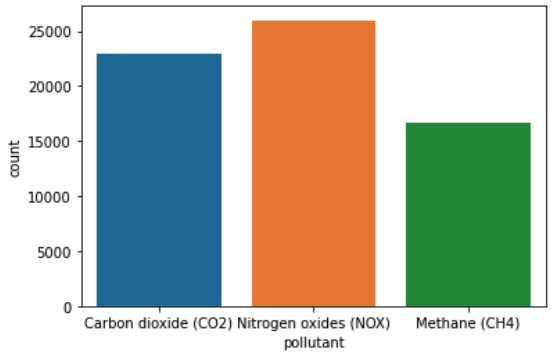
schneider-electric-challenge

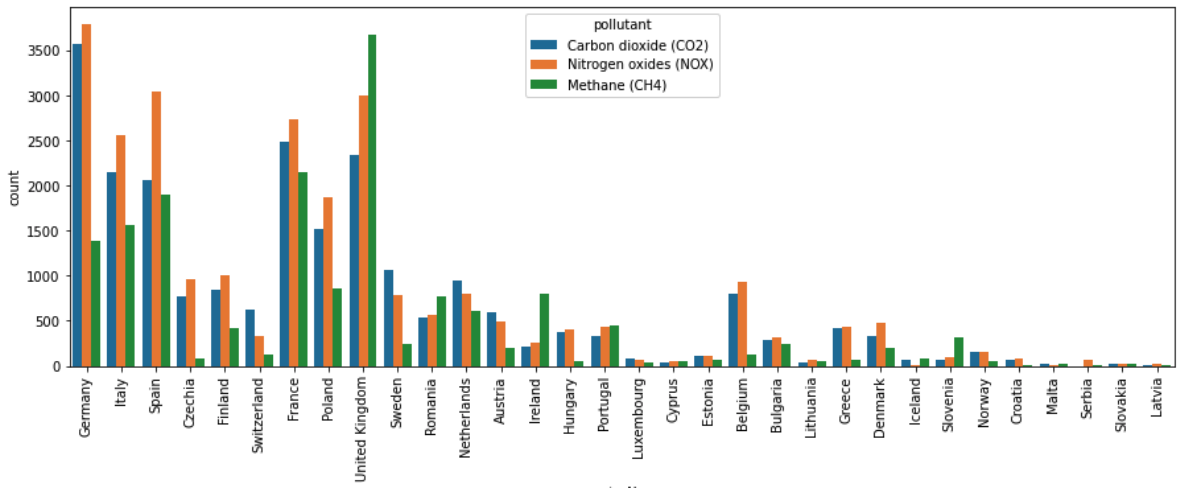
Group 36
Authors:
Arturo Lidueña - arturo.liduena@gmail.com
Alfredo Mariño - alfredoalejandro05@gmail.com
Date: 21/05/22

In order to determinate the type of pollutant emitted by the facility and predict it in new registers, the team has decided to explore the dataset, clean it and create a model that predict the type of pollutant (Nitrogen oxides (NOX), Carbon dioxide (CO2), Methane (CH4)).

Pollutant type Distribution:



Pollutant type by Countries



Stadistic

	count	mean	std	min	25%	50%	75%	max
reportingYear	90108.0	-35.059928	3.857469	-41.000000	-38.000000	-35.000000	-32.000000	-28.000000
MONTH	90108.0	6.493397	3.447282	1.000000	3.000000	7.000000	9.000000	12.000000
DAY	90108.0	14.543281	8.098556	1.000000	8.000000	15.000000	22.000000	28.000000
max_wind_speed	90108.0	15.512083	3.069792	8.011958	13.311790	15.505221	17.716792	22.991382
avg_wind_speed	90108.0	18.013150	2.311901	14.000100	16.013215	18.017877	20.017578	21.999973
min_wind_speed	90108.0	22.518903	3.060429	15.032589	20.331653	22.535910	24.708091	29.951401
max_temp	90108.0	9.464618	5.222313	-3.159766	5.880222	9.700750	13.296754	20.938266
avg_temp	90108.0	10.457452	5.093656	-0.199916	7.189392	10.711595	14.204277	19.999403
min_temp	90108.0	13.448275	5.228531	0.888248	9.880362	13.698286	17.293781	24.944638
DAY WITH FOGS	90108.0	2.238403	3.782805	0.000000	0.000000	1.000000	2.000000	19.000000
EPRTRSectorCode	52981.0	3.177252	2.048738	1.000000	1.000000	3.000000	5.000000	9.000000
test_index	24480.0	12239.500000	7066.911631	0.000000	6119.750000	12239.500000	18359.250000	24479.000000

