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

19-22 October 2021, Rome / virtual event

Machine learning for anomaly detection in datasets with categorical variables and skewed distributions[1]

Matteo Accornero and Gianluca Boscariol,
European Central Bank

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

EUROPEAN CENTRAL BANK

EUROSYSTEM

Machine learning for anomaly detection in datasets with categorical variables and skewed distributions

IFC-BIS Banca d'Italia
22/10/2021

Matteo Accornero & Gianluca Boscariol
European Central Bank – DG Statistics

Overview

1. Introduction and motivation
2. Anomaly detection algorithms used
3. Workflow and pipeline plumbing interventions

1

Introduction and motivation

Background

- MMSR (Money Markets Statistical Reporting) is a granular (transaction by transaction) dataset collecting data on money markets with daily frequency in the euro area
- In 2018 the ECB embarked in a new anomaly detection project to support the data quality checks connected with the euro short-term rate (€STR) production.
- In development phase, several challenges were identified, in particular with regards to:
  - Workflow: feedback loop needs an effective and sustainable flow of information among involved parties
  - Performance: daily data requires timely data quality information flows, quick reaction times, a lean process
  - Categorical variables: MMSR has few numerical variables, which poses challenges to analysis
  - Skewness: similarly to other granular financial datasets, the *exceptions* are the *rule*
  - Interpretability: how to guide users through the results, especially when using multitude of categorical values
  - Ensemble: results from multiple algorithms need to enter a single data quality process pipeline
- In 2019-20 the MMSR team tried to systematically tackle the issues encountered by extending and modifying existing functions, generalising the solutions identified

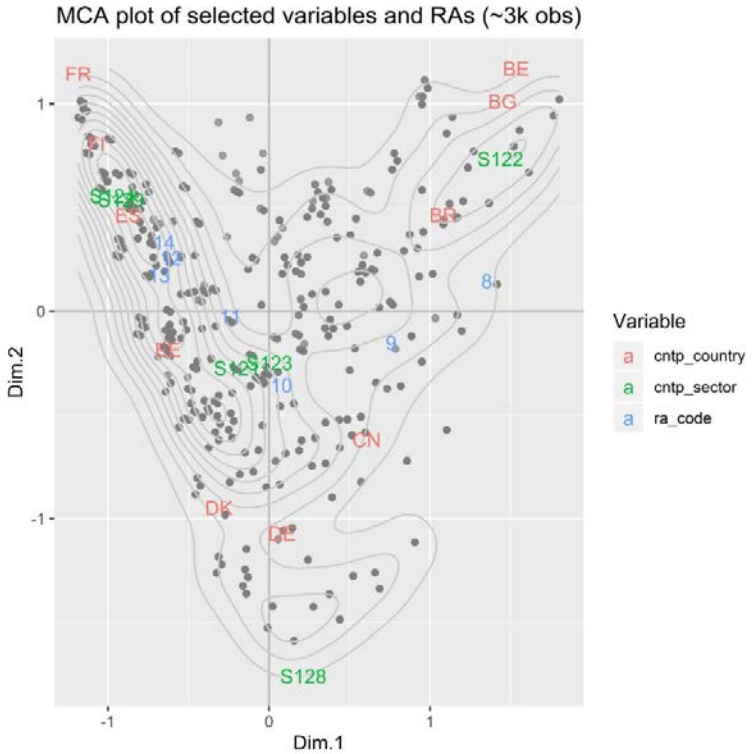
MMSR data quality management – an overview

- The Money Market Statistical Reporting (MMSR) is a daily data collection involving 47 banks, located in 10 different euro area countries
- Reporting agents report transaction-by-transaction data of their euro money market activity
- ~50,000 total transactional records received on a daily basis: ~30,000 Secured, ~15,000 Unsecured, ~5,000 FX Swap & Overnight Index Swap
- MMSR data are enriched with reference data (extra categorical values)
- 4 national central banks + ECB participate in data quality management
- Daily workflow implies limited budget of data quality inquiries
- Structured feedback: labelled anomalies datasets for training available

