Comparison of anomaly detection results - Notebook 2

This document presents the results of the Notebook entitled ‘Anomaly detection 2’, where we perform anomaly detection on the dataset data_clean.xlsx using : iForest, OneClassSVM and DBSCAN. We are going to detect anomalies in the entire dataset using all the variables after encoding the categorical ones.

1 - Isolation Forest

- In order the specify the “contamination” parameter c we are going to exploit the results obtained in the other study, we are going to set c as the weighted average of all the fractions of outliers in each mediaType, which gives c = 0.0035
- The percentage of total anomalies found in the dataset : ***0,35%***
- The detailed results :

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	0,14%	14	Austria (57,92%) Switzerland Zug (19,30%) India kalwa (17,76%) United State DA (2,70%) Portugal (1,54%) United State HU (0,39%) China (0,39%)
Heating	0%	-	-
ColdWater	0,33%	7	Switzerland Zug (51%) Switzerland SH (21,57%) Netherland(20,59%) Denmark(4,90%) Germany(1,96%)
WS_Brunnenwasser	0%	-	-
Electricity1	0,00%	-	-
Electricity2	0,00%	-	-
Power	0,27%	2	Germany (100%)
WS_Blindstrom	0,00%	-	-
ElectricityGenEmergency	0%	-	-

ElectricityGenPV	0%	-	-
NGas	6,32%	4	Switzerland SH (98,7%) Portugal (0,98%) United State HU (0,16%) United State PO (0,16%)
GeneralElectricity	0,14%	1	Switzerland SH (100%)
DistrictHeating	0%	-	-
Cooling	0%	-	-
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	0,00%	-	-
ColdGenerated	0%	-	-
ElectricitySupplyPV	0%	-	-
Diesel	0%	-	-

2 - OneClassSVM

- We are going to specify the parameter nu using the same method as we did for the “contamination” parameter , which gives nu = 0.0035
- The percentage of total anomalies found in the dataset : ***0,35%***
- Detailed results :

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	0,31%	62	India kalwa (71,78%) China (13,41%) Netherland (6,62%) Austria (2,26%) India GU (1,92%) Portugal (1,05%) Switzerland Zug (0,87%) United State FH (0,70%) United State JS (0,70%) United State SQ (0,35%) Switzerland SH (0,17%) United State DA (0,17%)

Heating	0%	-	-
ColdWater	0,01%	3	Germany (66,67%) Singapour (33,33%)
WS_Brunnenwasser	0,03%	1	Austria (100%)
Electricity1	33,78%	1	Austria (100%)
Electricity2	0,1%	1	Austria (100%)
Power	0,00%	-	-
WS_Blindstrom	0,00%	-	-
ElectricityGenEmergency	0,09%	1	India kalwa (100%)
ElectricityGenPV	0,33%	3	India kalwa (100%)
NGas	0,04%	1	United State PO (100%)
GeneralElectricity	0%	-	-
DistrictHeating	0,07%	1	Denmark (100%)
Cooling	0%	-	-
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	4,62%	1	China (100%)
ColdGenerated	0,41%	1	Switzerland SH (100%)
ElectricitySupplyPV	0%	-	-
Diesel	0%	-	-

3 - DBSCAN

- We are going to use the same parameters as the other study namely : eps = 1, min_samples = 2, to conserve the coherence of the results

- Using this configuration of DBSCAN with this approach of mixing all the mediaTypes at once doesn't give reasonable results, it considers ***97%*** of the data as anomalies.

- A potential explanation is that DBSCAN is a density based method and we have a very sparse dataset …

Observations :

- Using this approach iForest and OneClassSVM gave totally different results, and also in comparison with the first study