CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

P4 : Anticipez les besoins en consommation électrique de bâtiments

July 26, 2021

Carlos SACRISTAN

OpenClassrooms

Mission objective

Seattle

1. Predict the CO2 emissions and the total energies consumed

Table des matières

Index

Combining data

Seattle

1. **Unify ID**
2. Unify variable
3. df2015['Location']
4. Joining data sets

OSEBuildingID

Real number ($\mathbb{R}_{\geq 0}$)

UNIQUE

Lib: pandas_profiling

Combining data

1. Unify ID
2. **Unify variable**
3. df2015['Location']
4. Joining data sets

'ZipCode':'Zip Codes'

'Comments':'Comment'

'TotalGHGEmissions' :'GHGEmissions(MetricTonsCO2e)'

Combining data

Seattle

1. Unify ID
2. Unify variable
3. **df2015['Location']**
4. Joining data sets

ast.literal_eval

```
dfDATA_Original_15['Location'].iloc[0]
```

```
'{\'latitude\': \'47.61219025\', \'longitude\': \'-122.33799744\', \'human_address\'
\'{"address": "405 OLIVE WAY", "city": "SEATTLE", "state": "WA", "zip": "98101"}\'}'
```

Zip Codes	Latitude	Longitude	Address	City	State
98101	47.62500	-122.3125	405 OLIVE WAY	SEATTLE	0
98101	47.62500	-122.3125	724 PINE ST	SEATTLE	0

DataFrame: dfDATA_Original_15

Combining data

Seattle

1. Unify ID
2. Unify variable
3. df2015['Location']
4. **Joining data sets**

Update 2015:

combine_first

OSEBuildingID	2015	2016	Upgrade
20520	735,178	741,287	741,287
49990	nan	619,545	619,545
23937	2,509,232	2,506,928	2,506,928
49789	10,249,433	10,055,690	10,055,690
25217	1,866,800	1,968,907	1,968,907
25526	835,545	829,801	829,801
734	4,678,999	4,817,062	4,817,062
22585	766,628	845,044	845,044
794	1,796,889	1,951,856	1,951,856
765	0	16,246,106	16,246,106

DataFrame: dfDATA2

Dimensions : 3432 rows x 50 columns

Data cleansing

Seattle

1. **Homogenization of str values**
2. Last year of 'YearsENERGYSTARCertified'
3. Missing values

Fix to lowercases

Redundant white space

OSEBuildingID	2010 Census Tracts	Address	BuildingType
1	NaN	405 olive way	nonresidential
2	NaN	724 pine street	nonresidential
3	NaN	1900 5th avenue	nonresidential

DataFrame: dfDATA2

Data cleansing

Seattle

YearsENERGYSTARCertif...
Categorical

2016 15
20172016 8

Lib: pandas_profiling

1. Homogenization of str values
2. **Last year of 'YearsENERGYSTARCertified'**
3. Missing values

OSEBuildingID	
27685	NaN
20541	NaN
43	NaN
22361	NaN
26147	NaN
19949	NaN
49893	2017.0
28874	NaN
765	2014.0

