



City of Seattle

*Energy consumption*  
and  
*Greenhouse gases emission*  
**predictions**  
for the non-residential properties of Seattle.

Julien Le Boucher

03 - 31 - 2023

# Plan

- Context / motivation of this mission.
- Work on the dataset and modelling approach
- The best model and its performance.

# Seattle's Climate plan Objectives



## The city objectives :

- zero net core greenhouse gas emission by 2050.
- Sector-specific emissions reduction goals of 39% from building energy by 2030 comparing to 2008.

## My missions :

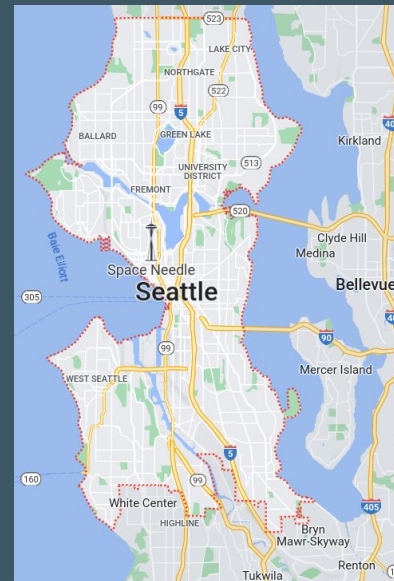
- Predict :
  - Energy Consumption
  - GHG emissionsof Non-residential Buildings from its structural attributes.
- Evaluate the impact of the Energy Star Score when used as an input in the models.

# Dataset Overview

---

# The dataset

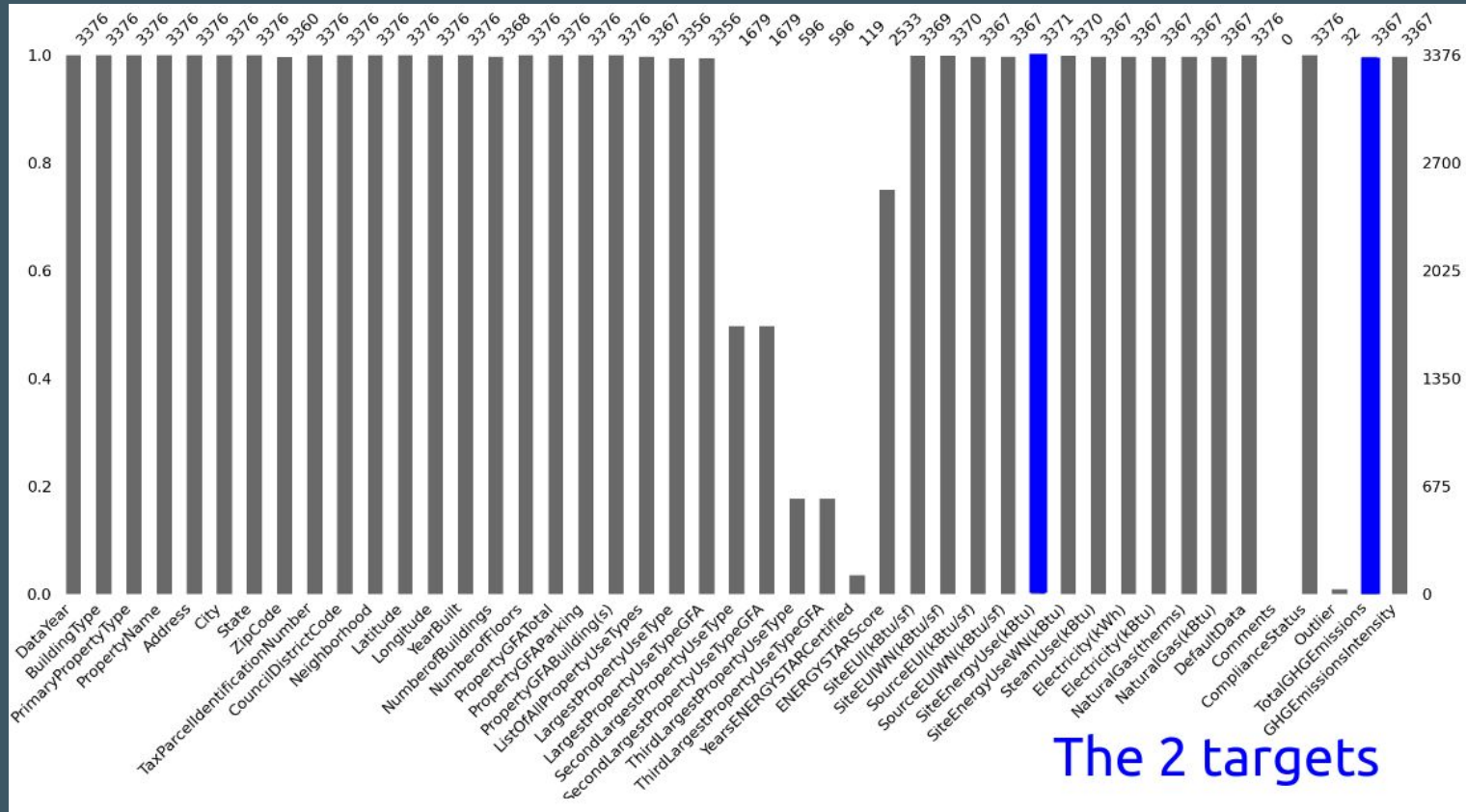
- information gathered in 2016. Mandatory for all Seattle's properties with a GFA > 20 000 feet square.
- 3376 properties of Seattle (rows).
- 45 features (columns).



	DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode
OSEBuildingID								
20890	2016	NonResidential	K-12 School	ASB School	6220 32nd Ave. NE	Seattle	WA	98115.0
49778	2016	NonResidential	Worship Facility	Center for Spiritual Living	5801 SAND POINT WAY NE	Seattle	WA	98105.0
27996	2016	Multifamily LR (1-4)	Low-Rise Multifamily	THE KENNEY	7125 FAUNTLEROY WAY SW	Seattle	WA	98136.0

A chunk of the dataset

# Missing values, Features groups and Targets.



The 2 targets

1 - Type and primary use

2 - Name and localization.

3 - Buildings' attributes.

4 - ENERGY STAR Score.

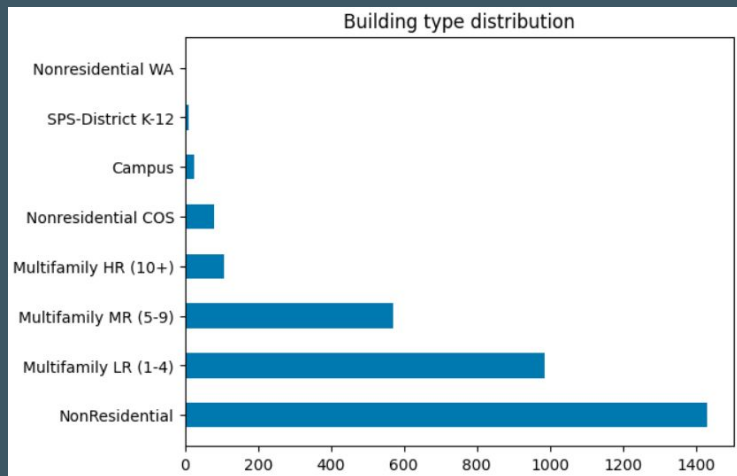
5 - Energy Consumption.

6 - Data Status.

7 - GHG Emissions.

# Filtering non-residential properties

- Compliant Status → 165 properties discarded (missing or default data)
- Filtered on a triplet of features to ensure only Non-Residential properties remained in the data.



	PrimaryPropertyType	LargestPropertyUseType	BuildingType
OSEBuildingID			
264	Mixed Use Property	Multifamily Housing	NonResidential
19445	Low-Rise Multifamily	Multifamily Housing	NonResidential
21122	Mixed Use Property	Multifamily Housing	NonResidential
21481	Low-Rise Multifamily	Multifamily Housing	Campus
23562	Mixed Use Property	Multifamily Housing	NonResidential
25222	Mixed Use Property	Multifamily Housing	NonResidential
25522	Mixed Use Property	Multifamily Housing	NonResidential
26834	Mixed Use Property	Multifamily Housing	NonResidential
27838	Mixed Use Property	Multifamily Housing	NonResidential
27969	Mixed Use Property	Multifamily Housing	NonResidential
29170	Mixed Use Property	Multifamily Housing	NonResidential

Example of inconsistencies found.

3376 properties

1487  
properties

# EDA on non-residential properties

---

Features selection motives



## Structural attributes

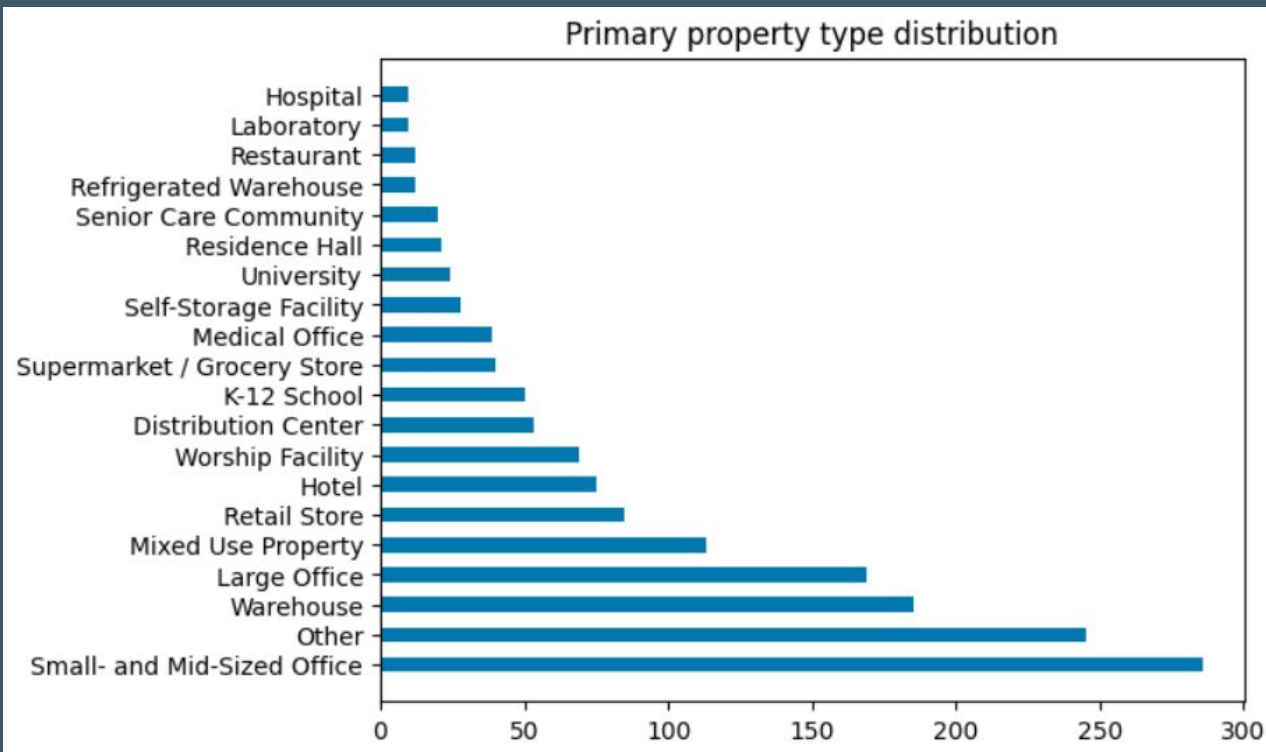


- 4 structural attributes out of 5 are correlated with the targets.  
Thus, should go in the inputs.
- Total GFA and buildings' GFA are highly correlated .  
→ Choose only 1 as input to avoid multicollinearity.