2

Anomaly detection algorithms used

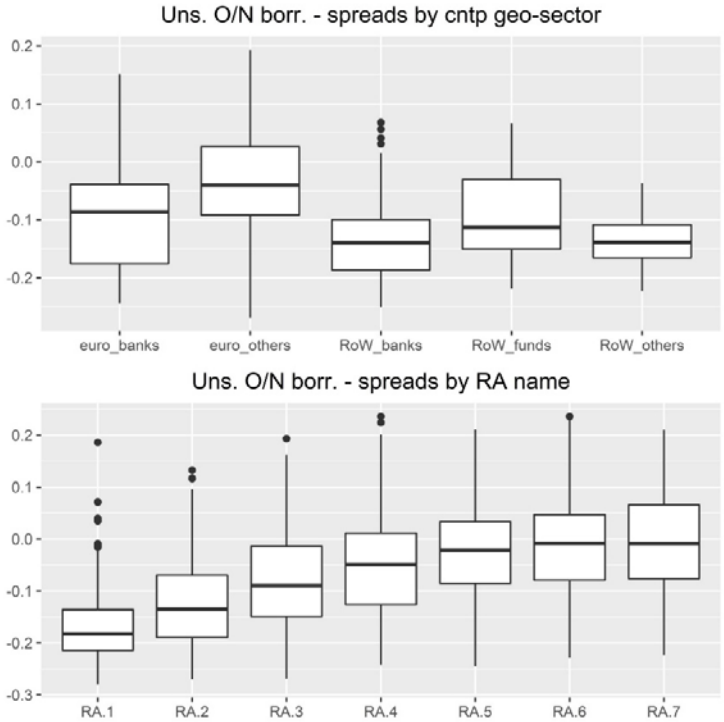
Multiple correspondence analysis (MCA)



Note: Illustrative analysis performed using synthetic data

- A data transformation previous to the application of ML techniques
- Categorical variables raise problems for ML algorithms
- 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

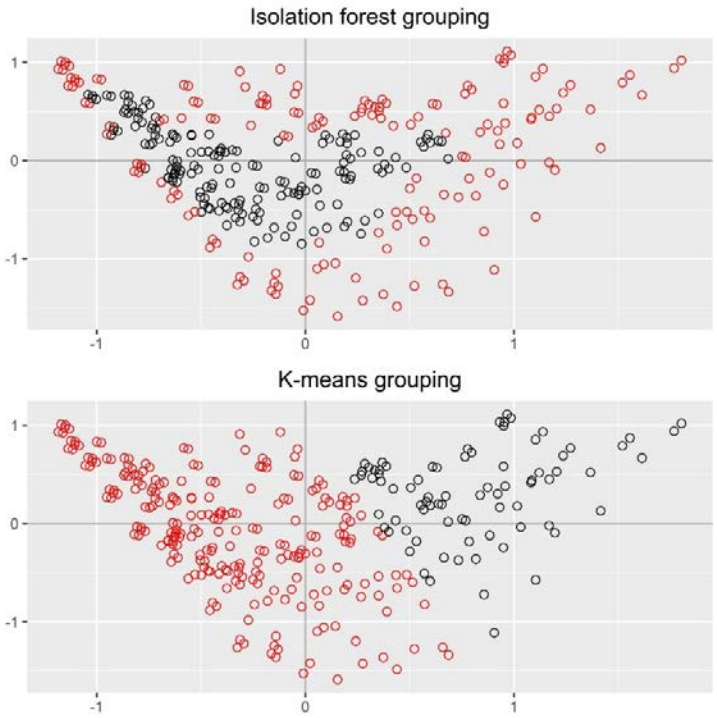
Anomaly detection – regression analysis



Note: Illustrative analysis performed using synthetic data

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

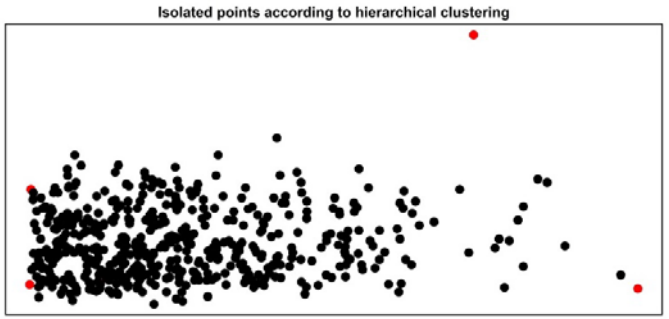
Anomaly detection – isolation forest

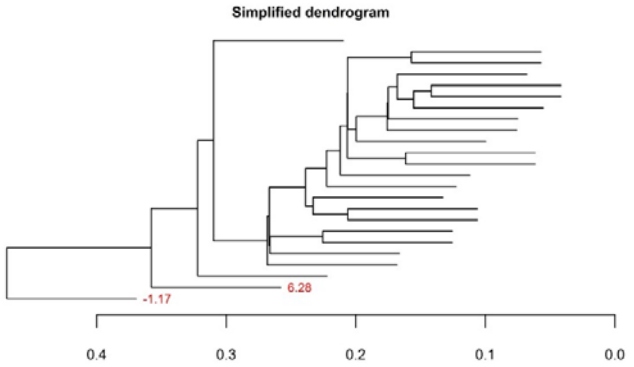


Note: Illustrative analysis performed using synthetic data

- 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
  - It is robust to "swamping" and "masking"

