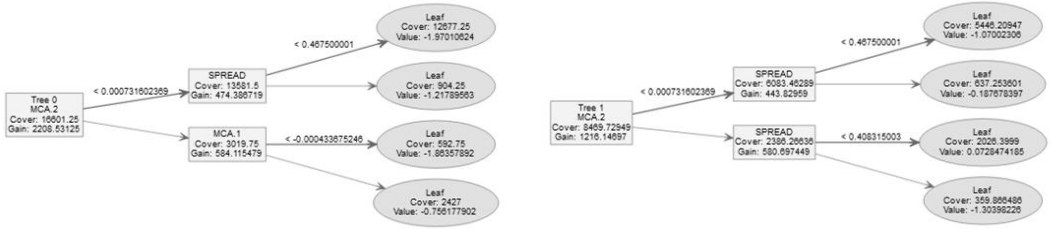



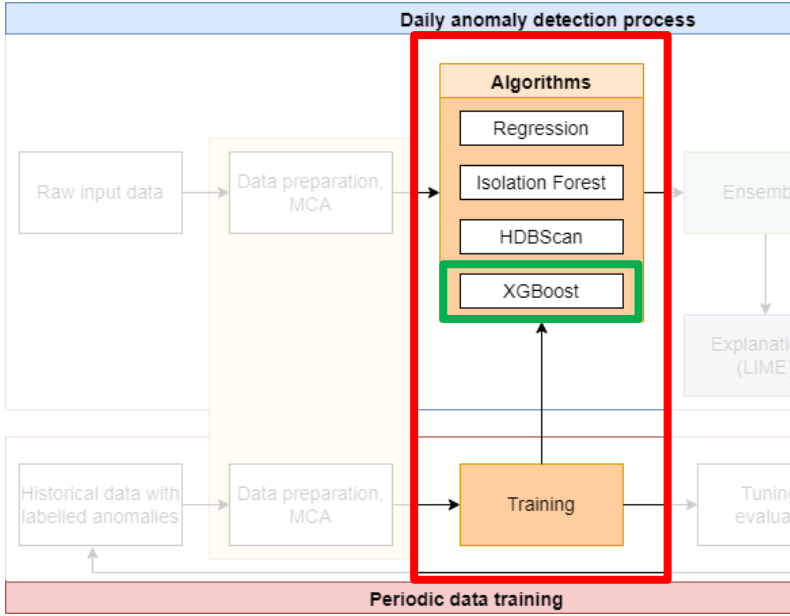



# A pipeline for daily anomaly detection: supervised

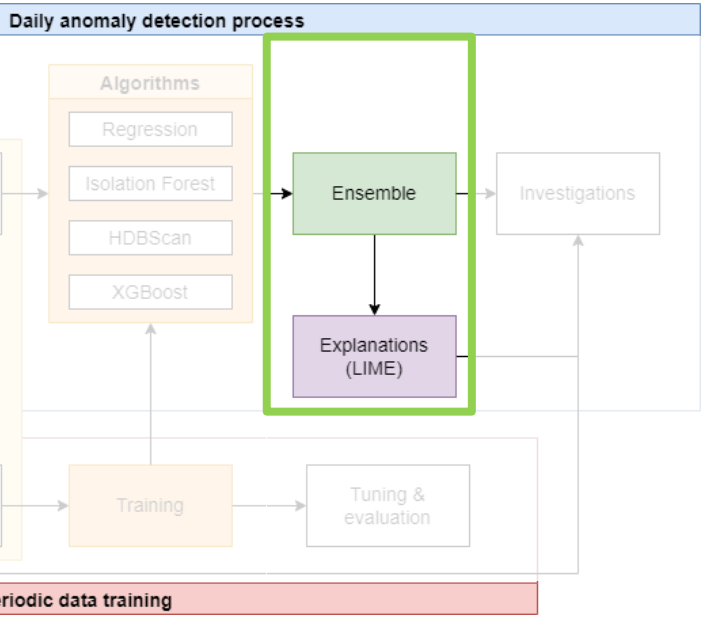
**XGBoost**: eXtreme Gradient Boosting

- Ensemble of multiple weak predictors (decision trees) to build a strong classifier
- Iterative improvement: subsequent addition of decision trees reduces the prediction error from previous weak predictors
- Anomalies are identified based on a training on past data