Retained Structural Features as inputs :

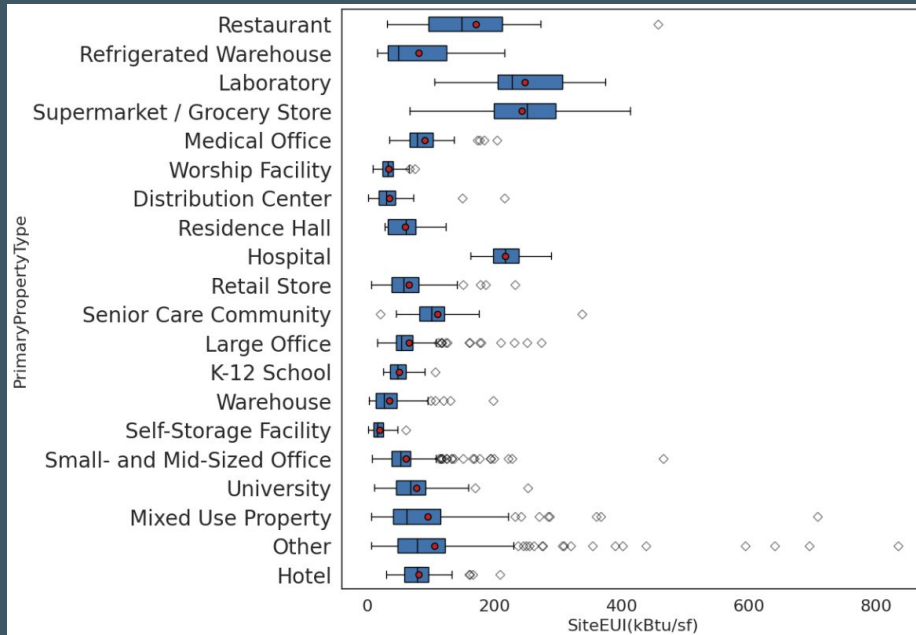
- 'PropertyGFABuilding(s)'
- 'NumberofFloors'
- 'NumberofBuildings'

# Grouping thanks to 'PrimaryPropertyType'

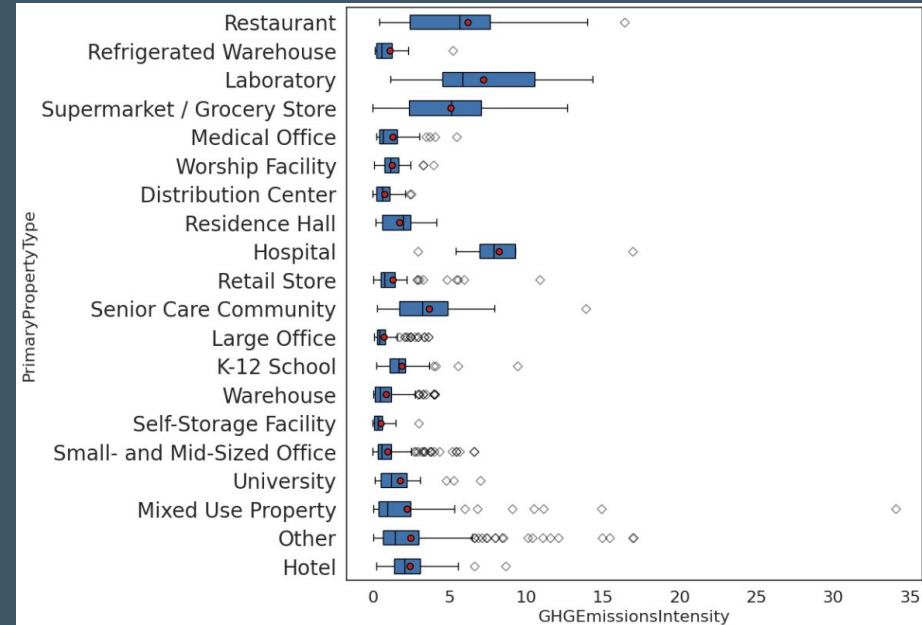


- Enables to split properties in several groups without being too specific (unlike 'LargestPropertyUseType').
- One-Hot Encoded dropping one column to avoid multicollinearity.
- Correlated with the targets.

# 'PrimaryPropertyType' correlations with the targets' intensities



eta squared : 0.318



eta squared : 0.264

The category has an influence regardless of the surface of the property.  
Must be feed in the model

# Features engineering

---

## 4 derived features.

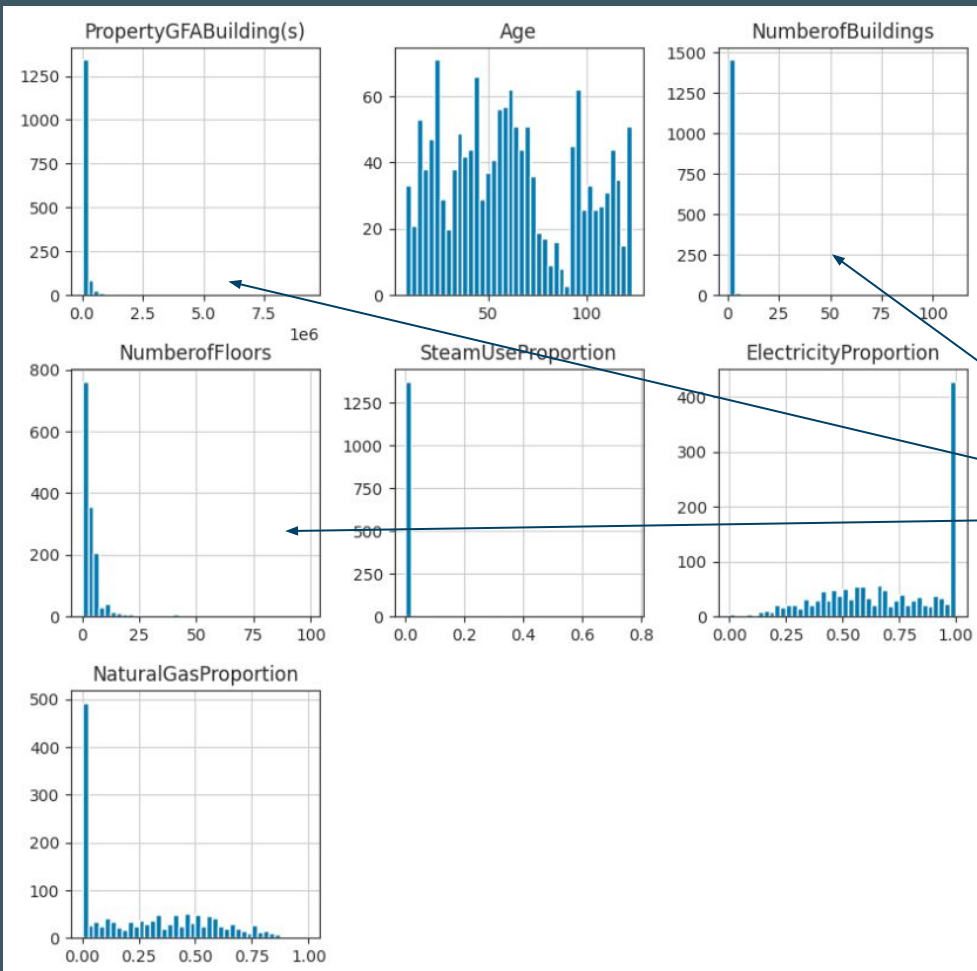
- From YearBuilt (date of construction or of a total renovation)  
→  $2023 - \text{YearBuilt} = \text{Age}$
- From SiteEnergyUse and the 3 features (Electricity, NaturalGas, SteamUse).  
→ Proportions of energy used per source type which can be seen as structural and should be assessable when building a new structure or renovating.
  - SteamUseProportion
  - ElectricityProportion
  - NaturalGasProportion.

Currently not 100% reliable : Outliers problem ?

# Models' inputs and targets

---

# 26 inputs



## 7 numericals inputs

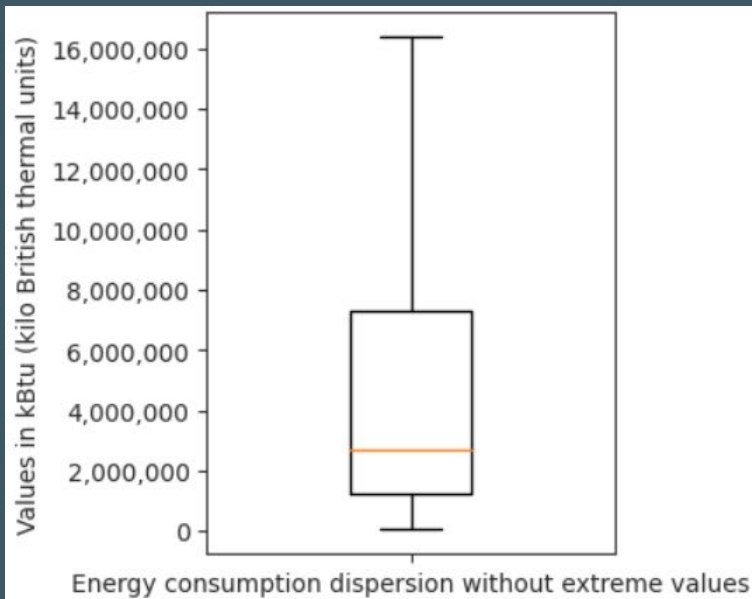
### Pre-processing possibilities

- The proportions are already scaled between 0 and 1.
- All the other features should be scaled? (MinMax, StandardScaler?)
- 3 skewed distributions. Transformation before going in the model? (log, quantile ?)