	count	unique	top	freq
countryName	90108	32	United Kingdom	12499
eprtrSectorName	90108	9	Energy sector	33725
EPRTRAnnexIMainActivityLabel	90108	71	Thermal power stations and other combustion in...	29562
FacilityInspireID	90108	7617	https://data.ied_registry.omgeving.vlaanderen....	55
facilityName	90108	8516	Enel Produzione S.p.A.	301
City	90108	5426	--	2713
targetRelease	90108	1	AIR	90108
pollutant	65628	3	Nitrogen oxides (NOX)	25982
CONTINENT	90108	1	EUROPE	90108
REPORTER NAME	90108	60488	Michael Smith	34
CITY ID	90108	5426	cfab1ba8c67c7c838db98d666f02a132	2713
EPRTRAnnexIMainActivityCode	52981	71	1(c)	17413

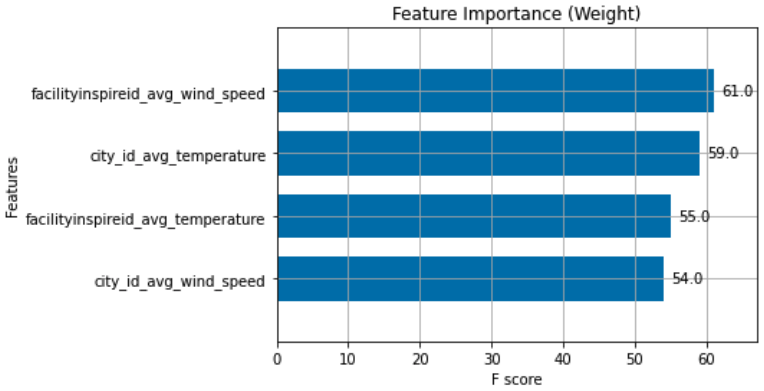
Decisions:

- **EPRTRSectorCode**: REMOVE - same as eprtrSectorName
- **EPRTRAnnexIMainActivityCode**: REMOVE - same as EPRTRAnnexIMainActivityLabel
- **targetRelease**: REMOVE - one unique value, it doesn't add valued information
- **City**: REMOVE - same as CITY ID
- **CONTINENT**: REMOVE - one unique value, it doesn't add valued information
- **REPORTER NAME** - REMOVE
- **countryName**: ONE HOT ENCONE
- **eprtrSectorName**: ONE HOT ENCONE
- **EPRTRAnnexIMainActivityLabel**: ONE HOT ENCONE

Feature generation:
Usign an avg function aggregation we add to a specific row the avg_temperture and avg_wind_speed by:

- **FacilityInspireID**: FacilityInspireID_avg_temperature, FacilityInspireID_avg_wind_speed
- **facilityName**: facilityName_avg_temperature, facilityName_avg_wind_speed
- **CITY ID**: CITY ID_avg_temperature, CITY ID_avg_wind_speed

Modelling:
Using XGBClassifier we created a model to predict the type of pollutant, those are the parameters using in the model: max_depth=3, min_child_weight=100, random_state=42, use_label_encoder=False, objective='multi:softprob', learning_rate=0.3, n_estimators=200.
-Parameters to fit the model: eval_metric="merror", eval_set=eval_set, verbose=True, early_stopping_rounds=10.



Result:
F1_score 0.6445