Comparison of anomaly detection results - Notebook 1

This document presents the results of the notebook entitled ‘Anomaly detection 1’ where we perform anomaly detection on the dataset data_clean.xlsx using : Z-score, iForest, OneClassSVM and DBSCAN. We are going to detect anomalies for each mediaType separately and based on the "data" column alone

0 - Missing data

- Percentage : 2,37 %
- Nb of Locations concerned : 20
- Number of meterIds concerned : 95
- Nb of mediaTypes concerned : 8

1 - Statistical approach

This method is based on the calculation of the Z-score; the z-score measures how far a data point is away from the mean as a signed multiple of the standard deviation. Large absolute values of the Z-score suggest an anomaly

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	0,07%	34	India kalwa (95,5%) Portugal (3%) India GU (1,5%)
Heating	2,66%	3	Austria (100%)
ColdWater	0,01%	2	Netherland (100%)
WS_Brunnenwasser	3,82%	5	Austria (100%)
Electricity1	0,00%	-	-
Electricity2	0,00%	-	-
Power	0,04%	1	Austria (100%)
WS_Blindstorm	0,00%	-	-
ElectricityGenEmergency	0,31%	2	India kalwa(100%)
ElectricityGenPV	0,23%	1	India kalwa(100%)
NGas	0,16%	5	United State SQ (25%) United State HE (18,75%) United State DA (18,75%) United State BE United (18,75%) State FH (18,75%)

GeneralElectricity	0,03%	1	Switzerland SH (100%)
DistrictHeating	0,11%	1	Denmark (100%)
Cooling	1,09%	3	Germany (100%)
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	0,00%	-	-
ColdGenerated	3,39%	1	Switzerland SH (100%)
ElectricitySupplyPV	3,42%	2	India kalwa(100%)
Diesel	0,00%	-	-

* Values that correspond to ElectricityGenCHP and HeatGenCHP mediaTypes are all either missing or equal to 0

2 - Machine learning approach

2.1 - Isolation forest

→ This algorithm requires a parameter called “contamination” that refers to the fraction of anomalies estimated in the data, we are going to estimate this parameter based on the results of the previous method
→ The other hyperparameters of the used model : n_estimators=50, max_samples='auto', max_features=1.0,random_state=42

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	0,07%	34	India kalwa (96%) Portugal (1,62%) India GU (1,62%) Switzerland Zug (0,81%)
Heating	2,65%	3	Austria (100%)
ColdWater	0,01%	2	Netherland (100%)
WS_Brunnenwasser	3,82%	4	Austria (100%)
Electricity1	0,00%	-	-
Electricity2	0,00%	-	-

Power	0,04%	1	Austria (100%)
WS_Blindstorm	0,00%	-	-
ElectricityGenEmergency	0,33%	2	India kalwa(100%)
ElectricityGenPV	0,23%	1	India kalwa(100%)
NGas	0,16%	5	United State SQ (25%) United State HE (18,75%) United State DA (18,75%) United State BE United (18,75%) State FH (18,75%)
GeneralElectricity	0,03%	1	Switzerland SH (100%)
DistrictHeating	0,11%	2	Denmark (50%) Netherland (50%)
Cooling	1,09%	3	Germany (100%)
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	0,00%	-	-
ColdGenerated	3,29%	1	Switzerland SH (100%)
ElectricitySupplyPV	3,42%	2	India kalwa(100%)
Diesel	0,00%	-	-

2.2 - OneClassSVM

→ This algorithm requires a parameter “nu” that specifies an approximation ratio of the outliers in the dataset, we are going to estimate this parameter in the same way we did for the “contamination”
→ The other hyperparameters of the used model are the default ones

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	2,02%	42	India kalwa (99,81%) Portugal (0,13%) India GU (0,053%)
Heating	2,65%	6	Austria (100%)
ColdWater	31,53%	30	Austria (67,67%) Germany (18,45%)

			Switzerland Zug (6,88%) Switzerland SH (2,31%) China (2,07%) Portugal (1,84%) Denmark (0,74%) Netherland (0,01%)
WS_Brunnenwasser	20,89%	10	Austria (100%)
Electricity1	65,5%	1	Austria (100%)
Electricity2	9,34%	1	Austria (100%)
Power	0,03%	1	Austria (100%)
WS_Blindstrom	13,24%	1	Austria (100%)
ElectricityGenEmergency	0,22%	3	India kalwa(90%) India BL (10%)
ElectricityGenPV	70,29%	6	Switzerland SH (60,32%) India kalwa(39,68%)
NGas	0,47%	6	United State DA (73,91%) United State SQ (8,70%) United State HE (6,52%) United State BE (4,35%) United State FH (4,35%) United States HU (2,17%)
GeneralElectricity	6,87%	1	Switzerland SH (100%)
DistrictHeating	0,1%	1	Denmark (100%)
Cooling	3,15%	8	Germany (100%)
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	30,36%	2	China (100%)
ColdGenerated	4,62%	1	Switzerland SH (100%)
ElectricitySupplyPV	90,99%	2	India kalwa(100%)
Diesel	50,83%	2	India kalwa(75,40%) India BL (24,60%)

2.3 - DBSCAN

→ This algorithm requires 2 parameters : min_samples and eps, that we set to min_samples = 2 and eps = 1 , using the previous results

mediaType	percentage of anomalies	number of meters concerned	locations concerned
Electricity	1,49%	84	India kalwa (59,45%) Austria (20,82%) Switzerland Zug(12,93%) Netherland (3,34%) Portugal (1,34%) China (0,58%) Germany (0,40%) India GU (0,36%) United State SQ (0,29%) United State DA (0,14%) United State LA (0,11%) United State PO (0,11%) United State HU (0,07%) Denmark (0,04%)
Heating	6,89%	8	Austria (100%)
ColdWater	0,04%	6	Netherland (36,36%) Portugal (36,36%) Austria (9,09%) United State FH (9,09%) China (9,09%)
WS_Brunnenwasser	0,34%	3	Austria (100%)
Electricity1	0,1%	1	Austria (100%)
Electricity2	0,1%	1	Austria (100%)
Power	0,24%	2	Austria (63,63%) Germany (36,36%)
WS_Blindstorm	0,1%	1	Austria (100%)
ElectricityGenEmergency	0,35%	5	India kalwa(68,75%) India BL (31,25%)
ElectricityGenPV	13,98%	4	India kalwa(99,81%) Switzerland SH (0,18%)
NGas	0,42%	9	Switzerland SH (41,46%) United State HU (14,63%) United State SQ (9,75%) United State HE (9,75%) United State DA (7,31%) United State BE (7,31%) United State FH (7,31%) United State JS (2,44%)
GeneralElectricity	3,68%	3	Switzerland SH (100%)
DistrictHeating	6,65%	10	Germany (64,77%) Netherland (27,71%)

			Denmark (7,51%)
Cooling	2,1%	8	Germany (100%)
ElectricityGenCHP	0,00%*	-	-
HeatGenCHP	0,00%*	-	-
IntervalGas	0,00%	-	-
ColdGenerated	3,49%	1	Switzerland SH (100%)
ElectricitySupplyPV	7,76%	2	India kalwa(100%)
Diesel	0,00%	-	-

Observations :

- The Z-score method and iForest give very similar results
- OneClassSVM are less similar (about 50% similarity) they tend to detect more anomalies in addition to the ones detected by the first 2 methods