## + 19 binary inputs

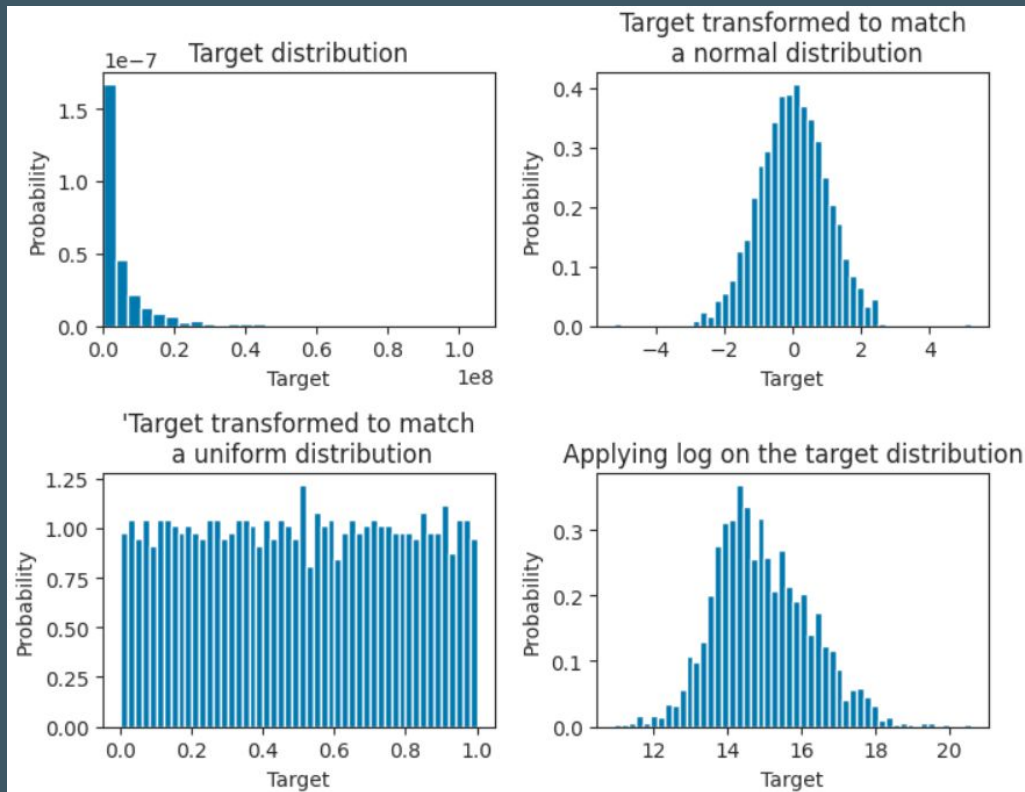
Primary property type One-Hot encoded feature.

# 1st target : Total energy consumption (kBtu)



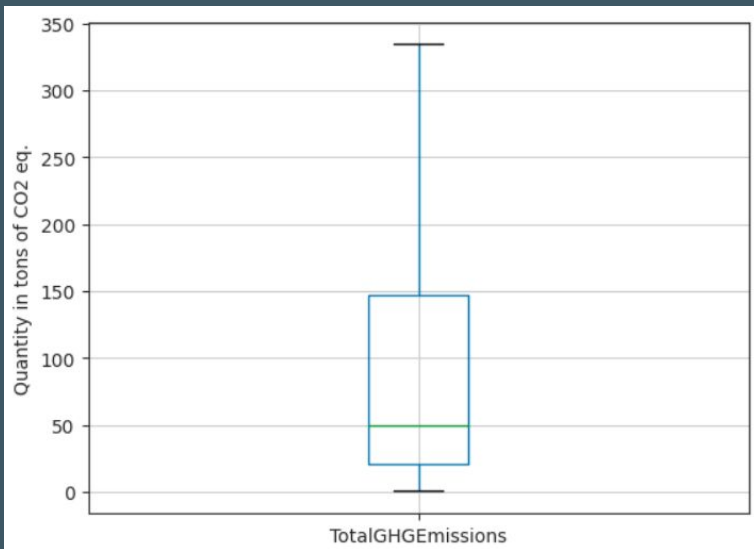
Min : 5 713 kBtu

Max : 874 000 000 kBtu

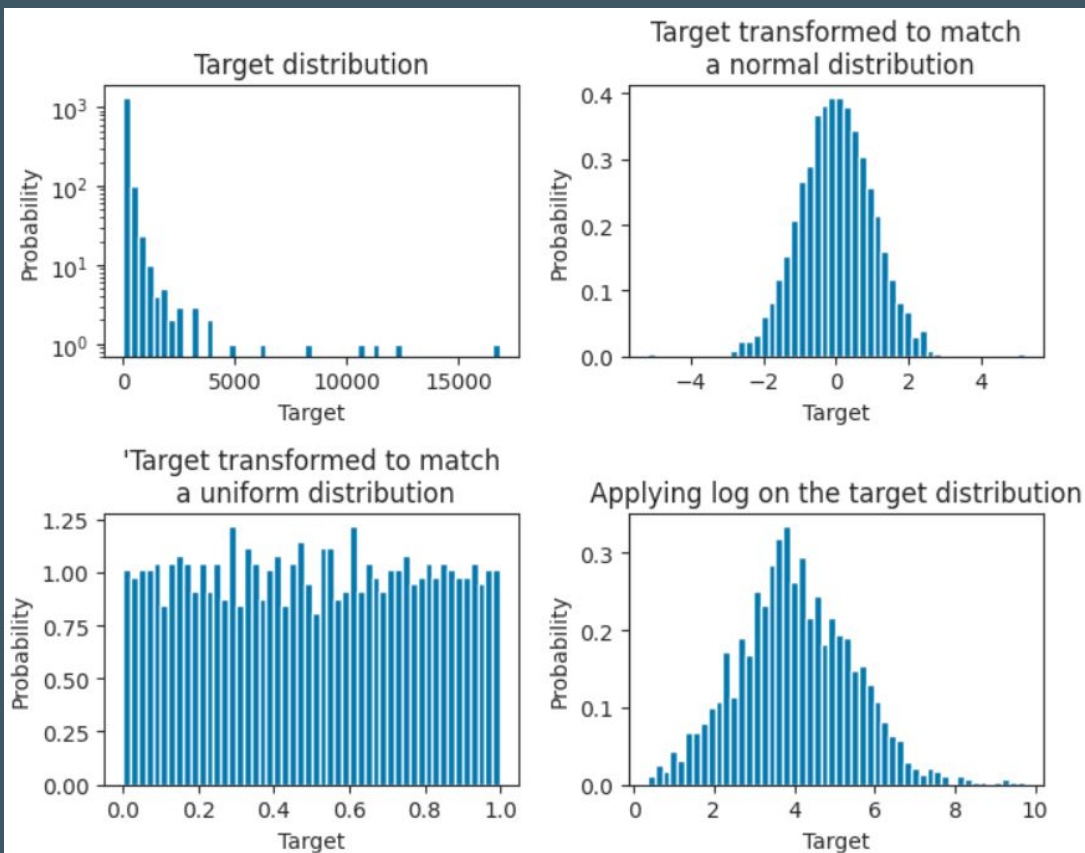




## 2nd target : Greenhouse gases emissions (in tons of CO2 eq)



Min : 0.4 tons CO2 eq.  
Max : 16 870 tons CO2 eq.

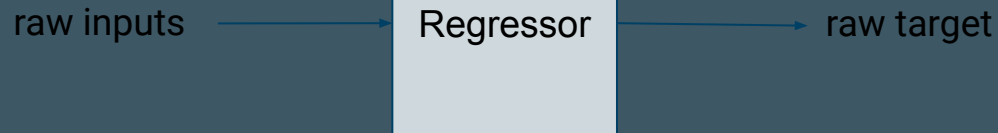


# Modelling approach

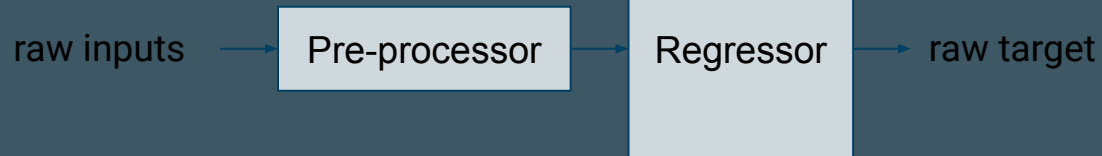
---

Searching the best model

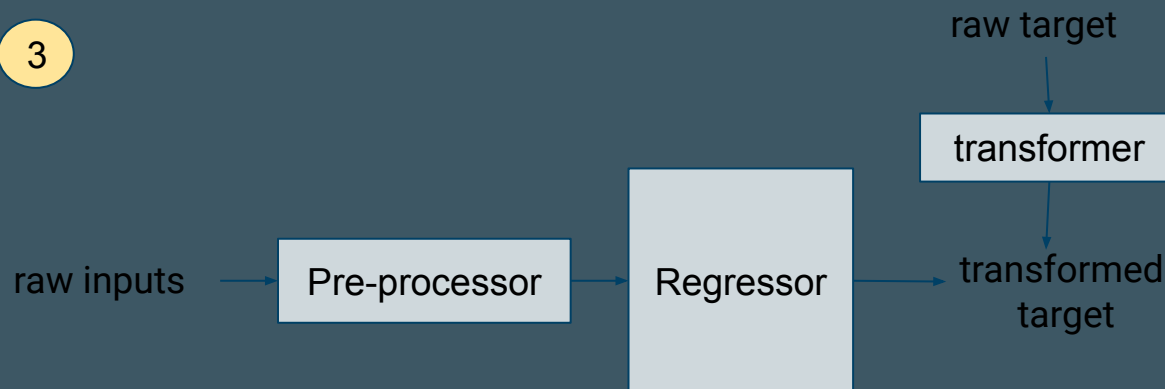
1



2



3



- Playing on the modelling steps complexity (1, 2, or 3) and permuting the following elements :

- REGRESSOR :

- OLS regressor
- Lasso
- Ridge
- ElasticNet
- Knn
- Random Forest
- XGBoost (gbtree)

- PRE-PROCESSOR :

- passthrough
- MinMax
- Standard
- log transformer
- log into the 2 firsts.
- quantile transformer

- TRANSFORMER :

- log transformer
- quantile transformer fit on the train set.