DataFrame: dfDATA2

Data cleansing

Seattle

1. Homogenization of str values
2. Last year of 'YearsENERGYSTARCertified'
3. **Missing values**

	count	unique	freq	mean	std	min	max	%full	%unique
Comment	13.0	13.0	1.0	nan	nan	nan	nan	0.4	0.0
DefaultData	56.0	2.0	43.0	nan	nan	nan	nan	1.6	96.4
Outlier	97.0	2.0	49.0	nan	nan	nan	nan	2.8	97.9
YearsENERGYSTARCertified	123.0	nan	nan	2,015.0	2.4	2,007.0	2,017.0	3.6	nan
City Council Districts	213.0	nan	nan	1.3	0.5	1.0	2.0	6.2	nan
2010 Census Tracts	224.0	nan	nan	123.1	5.8	116.0	135.0	6.5	nan
ThirdLargestPropertyUseTypeGFA	607.0	nan	nan	11,675.7	29,083.5	0.0	459,748.0	17.7	nan
ThirdLargestPropertyUseType	607.0	44.0	112.0	nan	nan	nan	nan	17.7	92.8
SecondLargestPropertyUseTypeGFA	1,704.0	nan	nan	28,475.2	54,377.3	0.0	686,750.0	49.7	nan
SecondLargestPropertyUseType	1,704.0	50.0	992.0	nan	nan	nan	nan	49.7	97.1
ENERGYSTARScore	2,656.0	nan	nan	68.0	26.9	1.0	100.0	77.4	nan
GHGEmissionsIntensity(kgCO2e/ft2)	3,330.0	nan	nan	1.0	1.6	0.0	31.4	97.0	nan
OtherFuelUse(kBtu)	3,330.0	nan	nan	7,142.3	196,279.0	0.0	8,269,669.0	97.0	nan
rtment Micro Community Policing Plan Areas	3,338.0	nan	nan	32.4	19.9	1.0	61.0	97.3	nan
SPD Beats	3,338.0	nan	nan	24.8	15.0	1.0	51.0	97.3	nan
GHGEmissionsIntensity	3,367.0	nan	nan	1.2	1.8	-0.0	34.1	98.1	nan
LargestPropertyUseType	3,402.0	57.0	1,689.0	nan	nan	nan	nan	99.1	98.3
LargestPropertyUseTypeGFA	3,402.0	nan	nan	79,170.6	200,946.8	5,656.0	9,320,156.0	99.1	nan
ListOfAllPropertyUseTypes	3,408.0	472.0	875.0	nan	nan	nan	nan	99.3	86.2
Zip Codes	3,416.0	nan	nan	98,116.7	18.6	98,006.0	98,272.0	99.5	nan

DataFrame: dfTAUX

First filtering

1. Missing values
2. Drop missing values < 50%
3. Unique = 1
4. Duplicate rows
5. Irrelevant features

	count	unique	freq	mean	std	min	max	%full	%unique
Comment	13.0	13.0	1.0	nan	nan	nan	nan	0.4	0.0
DefaultData	56.0	2.0	43.0	nan	nan	nan	nan	1.6	96.4
Outlier	97.0	2.0	49.0	nan	nan	nan	nan	2.8	97.9
YearsENERGYSTARCertified	123.0	nan	nan	2,015.0	2.4	2,007.0	2,017.0	3.6	nan
City Council Districts	213.0	nan	nan	1.3	0.5	1.0	2.0	6.2	nan
2010 Census Tracts	224.0	nan	nan	123.1	5.8	116.0	135.0	6.5	nan
ThirdLargestPropertyUseTypeGFA	607.0	nan	nan	11,675.7	29,083.5	0.0	459,748.0	17.7	nan
ThirdLargestPropertyUseType	607.0	44.0	112.0	nan	nan	nan	nan	17.7	92.8
SecondLargestPropertyUseTypeGFA	1,704.0	nan	nan	28,475.2	54,377.3	0.0	686,750.0	49.7	nan
SecondLargestPropertyUseType	1,704.0	50.0	992.0	nan	nan	nan	nan	49.7	97.1
ENERGYSTARScore	2,656.0	nan	nan	68.0	26.9	1.0	100.0	77.4	nan
GHGEmissionsIntensity(kgCO2e/ft2)	3,330.0	nan	nan	1.0	1.6	0.0	31.4	97.0	nan
OtherFuelUse(kBtu)	3,330.0	nan	nan	7,142.3	196,279.0	0.0	8,269,669.0	97.0	nan
rtment Micro Community Policing Plan Areas	3,338.0	nan	nan	32.4	19.9	1.0	61.0	97.3	nan
SPD Beats	3,338.0	nan	nan	24.8	15.0	1.0	51.0	97.3	nan
GHGEmissionsIntensity	3,367.0	nan	nan	1.2	1.8	-0.0	34.1	98.1	nan
LargestPropertyUseType	3,402.0	57.0	1,689.0	nan	nan	nan	nan	99.1	98.3
LargestPropertyUseTypeGFA	3,402.0	nan	nan	79,170.6	200,946.8	5,656.0	9,320,156.0	99.1	nan
ListOfAllPropertyUseTypes	3,408.0	472.0	875.0	nan	nan	nan	nan	99.3	86.2
Zip Codes	3,416.0	nan	nan	98,116.7	18.6	98,006.0	98,272.0	99.5	nan