# A pipeline for daily anomaly detection: explaining outliers

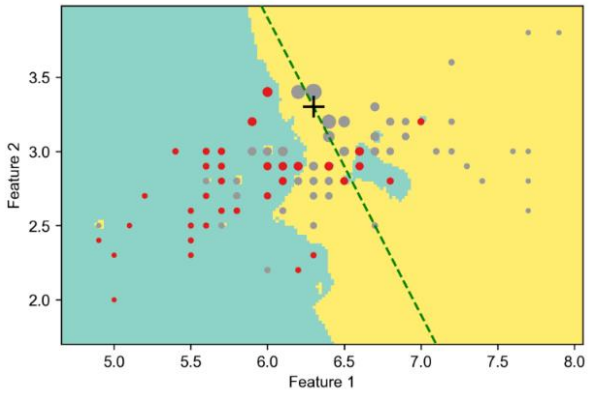



## Ensemble

- the results from the algorithms part are gathered and some outliers are selected for the daily investigations

## Explaining anomalies

- Investigations are better conducted with an insight of which reported field may be the erroneous ones
- LIME* works as a surrogate regression model, making assumptions on which features contributed the most to the outlying nature of the observation



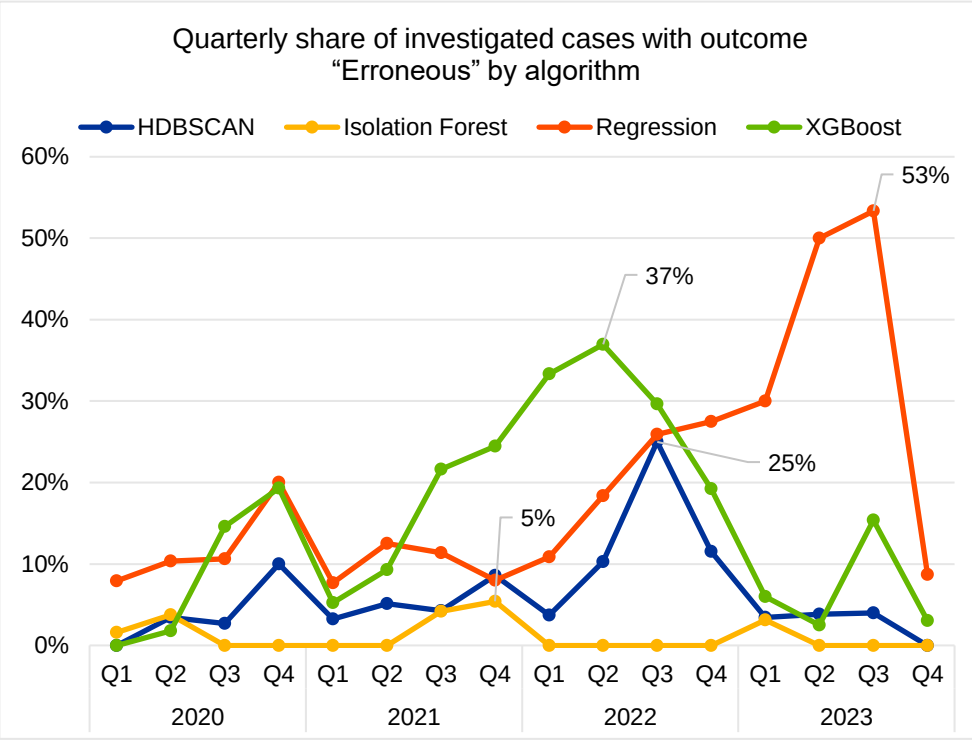

* Local Interpretable Model-agnostic Explanations

# 3

# Results and comparison of techniques

# XGBoost and Regression are the best performers when considering share of detected erroneous trades

- **XGBoost** and **Regression** perform generally better than HDBSCAN and Isolation Forest
- The **Regression** seems to be responsive to erroneous reporting that occurred during ECB key rates hikes
- Reporting agents learn from their mistakes, increasingly reducing anomalous transactions