# Hyperparameters tuning + evaluation of models' performances.



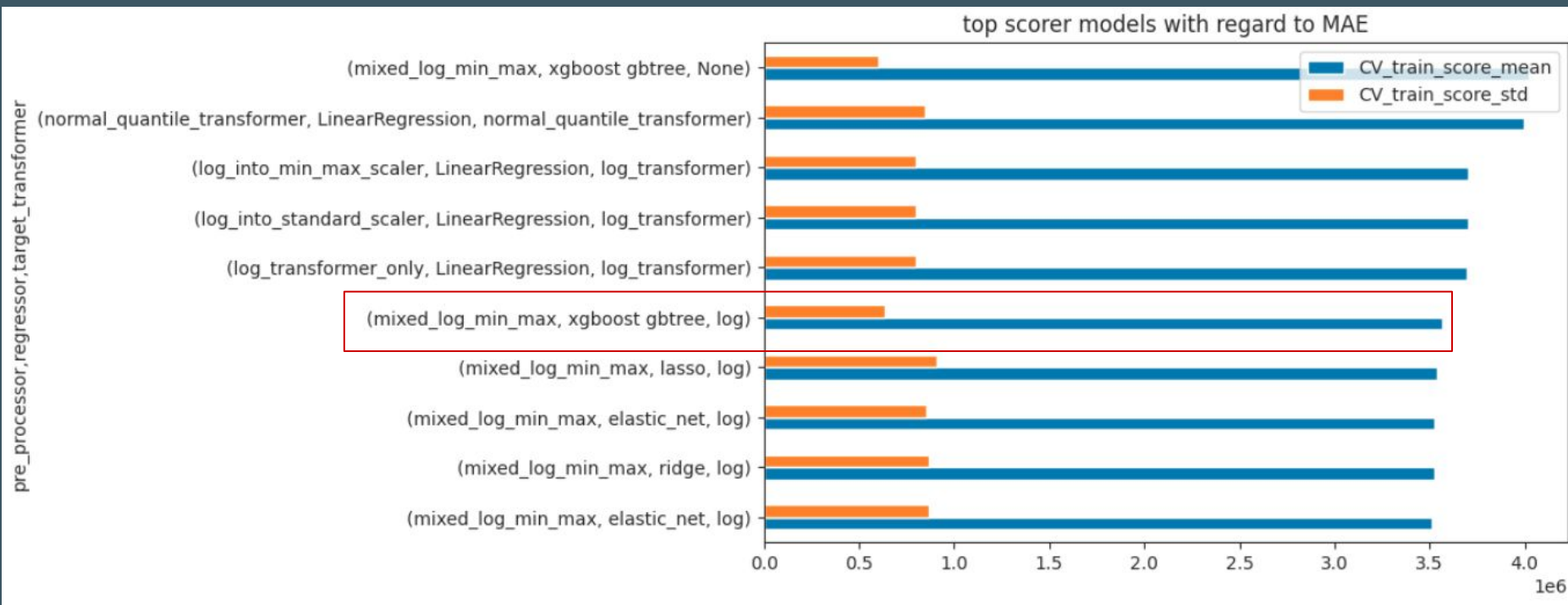
*sklearn's illustration of a 5-fold cross validation*

- Test data : 25 % of the data.
- **Hyperparameters tuning :**
  - Optimized sequentially thanks to grid searches.
  - Choices made with regard to 3 different metrics : R2, MAE, RMSE.  
(can lead to model slightly different)
  - Optimized the mean score on 5 folds.  
Kept in mind that minimizing the variance was also important.
- **Evaluation of the model :**
  - Saved the mean score on folds and the variance.
  - Compared fitting time on the train set.
  - Plotted predictions against the real values of the target on the test data.

# Top scorers models

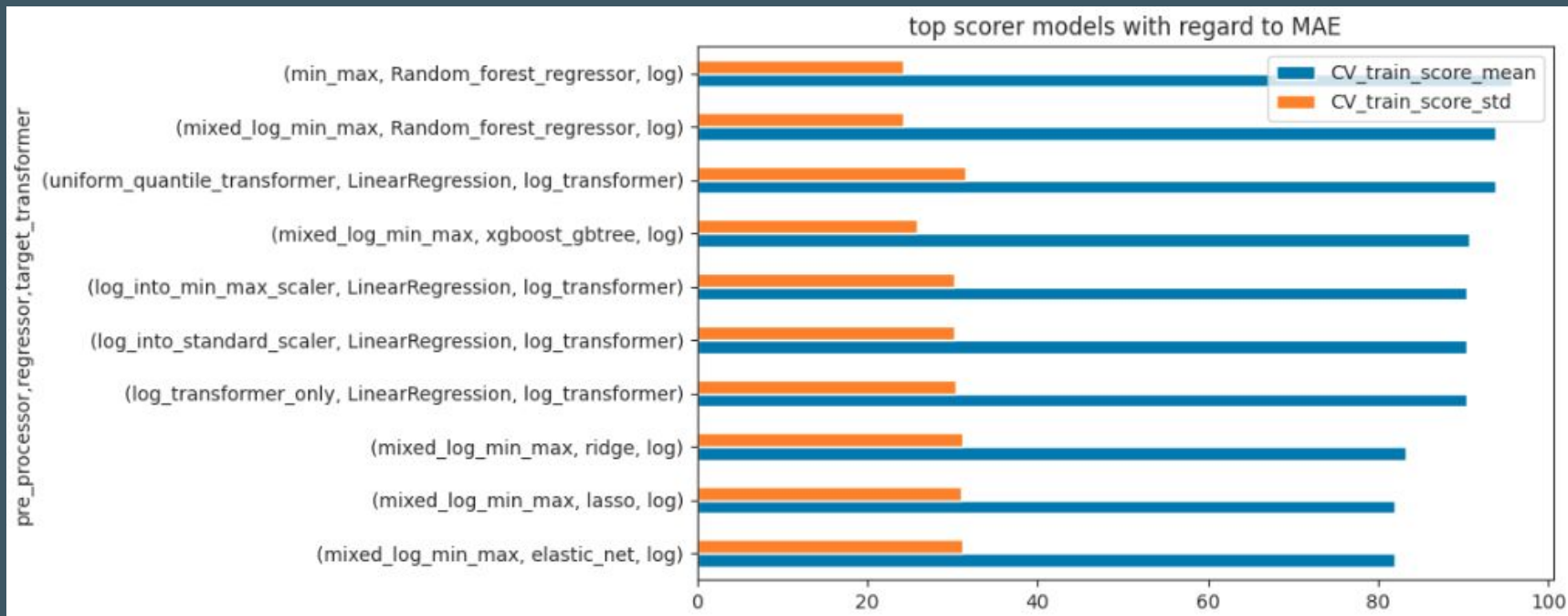
---

# Energy prediction (MAE)



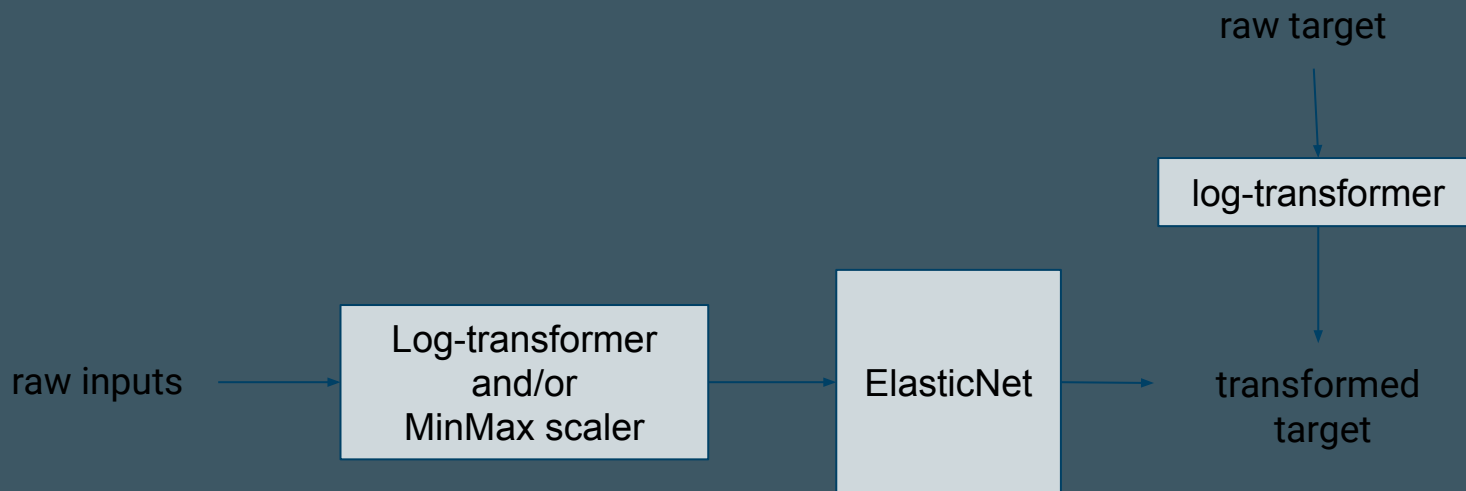
- Dummy regressor : MAE ~ 8e6 kBtu
- Regardless of the regressor, a log transformation on both skewed inputs and the target gave the best results.
- Linear regression models gave the best results.
- Interesting to note that the Xgboost model coming close after linear models has a lower variance

# GHG emissions (MAE)



- Dummy Regressor : MAE ~ 187 tons CO2 eq
- log transformation also very important (inputs and target)
- regularized version of linear models outperform the rest more significantly.

# Best model for both predictions



Note that results are fairly consistent when evaluating with the R2 , MAE and the RMSE metric.