DataFrame: dfTAUX

First filtering

Seattle

	count	unique	freq	mean	std	min	max	%full	%unique
Comment	13.0	13.0	1.0	nan	nan	nan	nan	0.4	0.0
DefaultData	56.0	2.0	43.0	nan	nan	nan	nan	1.6	96.4
Outlier	97.0	2.0	49.0	nan	nan	nan	nan	2.8	97.9
YearsENERGYSTARCertified	123.0	nan	nan	2,015.0	2.4	2,007.0	2,017.0	3.6	nan
City Council Districts	213.0	nan	nan	1.3	0.5	1.0	2.0	6.2	nan
2010 Census Tracts	224.0	nan	nan	123.1	5.8	116.0	135.0	6.5	nan
ThirdLargestPropertyUseTypeGFA	607.0	nan	nan	11,675.7	29,083.5	0.0	459,748.0	17.7	nan
ThirdLargestPropertyUseType	607.0	44.0	112.0	nan	nan	nan	nan	17.7	92.8
SecondLargestPropertyUseTypeGFA	1,704.0	nan	nan	28,475.2	54,377.3	0.0	686,750.0	49.7	nan
SecondLargestPropertyUseType	1,704.0	50.0	992.0	nan	nan	nan	nan	49.7	97.1

DataFrame: dfTAUX

1. Missing values
2. **Drop missing values $< 50\%$**
3. Unique $= 1$
4. Duplicate rows
5. Irrelevant features

FEATURES DALETED :
'SecondLargestPropertyUseType', 'SecondLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseType', 'ThirdLargestPropertyUseTypeGFA', '2010 Census Tracts', 'City Council Districts', 'Outlier', 'DefaultData', 'Comment'

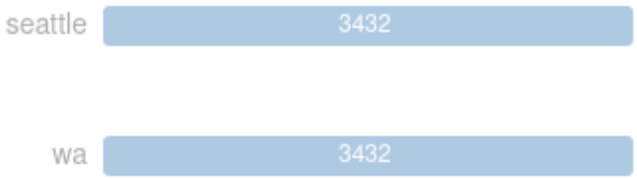
FEATURES EXEPTION :
'YearsENERGYSTARCertified'

Dimensions : 3432 rows x 41 columns

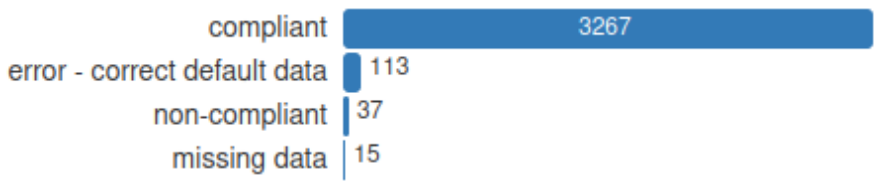
First filtering

1. Missing values
2. Drop missing values < 50%
3. **Unique = 1**
4. Duplicate rows
5. Irrelevant features

FEATURES DALETED : 'City', 'State', 'Date Year'



FEATURES DALETED : 'ComplianceStatus feature' (only 'compliant' value selected)