# Future developments

- **Rationalisation of applied methods**, focusing on regression and XGBoost
- **Improvement of supervised learning** method, by starting from scratch with the model construction
- **Enhancement of data pre-processing** by leveraging additional data sources

# EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

## 2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

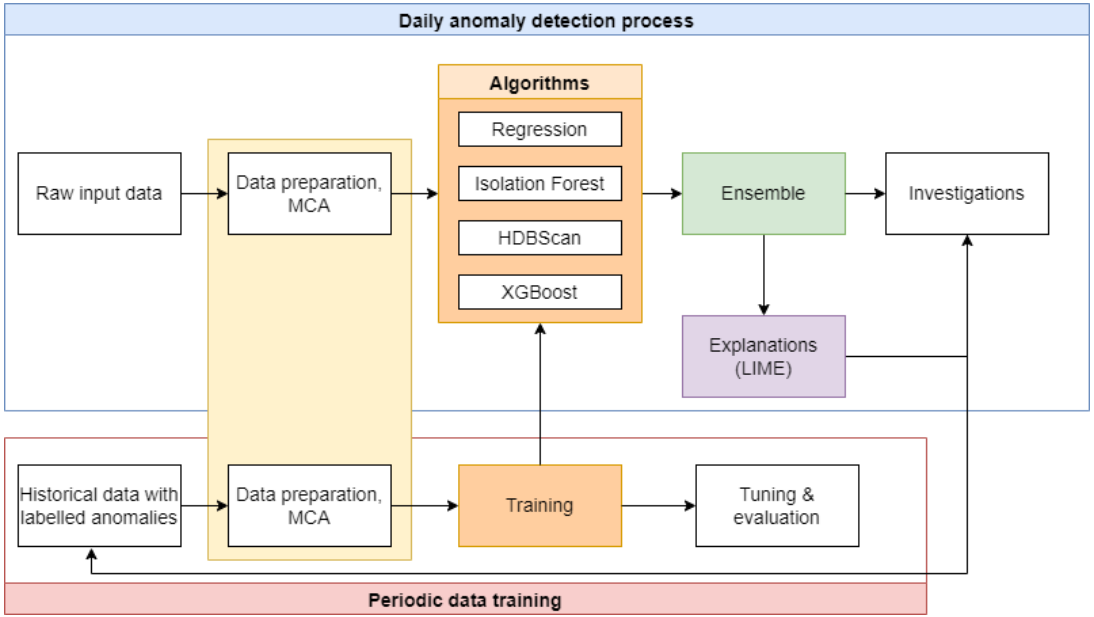
eurostat

The conference is partly financed by the European Union

# Reserved slides

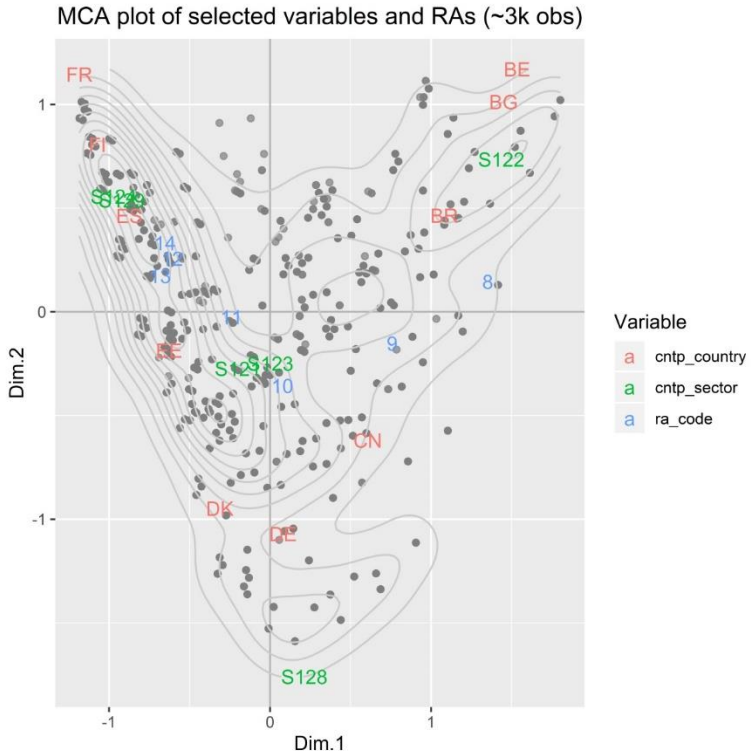
# A pipeline for daily anomaly detection