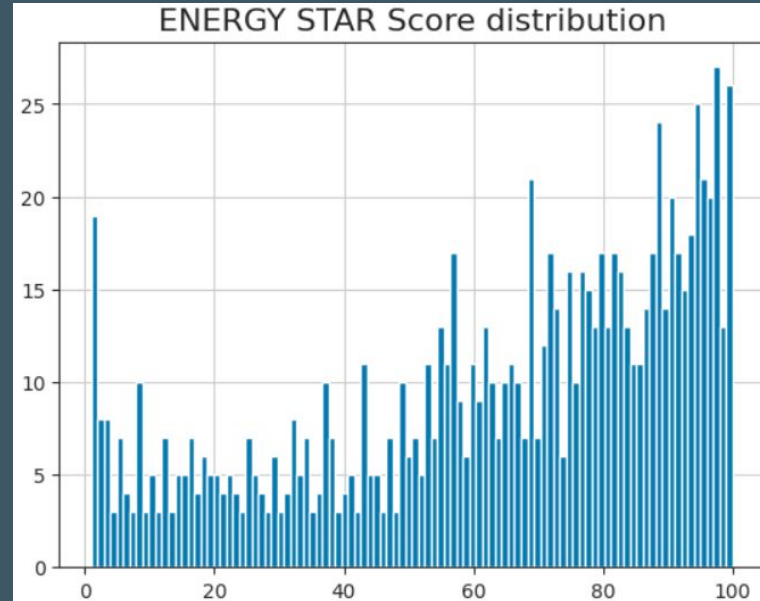


# Energy Star score

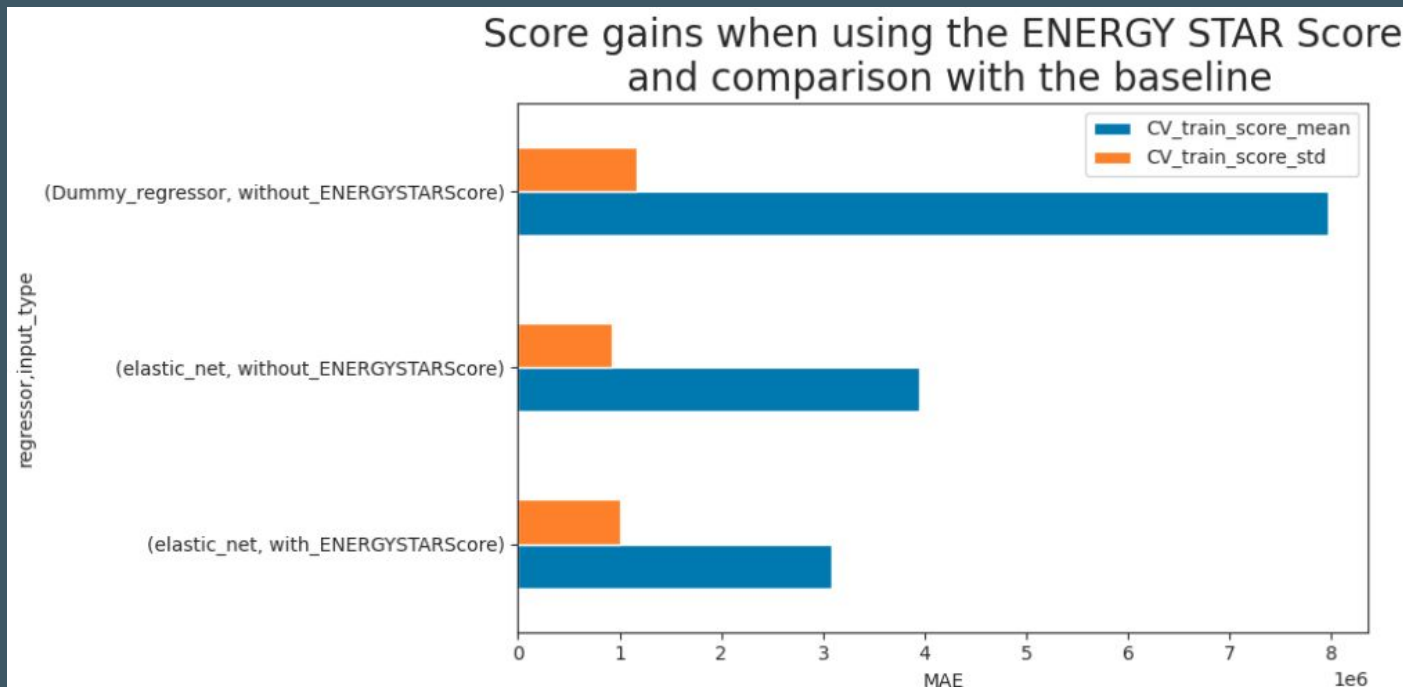
---

# What is the Energy Star Score ?

- available for 65 % of the selected properties.
- currently being generated by colleagues (expensive/time consuming).
- range between 0 and 100



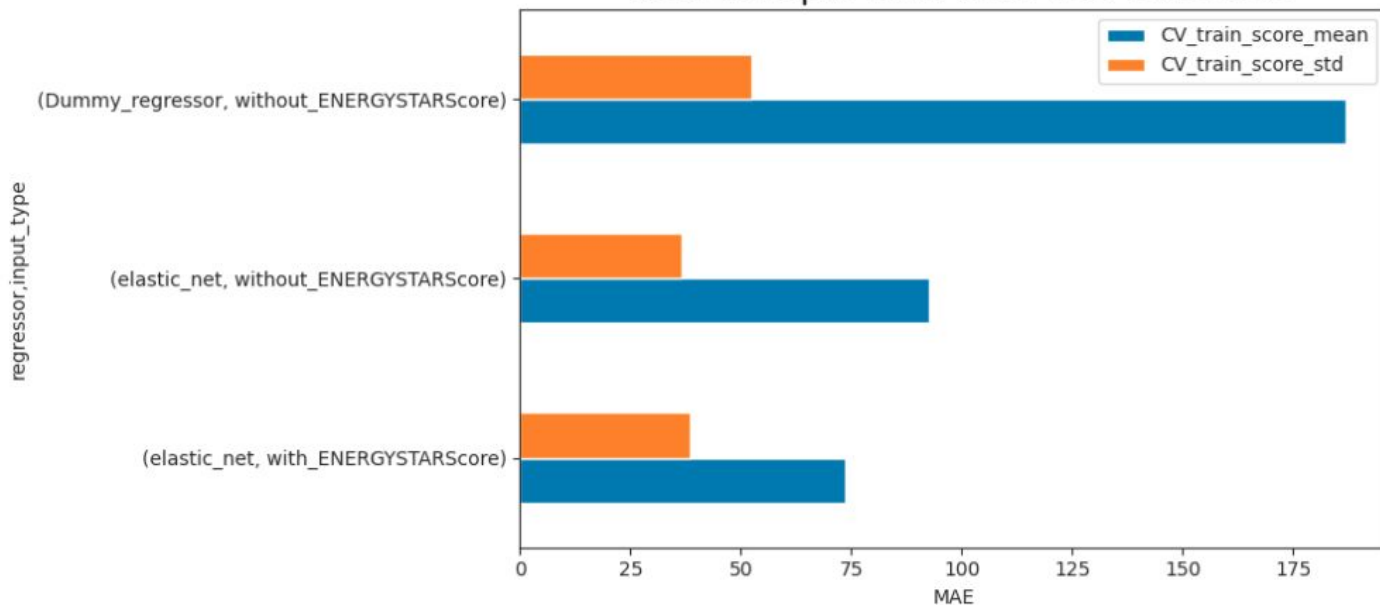
# Impact on the best model : ElasticNet (Energy consumptions)



- Adding the Energy Star Score to the inputs of the model increases the performance (~21,9% gain)
- Ultimately, the model has an MAE which is ~38,7% of the baseline's MAE.

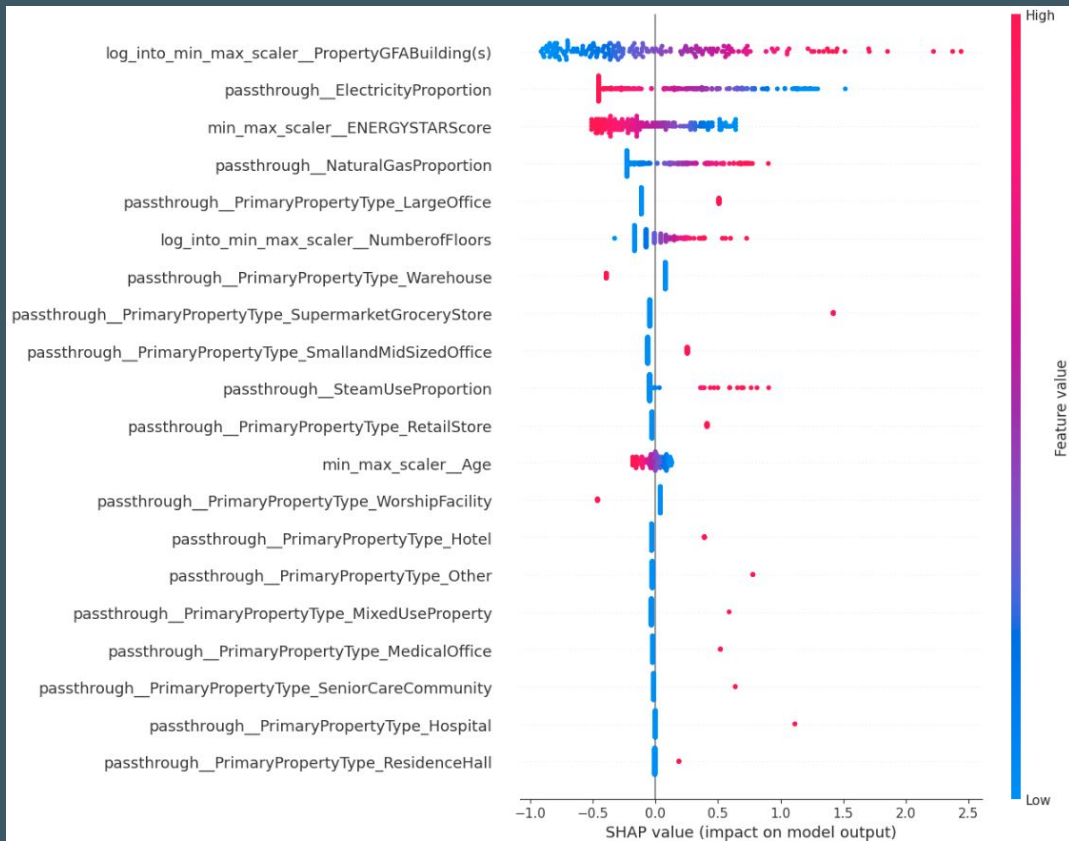
# Impact on the best model : ElasticNet (GHG emissions)

Score gains when using the ENERGY STAR Score  
and comparison with the baseline



- Adding the Energy Star Score to the inputs of the model increases the performance (~20,4% gain)
- Ultimately, the model has an MAE which is ~39,4% of the baseline's MAE.