Anomaly detection – hierarchical clustering





Note: Illustrative analysis performed using synthetic data

- 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
- Advantages:
  - Performance (limited complexity)
  - Parsimony in parameters (minimum cluster size is intuitive)
  - Robust to “chaining phenomenon” and other drawbacks of single-linkage
  - Association with GLOSH (Global-Local Outlier Score from Hierarchies)

Anomaly detection – XGBoost

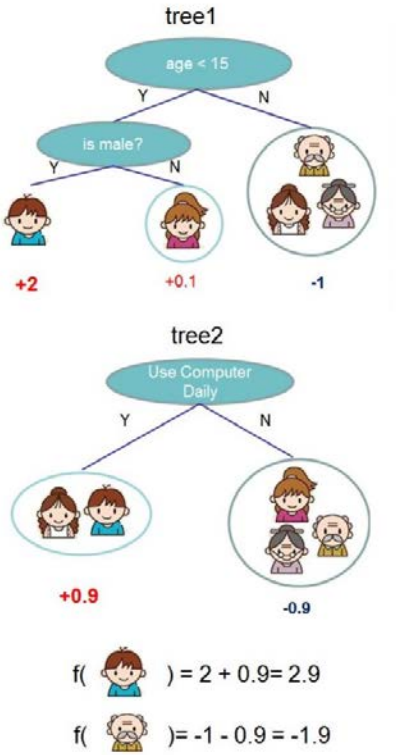
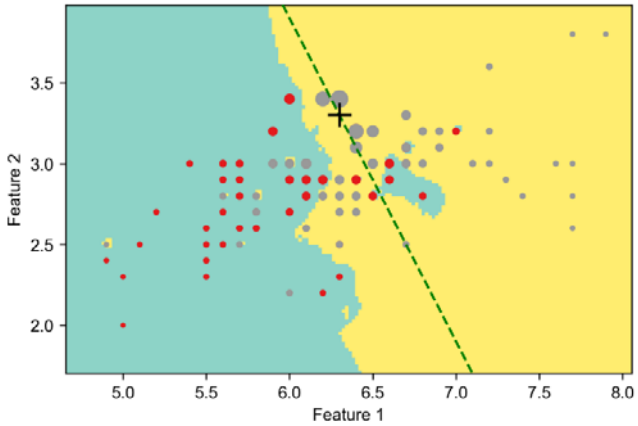


Figure from XGBoost: Chen Guestrin 2016, A Scalable Tree Boosting System. (Does the person like computer games?)

- In XGBoost weak predictors are employed to solve classification problems
- Anomalies are identified on the basis of a training on past data (supervised learning)
- Being a supervised learning algorithm, it requires a dataset of labelled anomalies to be trained with
- Advantages: award-winning algorithm, excelling in both efficiency and accuracy
- Success of the algorithm relies in the quality of the labelled anomalies dataset
- Necessity of structured and ordinate feedback from RAs for enquiries regarding outlying observations
- The verification workflow, integrated with the MMSR DQM aims at “automatizing” and simplifying the internal communication and the storage of information (including pre-defined set of feedback options)

Explaining detected anomalies: LIME algorithm

- LIME (Local Interpretable Model-agnostic Explanations) is an algorithm for the explanation of black-box algorithms results
- LIME is a model-agnostic method: it is equally applicable to every model.
- LIME provides approximated results: for this reason it is reasonably quick, but also somewhat volatile.



Graphical intuition of how LIME works

The black-box model f (unknown to LIME) is represented by the green/yellow background.

The bold black cross is the instance being explained.

LIME samples instances, gets predictions using f and weighs them by the proximity to the instance being explained (represented here by size of dots).

The dashed line is the learned explanation that is locally (but not globally) faithful.

