EUROPEAN CENTRAL BANK

EUROSYSTEM

Nicola Benatti
European Central Bank

A machine learning approach to outlier detection and imputation of missing data

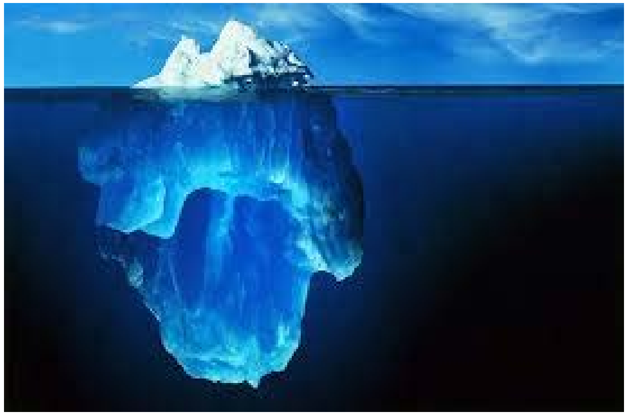
9th IFC Conference 30-31 Aug 2018, Basel

DISCLAIMER: This paper should not be reported as representing the views of the European Central Bank. The views expressed in this paper are those of the authors and do not necessarily reflect those of the European Central Bank.

Overview

What is an outlier?

- An outlier is an observation which is significantly distant from the other considered observations.



- Often outliers are identified by assuming the true distribution of each variable separately to be a known one.

- Alternatively, distributional methods are used but they do not suggest the true values of the Observation.

- It is very important that outliers are not automatically considered as errors since extreme cases can still be justified.

- The aim of this analysis is to rank observations that need to be assessed by their likelihood of being errors.

The iBACH dataset

- Balance sheet and profit and loss data of firms collected by the European Committee of Central Balance Sheet Data Offices (ECCBSO) within its WG on Bank for the Accounts of Companies Harmonized (BACH).
- Aggregate database available since several years but firm level data (iBACH) available to participating countries since February 2018.
- 66 numeric variables taken into consideration in the analysis I carry out

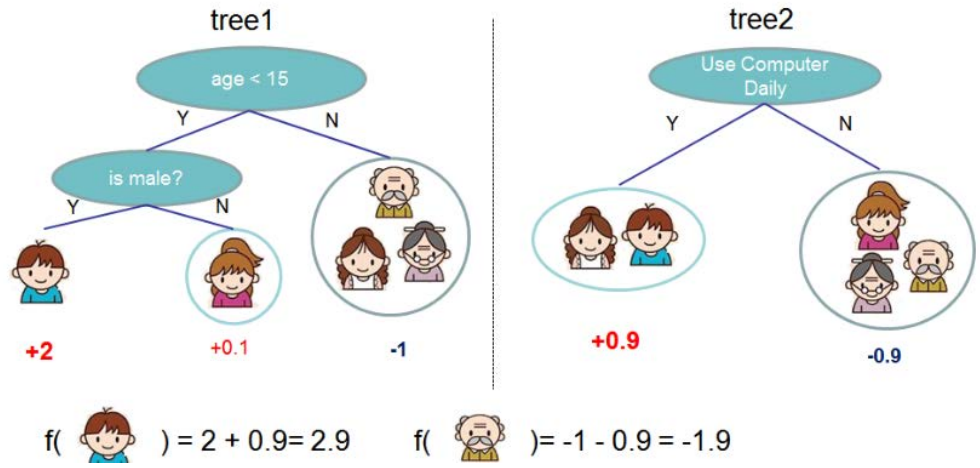
Number of entities

	dyear													
dcountry	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
BE		204,825	218,707	233,180	250,392	264,474	284,327	297,899	326,480	344,480	362,762	377,386	382,669	349,034
ES							450,538	447,540	459,076	450,146	443,527	583,081	560,570	324,701
FR	184,812	192,206	198,107	208,534	225,408	233,267	233,865	244,843	260,670	260,565	250,048	253,758	257,950	261,051
IT	492,472	517,464	540,517	560,140	582,993	600,656	613,021	624,235	634,278	629,865	627,317	621,722	618,177	464,353
PT	16,920	17,547	15,176	342,588	357,480	367,237	366,806	365,821	373,230	373,500	378,731	382,779	390,730	392,030
SK												99,389	99,584	97,869

Sum of n_entities broken down by dyear vs. dcountry. Color shows sum of n_entities. The marks are labeled by sum of n_entities.

Estimation: XGBoost, Gridsearch

- The estimation technique used is extreme gradient boosting (Chen 2016, in the python package xgboost)



- The hyperparameters are set using a Gridsearch algorithm (M. Claesen, B. De Moor 2015, in the python package Gridsearch) which iterates over a tuple of values and choses the optimal set for the following hyperparameters of xgboost: max depth, eta, subsample, number of estimators

Detection: Distance measures and importance averaging

- **Outlier flagging methods using estimation residuals:**
  - K-nearest neighbour on absolute and relative distance from true value
  - Distribution based on both absolute and relative distance from true value

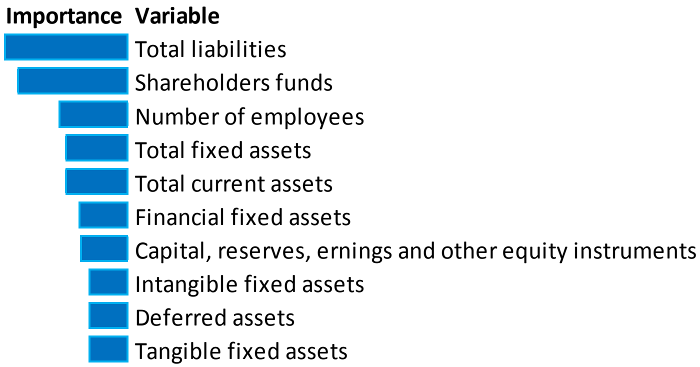
- **Importance averaging:**
  - While causality is confirmed, it might not be clear where the error comes from