# Features importance in the model : (GHG emissions)



*(Read the most important features from top to bottom.)*

- Energy Star Score ranked 3.

-> Not surprising considering the gains.

- Energy sources proportions are important.

# Conclusion

- For both predictions, the ElasticNet gives the best performance.
- In both cases, adding the Energy Star Score in the model's inputs improves the performance (~20% gain).

# Going further ?

Some ideas to improve the performance ?

- Address the energy sources proportions problem on some properties.
- Introduce location information.
- Retrieve number of buildings (very correlated with GFA).
- Continue to build the Energy Star Score for the missing non-residential properties.

Feature engineering :

- map each class of Primary, secondary and third type use (quite detailed) with the one-hot encoded variable classes AND weight the 1 with the respective GFA proportion.

# Thank you for listening

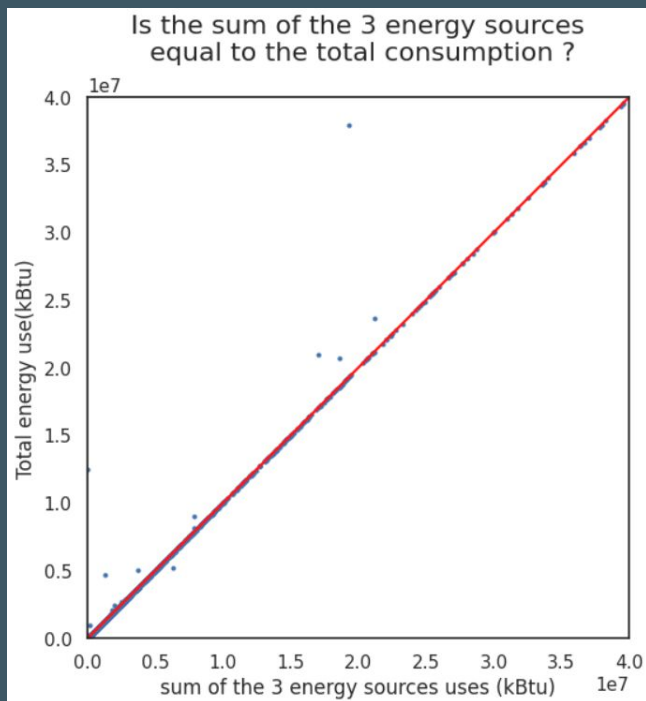
---



# Appendices

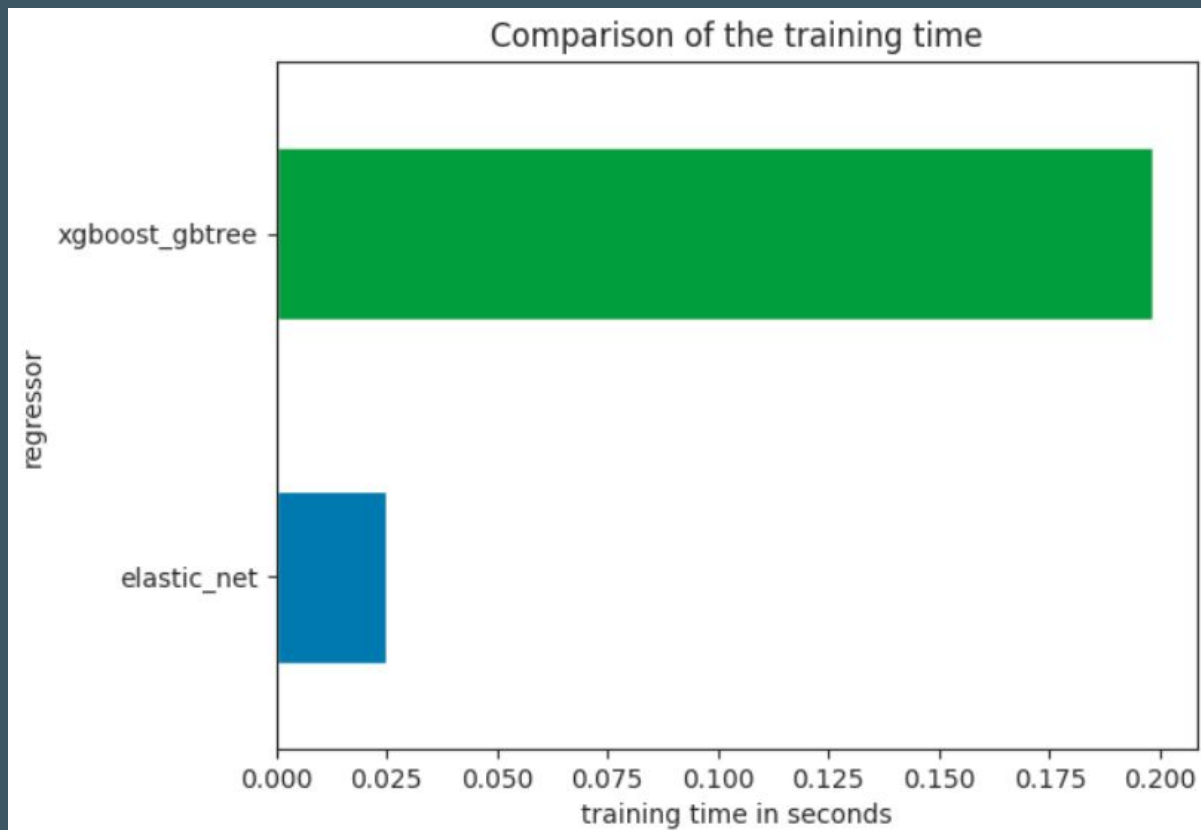
---

# Energy source proportions outliers ?



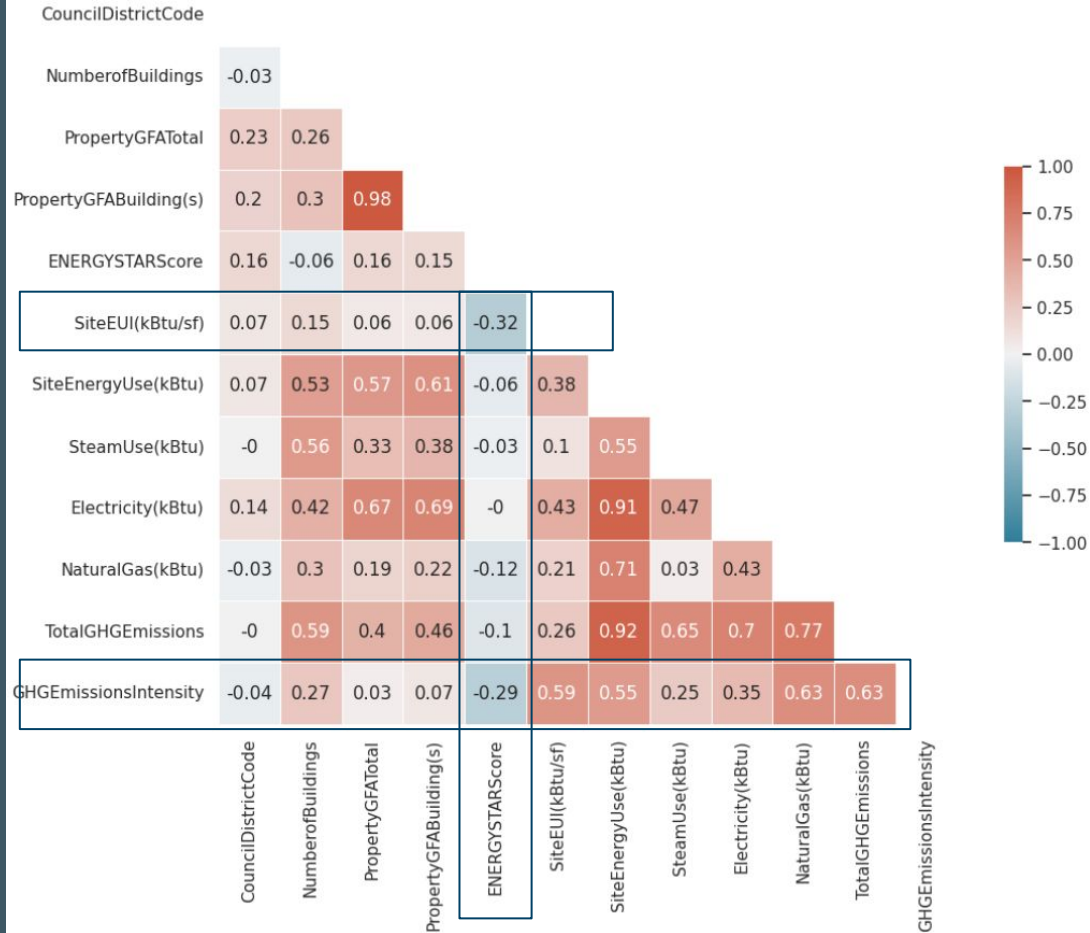
OSEBuildingID	PrimaryPropertyType	SiteEnergyUse(kBtu)	energies_sum	energy_ratio
103	Other	2.365898e+07	21200976.0	0.896107
106	Other	2.095503e+07	17016015.0	0.812025
112	Other	2.072325e+07	18649906.0	0.899951
328	Large Office	4.084775e+07	41377888.0	1.012978
561	Large Office	9.058916e+06	7877393.0	0.869573
700	Supermarket / Grocery Store	1.252517e+07	0.0	0.000000
757	Large Office	5.177270e+06	6287167.0	1.214379
803	Small- and Mid-Sized Office	3.795171e+07	19295187.0	0.508414
21436	Other	2.485521e+06	2021951.0	0.813492
24216	Small- and Mid-Sized Office	2.107744e+06	1815369.0	0.861285
26849	Retail Store	9.772303e+05	204995.0	0.209771
26973	Mixed Use Property	4.729846e+06	1323792.0	0.279881
49784	Small- and Mid-Sized Office	3.427261e+05	-115417.0	-0.336762
49967	University	8.739237e+08	742059629.0	0.849113
49968	University	5.000717e+06	3719217.0	0.743737
49972	University	5.116831e+07	28614613.0	0.559225

I Discarded the 3 properties highlighted here. Not all, in order to keep some universities in the dataset.  
Explanations for the other ratios ?