- **Data preparation**: data enrichment, preprocessing of categorical variables, incorporation of reference rates
- **Application of algorithms**: each method assigns a score to the input data
- **Ensemble**: harmonization of scores and selection of top-scoring observations to be investigated
- **Explanation**: a human-readable explanation of the anomaly is prepared with LIME
- **Feedback loop**: investigations' results build the dataset with labelled data for training purposes



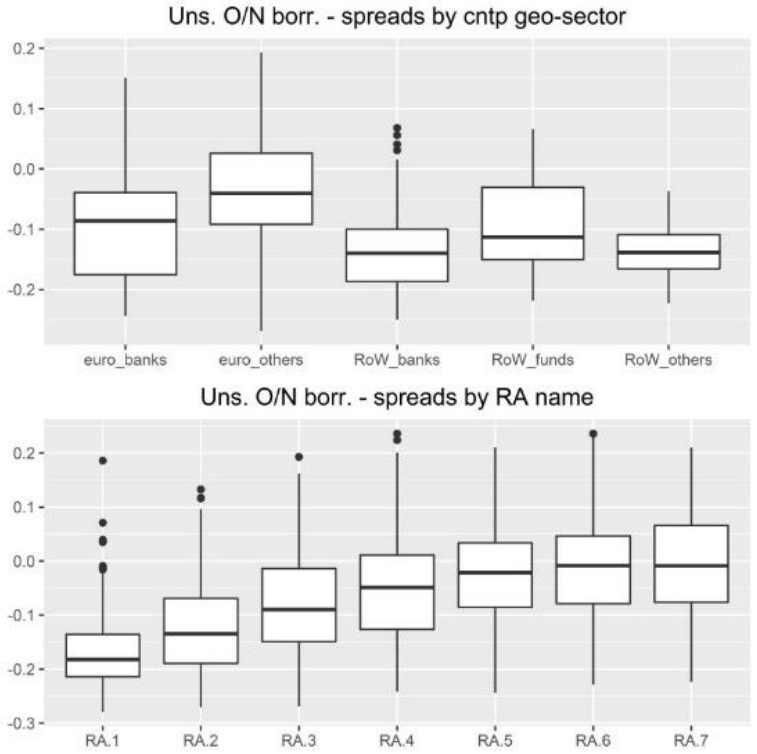
# Multiple Correspondence Analysis (MCA)

- A data transformation before the application of ML techniques
- Categorical variables raise problems for ML algorithms: one-hot encoding is too expensive on resources
- MCA is used to convert categorical variables into numerical values
- MCA exploits the “correlation” between features represented in different categorical variables
- The obtained numerical variables represent observations in a multidimensional space where frequently associated features appear clustered together