Lib: pandas_profiling

Dimensions : 3267 rows x 37 columns

First filtering

1. Missing values
2. Drop missing values < 50%
3. Unique = 1
4. **Duplicate rows**
5. Irrelevant features

0 duplicate rows

```
(3267, 37)
(3267, 37)
```

DataFrame: dfDATA2

Dimensions : 3432 rows x 37 columns

First filtering

1. Missing values
2. Drop missing values < 50%
3. Unique = 1
4. Duplicate rows
5. **Irrelevant features**

IRRELEVANT FEATURES DALETED (16) :

- 'Electricity(kWh)', 'GHGEmissionsIntensity', 'GHGEmissionsIntensity(kgCO2e/ft2)', 'NaturalGas(therms)', 'SiteEUIWN(kBtu/sf)',
- 'SiteEnergyUseWN(kBtu)', 'SourceEUI(kBtu/sf)', 'SourceEUIWN(kBtu/sf)', 'SiteEUI(kBtu/sf)'
- 'PropertyName', 'LargestPropertyUseType', 'LargestPropertyUseTypeGFA', 'ListOfAllPropertyUseTypes',
- 'TaxParcelIdentificationNumber'
- 'Address', 'Seattle Police Department Micro Community Policing Plan Areas'

Dimensions : 3432 rows x 21 columns

Physically impossible values

Seattle

- "Energy values" (Electricity, GHGEmissions, ...) >0
- 'CouncilDistrictCode', 'Zip Codes', 'YearBuilt', 'Latitude'
- 'NumberofBuildings', 'NumberofFloors', 'ENERGYSTARScore', 'PropertyGFABuilding(s)'
- **FEATURES EXEPTION :** 'Longitude'

1. **Features value > 0**
2. Energy < SiteEnergyUse
3. $\sum$Energy < SiteEnergyUse

OSEBuildingID	BuildingType	CouncilDistrictCode	ENERGYSTARScore	Electricity(kBtu)	GHGEmissions(MetricTonsCO2e)
1	nonresidential	7	60	3946027.0	249.98
2	nonresidential	7	61	3242851.0	295.86
3	nonresidential	7	43	49526664.0	2089.28
5	nonresidential	7	56	2768924.0	286.43
8	nonresidential	7	75	5368607.0	505.01
...	...	...	...	...	...
50049	multifamily lr (1-4)	2	58	1248599.0	8.7
50055	multifamily mr (5-9)	4	96	1189427.0	31.46
50057	multifamily hr (10+)	7	79	9164908.0	395.26
50058	multifamily lr (1-4)	4	80	783346.0	5.46
50059	multifamily lr (1-4)	4	70	966812.0	6.74

2444 rows × 21 columns

DataFrame: dfDATA2

Dimensions : 2444 rows x 21 columns

Physically impossible values

Seattle

1. Features value > 0
2. **Energy $<$ SiteEnergyUse**
3. $\sum$Energy $<$ SiteEnergyUse

OSEBuildingID	Electricity(kBtu)	SiteEnergyUse(kBtu)
266	1212601.0	1212551.0
325	9899135.0	9898724.0
442	16760914.0	16760217.0
490	3632614.0	3632613.75