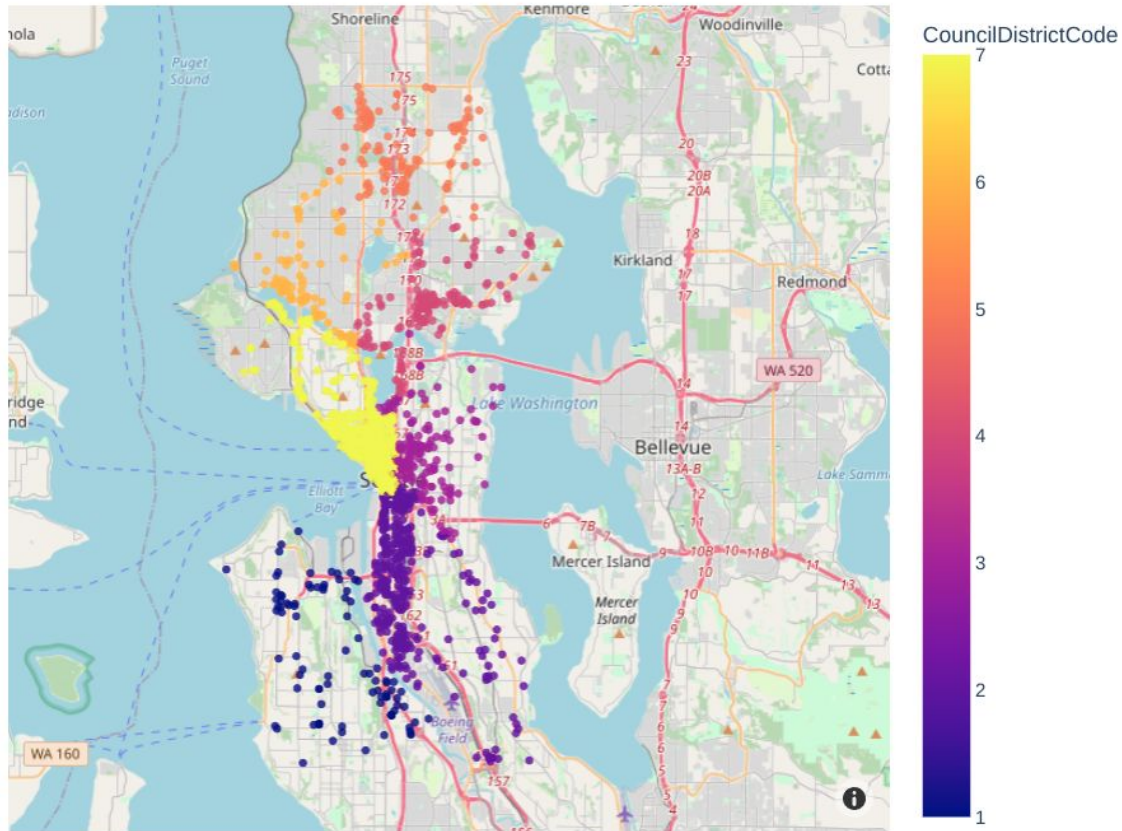
The Xgboost which could be a reasonable alternative to linear models is about 8 times longer to train on the training set.

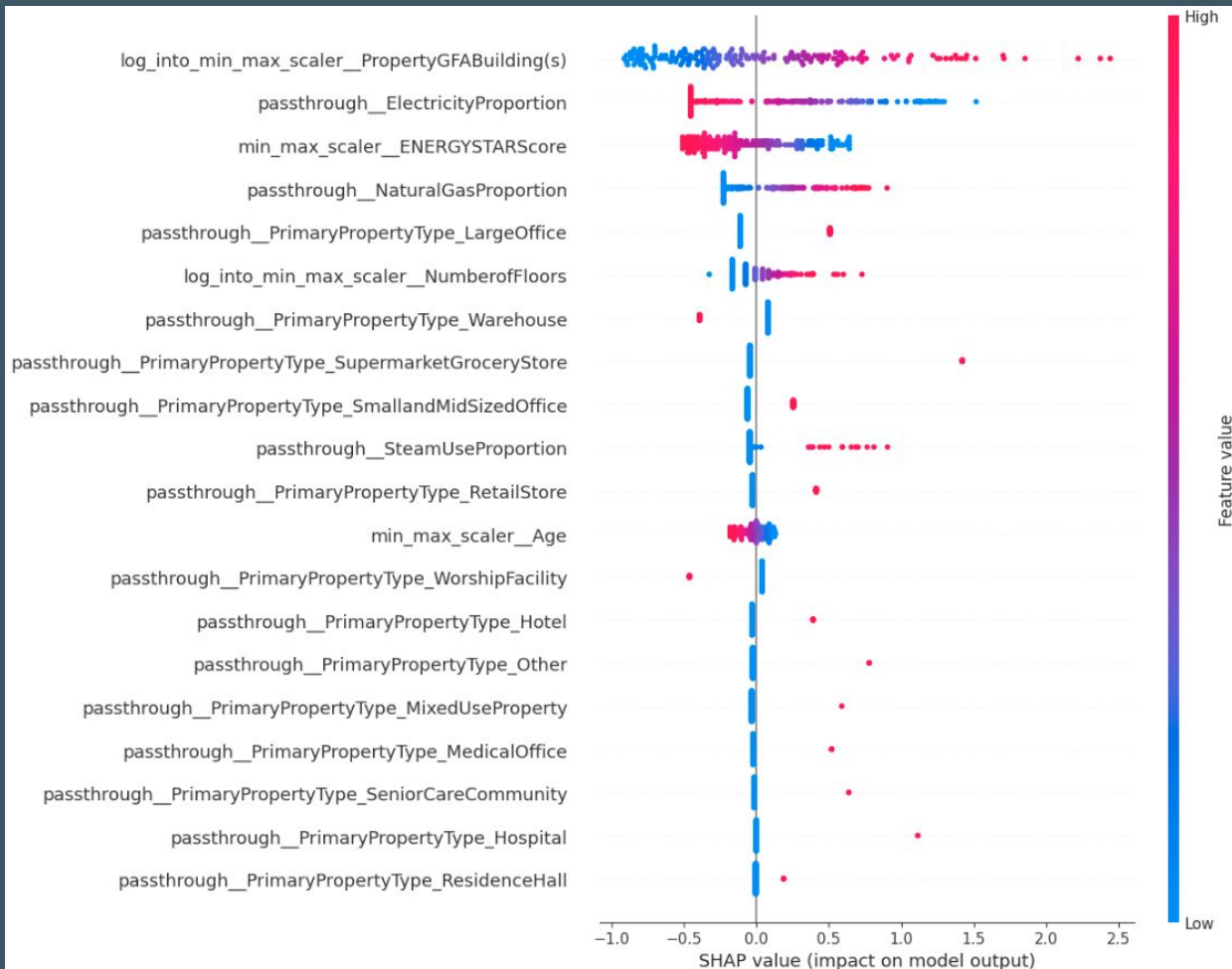
Correlation Matrix



High Energy Star score  
could mean lower values for both  
predictions.

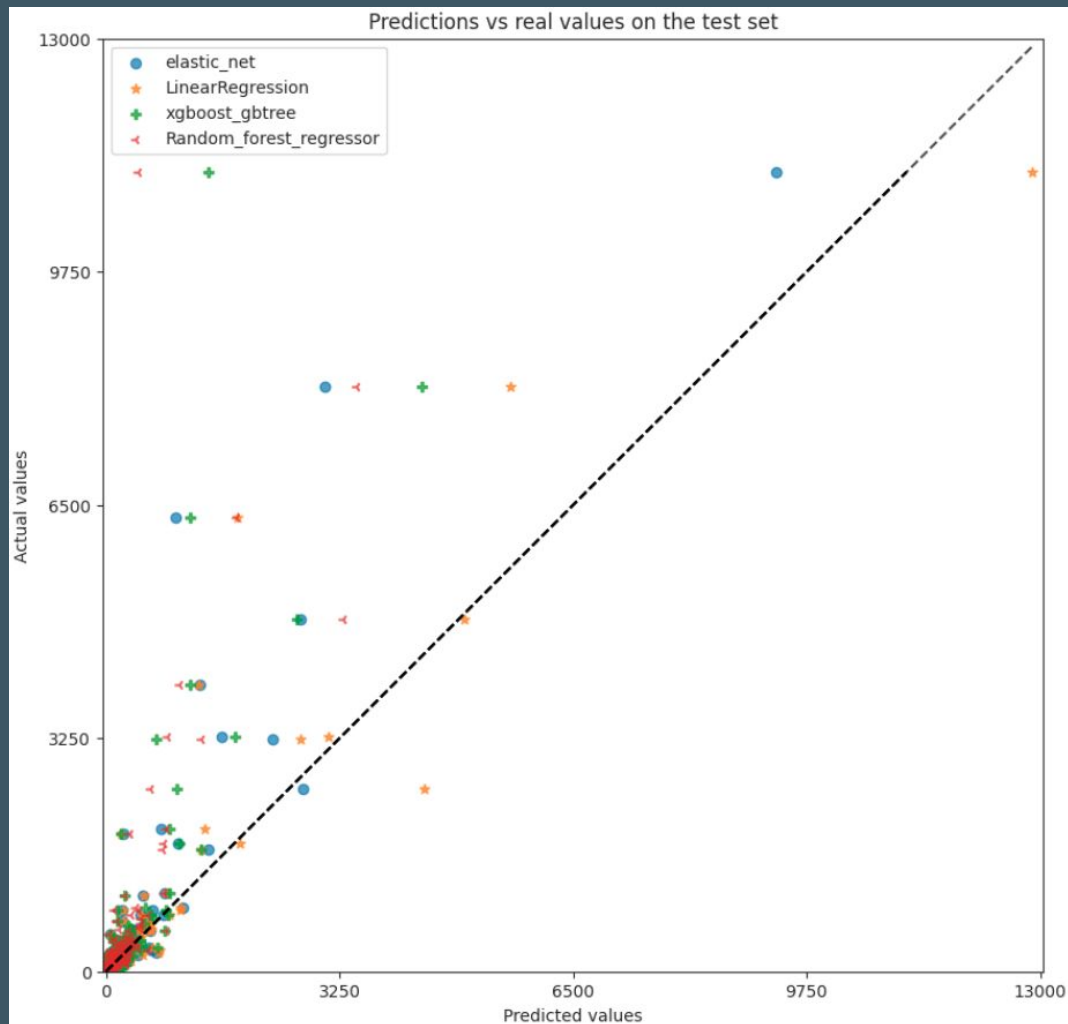
## Non-Residential Properties localizations





*Features importance and explainability with shap in the ElasticNet configured for GHG emission predictions*

It confirms the assumption made with the correlation matrix about the Energy StarScore. The higher the Energy Star Score, the less polluter is the the property.