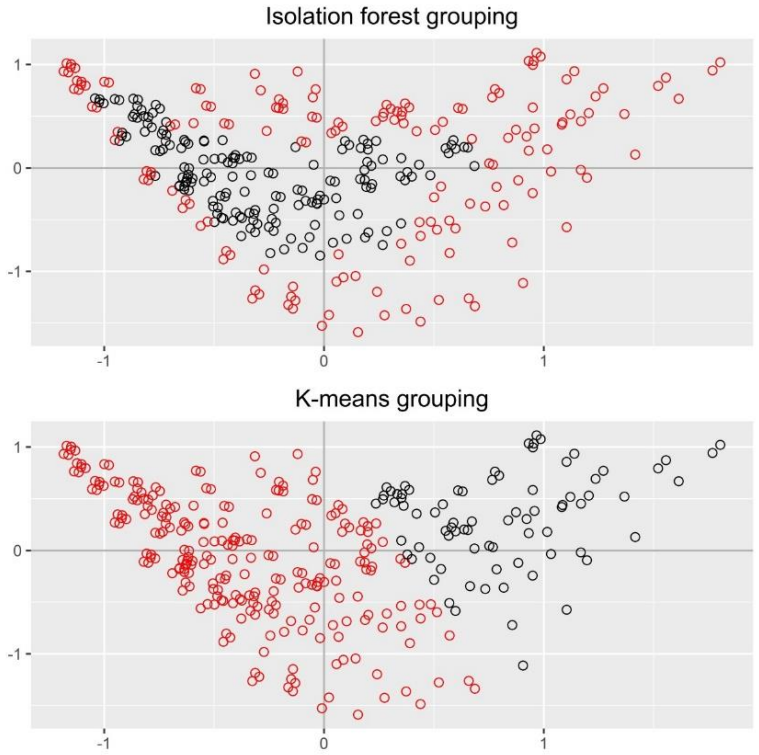
# Regression analysis

- Model based: anomalies are defined as transactions far-off the prediction of a model
- Model: $s_i = \alpha + \boldsymbol{g}_i'\boldsymbol{\beta} + \boldsymbol{m}_i'\boldsymbol{\gamma} + \delta \log(vol_i) + \varepsilon_i$ based on descriptive evidence of typical trading pattern
- Dependent variable: spread between deal rate and benchmark rate
- Explanatory variables: geographical sector (*g*), maturity (*m*), and log of transactional nominal amount (*vol*)
- Estimation: weighted least squares
- Anomalies: transactions having the highest studentized residuals



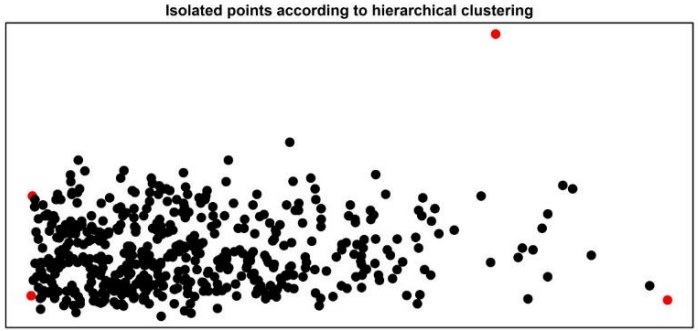
# Unsupervised method: Isolation Forest

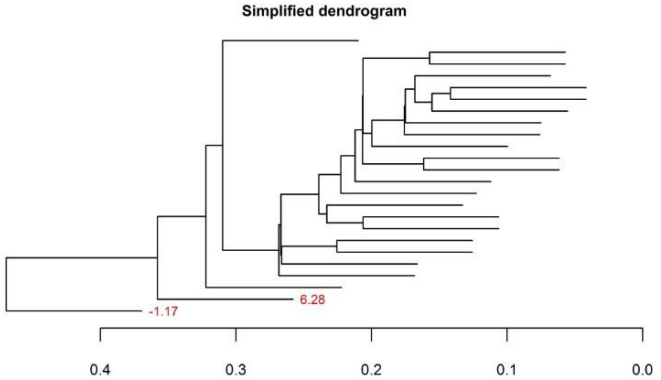
- Isolation forest only works with numerical variables
- It consists in a repeated random partition of the data until all data points in the sample are isolated
- Data points are considered anomalies when the number of partitions required for their isolation is small
- Advantages:
  - It has low linear time complexity and a small memory requirement (it samples)
  - Identifies both scattered and clustered anomalies
  - Sub-sampling, which makes it robust to "swamping" and "masking", i.e., to false positives



# Unsupervised method: Hierarchical clustering

- Hierarchical clustering identifies data points isolated and poorly connected to other data points
- Algorithm used: HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise
- Features:
  - Performance (limited complexity)
  - Parsimony: minimum cluster size is the only parameter
  - Robust to “chaining phenomenon” and other drawbacks of single-linkage
  - Outputs a GLOSH score (Global-Local Outlier Score from Hierarchies) that can be used to identify anomalies



# Supervised method: XGBoost

- In XGBoost, an ensemble of weak predictors are employed to solve classification problems
- Supervised learning: anomalies are identified based on a training on past data
- Being a supervised learning algorithm, it requires a dataset of labelled anomalies to be trained with
  - The quality of the labelled dataset plays a key role in the algorithm accuracy
  - Thanks to the employed pipeline, the labelled dataset is updated regularly
- Advantages: award-winning algorithm (Kaggle), excelling in both efficiency and accuracy