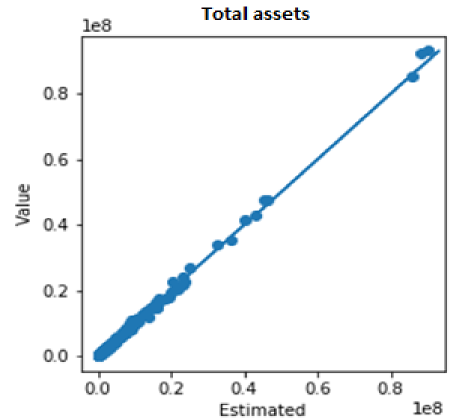
A=B-(C*D) is false
Which variable among A, B, C and D is wrong?

  - For each firm/year I sum the contribution of each variable to the model of detected outliers and create a ranking of “most-likely-to be wrong”.

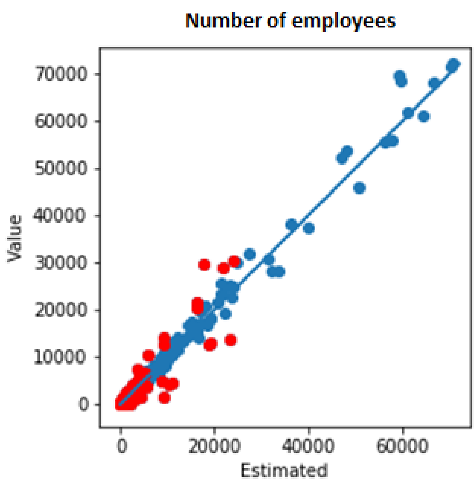
Results

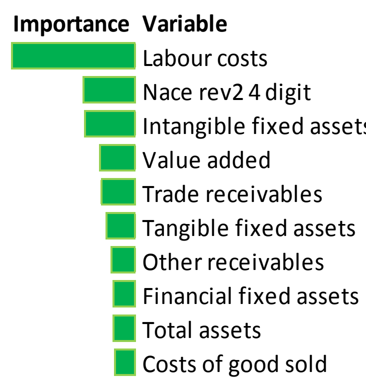
The algorithm allows to accurately estimate all variables analysed.





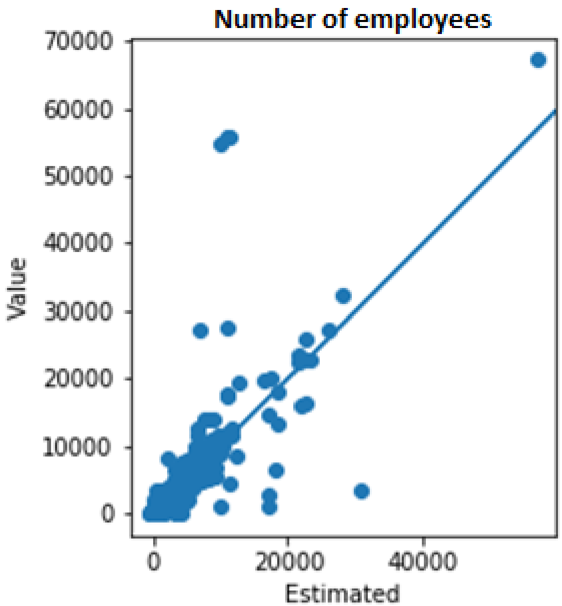
The outliers detected re sent to the NCBs to be investigated, ranked by likelihood of being errors.

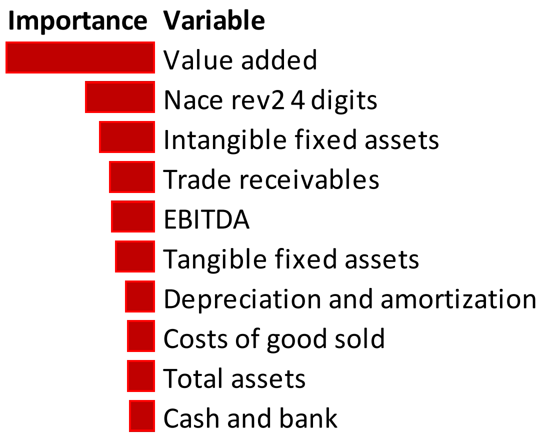




Imputation

The same methodology can be used to estimate the missing values in the dataset. As an exercise, when estimating employment, forcing out the labour costs variable, the estimation still over-performs the methodology used previously internally.





Conclusions

- This paper presents an application of a combination of supervised and unsupervised machine learning with a final feature-additive ranking technique in order to spot mistakes in outlying datapoints.
- The methodology described seems to be useful also for additional steps of data quality improvement such as data imputation.
- This technique also provides guidance for the construction of new data quality checks that could prevent the submissions of mistakes.

Further improvements:

- The increase in the sample size.
- The inclusion of lagged variables would allow for using long-term-short-term memory frameworks.
- The comparison of the results with neural networks and multi-target regressions.
- Inclusion of the confirmation on whether a spotted potential mistake is actually an error or not to transform the distance measure into a classification problem.