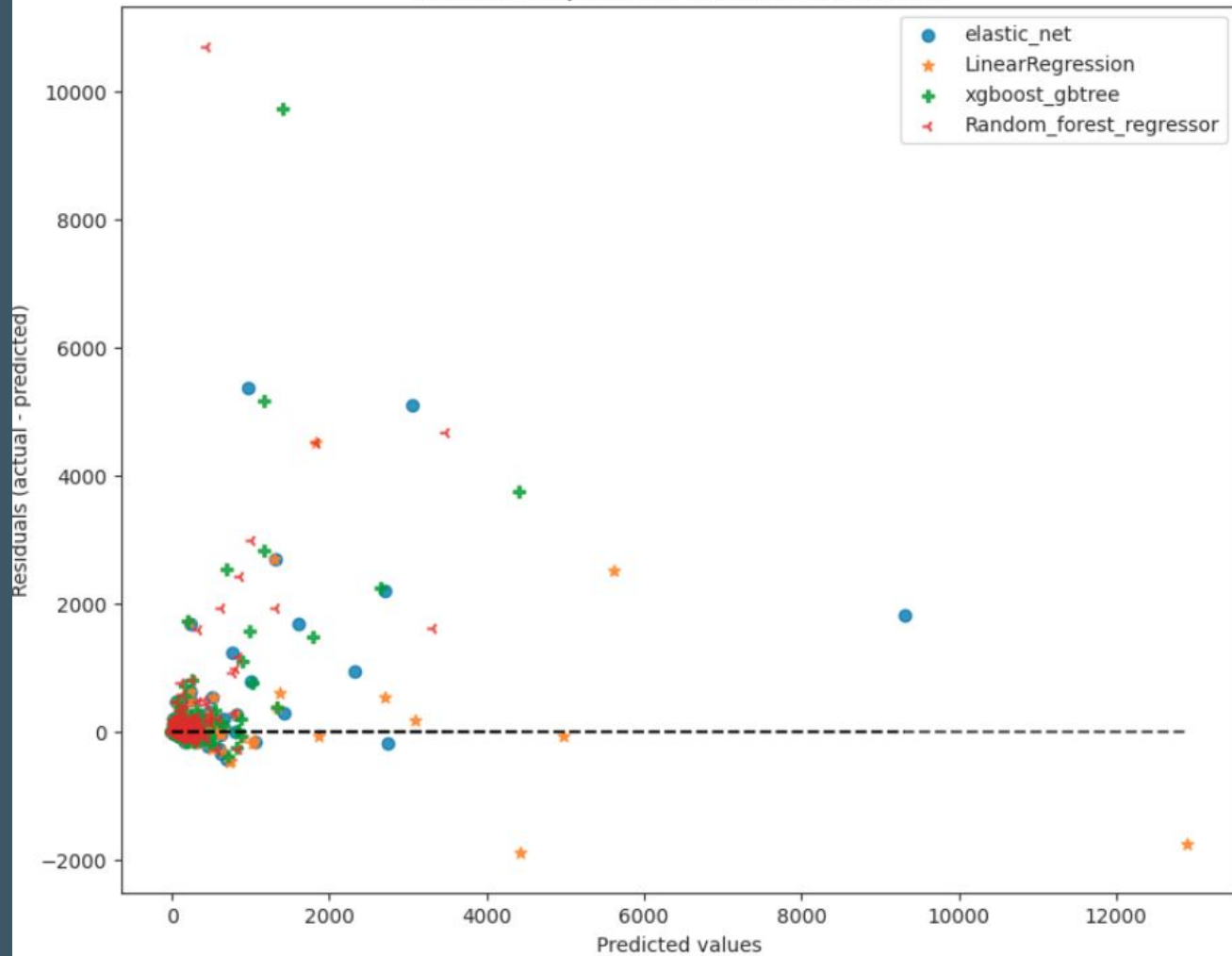
Linear models perform better for high values because they are not well represented in the dataset ?

Better capacity to extrapolate.

Would more data help tree-based to perform better?

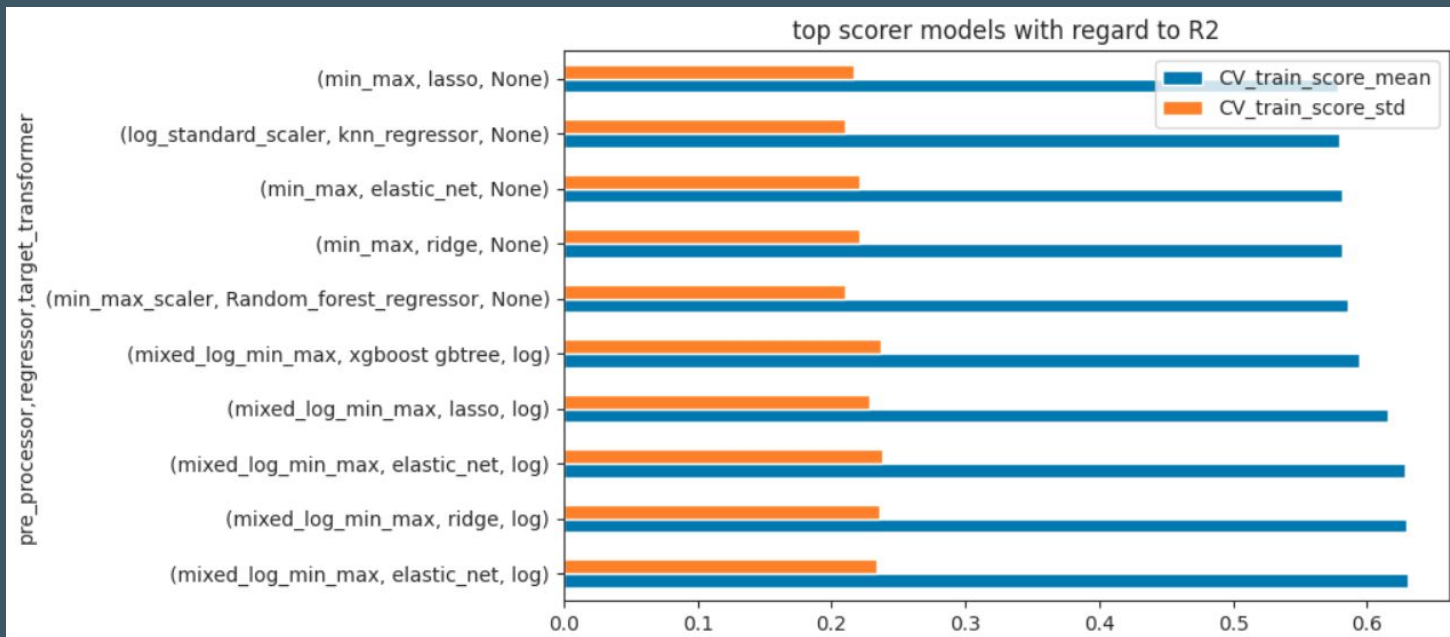
-> We will not get more data, because such properties are not going to be built massively.

Residuals vs predicted values on the test set



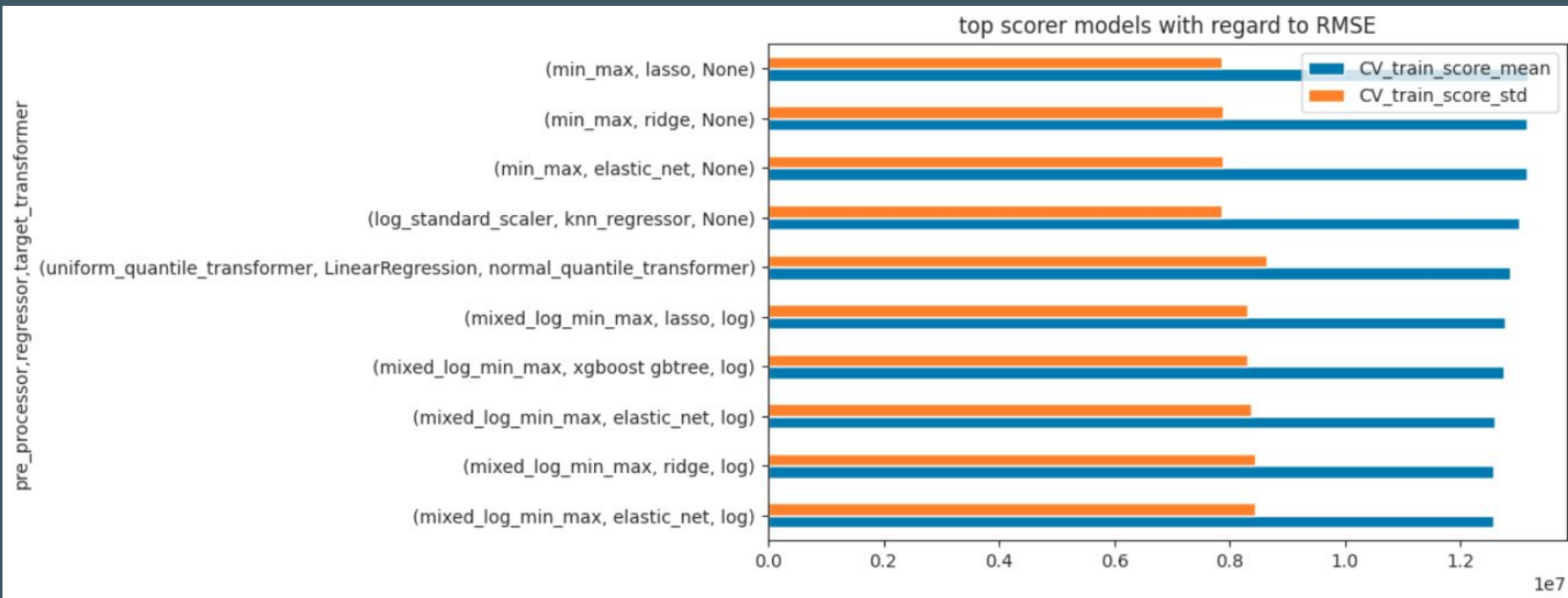


# Energy prediction (R2)



- Dummy regressor :  $R^2 \sim 0$
- Regardless of the regressor, a log transformation on both skewed inputs and the target gave the best results.
- Linear models give better results than tree-based models.

# Energy prediction (RMSE)



- Dummy regressor : RMSE  $\sim 3e7$  kBtu
- Idem