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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Data</td>
</tr>
<tr>
<td>3</td>
<td>Methodology</td>
</tr>
<tr>
<td>4</td>
<td>Results</td>
</tr>
<tr>
<td>5</td>
<td>Imputation</td>
</tr>
<tr>
<td>6</td>
<td>Conclusions</td>
</tr>
</tbody>
</table>
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

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

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
    
    $A = B - (C \times D)$ 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”.
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.
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.