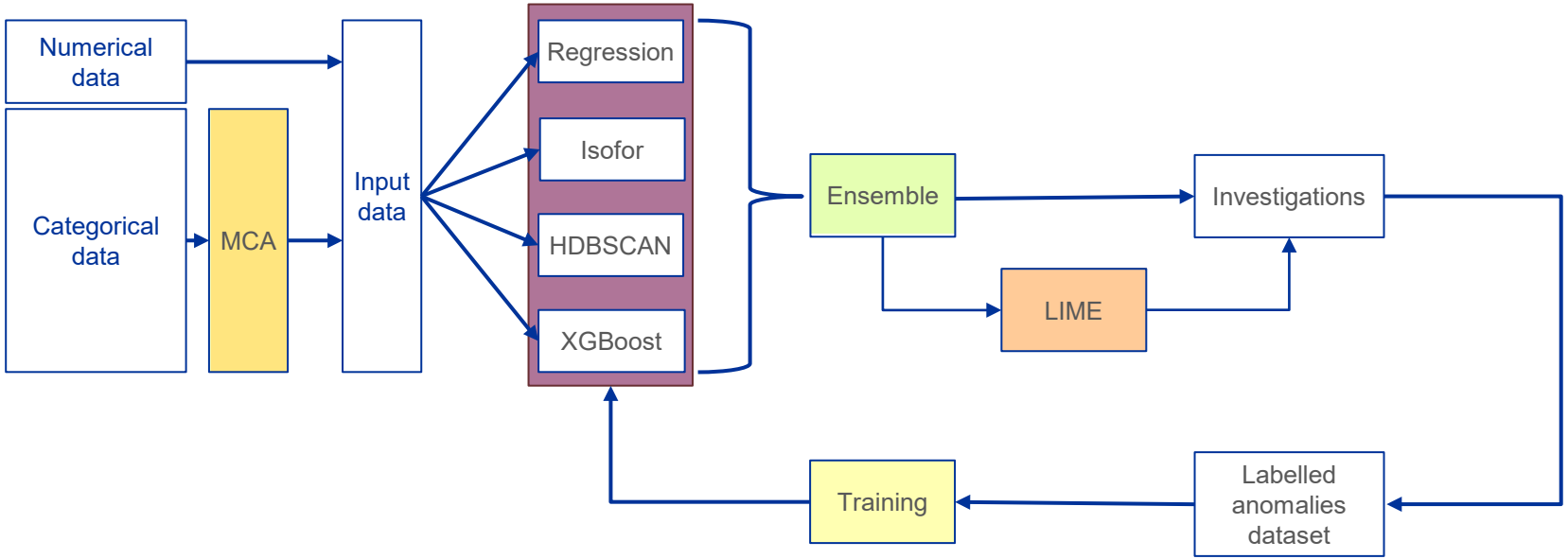
Note: Illustrative analysis performed using synthetic data

3

Workflow and pipeline plumbing interventions

Workflow definition

- Workflow defined to accommodate feedback loop and ensemble of algorithms



Improving MCA performance

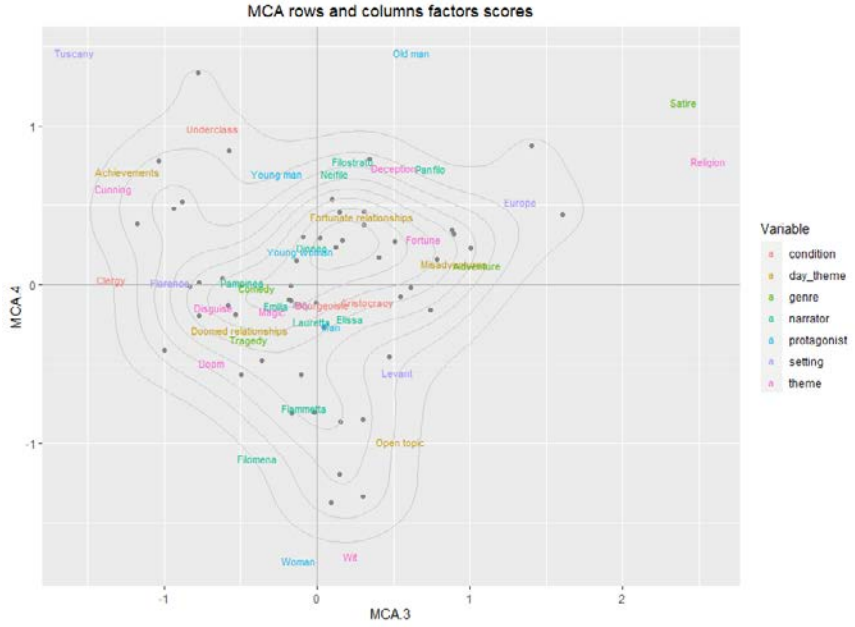
- Reduced weight of MCA objects by exploitation of *conversion formula*[(1)]:
  - Available MCA functions applied to *n* x *k* matrices produce MCA objects including elements having *n* observations
  - This increases extremely the size of these objects when MCA is applied to large datasets
  - MCA in our setting is used to convert *rows* into *row factors*, so there is no interest in column factors
  - Consequently, in the solution adopted all n-dimensional elements are dropped from the stored object
  - This speeds up the saving and the loading in memory of MCA objects
  - The prediction of row factors related to the daily serving input dataset can be then obtained with highly increased speed because of the reduced size of MCA objects:

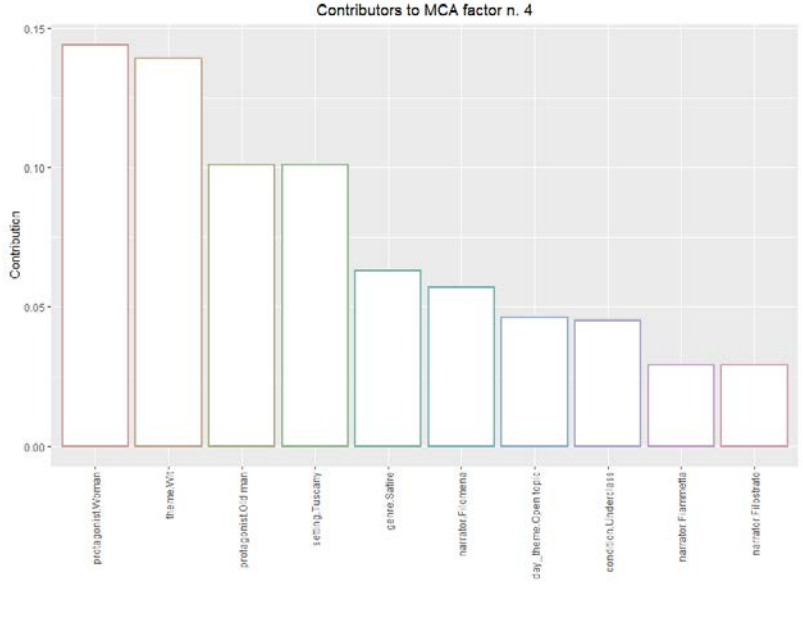
```
#> [1] "Size of compact file:3438"
#> [1] "Size of non-compact file:76914"
```

(1) See H. Abdi, J. Williams “Correspondence Analysis” in N. Salkind (Ed.) *Encyclopaedia of Research Design*

Improving MCA visualization

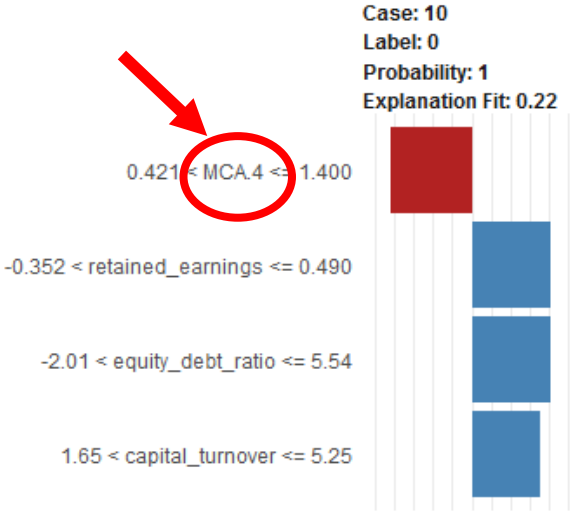
- Improved charting facilities aimed at better linking original variables and factors (focus on the contributors to the factors)





Improving interpretation of results

- LIME adopted for the explanation of results obtained via ML algorithms
- Improved readability of LIME results by means of:
  - Conversion of LIME results to textual explanations, to be easily integrated within the normal communication channels
  - Conversion of LIME results involving MCA factors to assessments related to the original categorical variables




```
# view explanation
a$explanation$lime_light_exp
#>  [1] "retained_earnings: 45%; industry: 22%; working_capital: 20%"
#>  [2] "equity_debt_ratio: 51%; industry: 30%; working_capital: 16%"
#>  [3] "industry: 41%; capital_turnover: 33%; working_capital: 16%"
#>  [4] "equity_debt_ratio: 26%; retained_earnings: 24%; industry: 19%"
```

Improving use of ensemble of algorithms

- An ensemble of algorithms is employed
- A comparison among heterogeneous approach is required to make sense of the results (scores)
- The ranking of the scores obtained from heterogeneous algorithms is obtained as the weighted average of the provided scores.
- Comparability among algorithms is pursued through robust standardisation
- Feasibility is ensured by means of a cap on the overall workload
- Relevance is tested against labelled anomalies dataset

Thank you for your attention

For any question or comment please feel free to contact us:

matteo.accornero@ecb.europa.eu

gianluca.boscariol@ecb.europa.eu