DataFrame: dfDATA2

Dimensions : 2043 rows x 21 columns

Physically impossible values

1. Features value > 0
2. Energy < SiteEnergyUse
3. $\sum$**Energy < SiteEnergyUse**

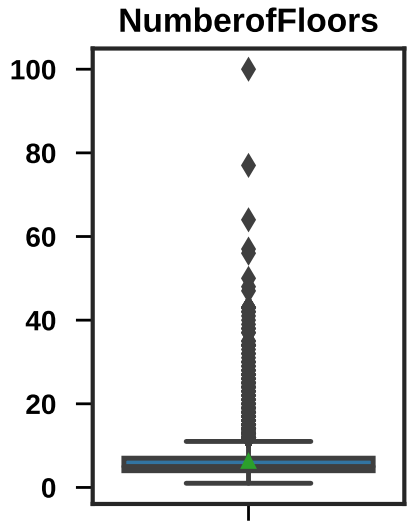
OSEBuildingID	Electricity(kBtu)	NaturalGas(kBtu)	OtherFuelUse(kBtu)	SteamUse(kBtu)	TotalEnergy(kBtu)
1	3946027.0	1276453.0	0	2003882.0	7226362.0
2	3242851.0	5145082.0	0	0.0	8387933.0
3	49526664.0	1493800.0	0	21566554.0	72587018.0
5	2768924.0	1811213.0	0	2214446.25	6794583.25
8	5368607.0	8803998.0	0	0.0	14172605.0

DataFrame: dfDATA2

Dimensions : 2026 rows x 22 columns

Outliers

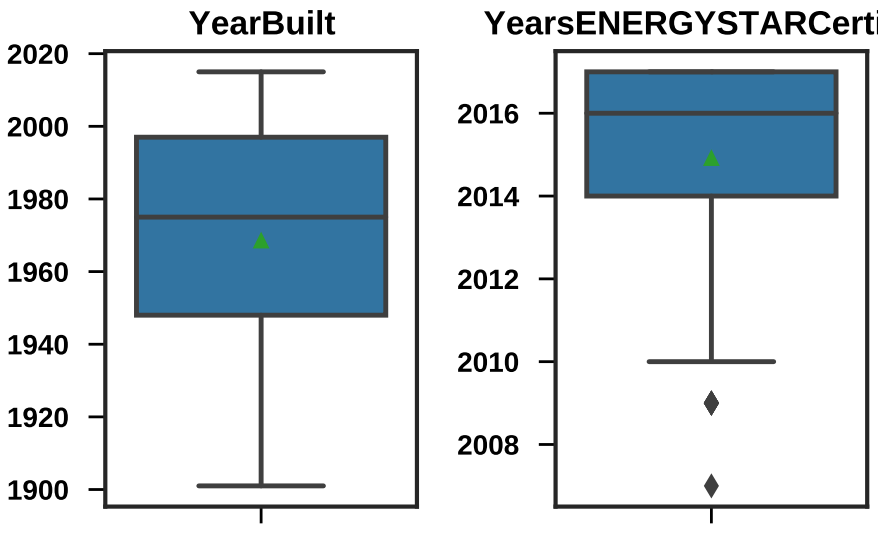
1. **'NumberofFloors' +1**
2. 'YearsENERGYSTARCertified', 'YearBuilt' > 1900
3. 'Zip Codes' > 80000
4. 'NumberofBuildings' = 1
5. Decade classification



DataFrame: dfDATA2

Outliers

1. 'NumberofFloors' +1
2. **'YearsENERGYSTARCertifie**
   'YearBuilt' > 1900
3. 'Zip Codes' > 80000
4. 'NumberofBuildings' = 1
5. Decade classification

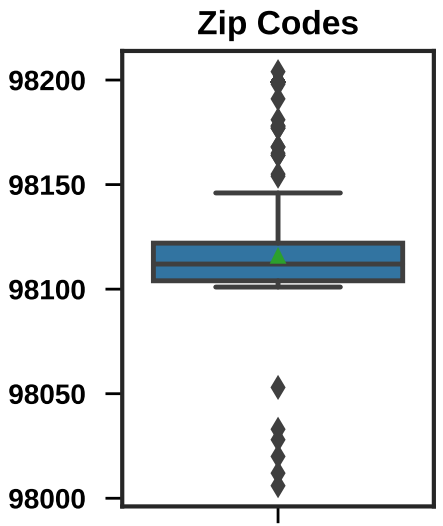


DataFrame: dfDATA2

Outliers

Seattle

1. 'NumberofFloors' +1
2. 'YearsENERGYSTARCertified', 'YearBuilt' > 1900
3. **'Zip Codes' > 80000**
4. 'NumberofBuildings' = 1
5. Decade classification

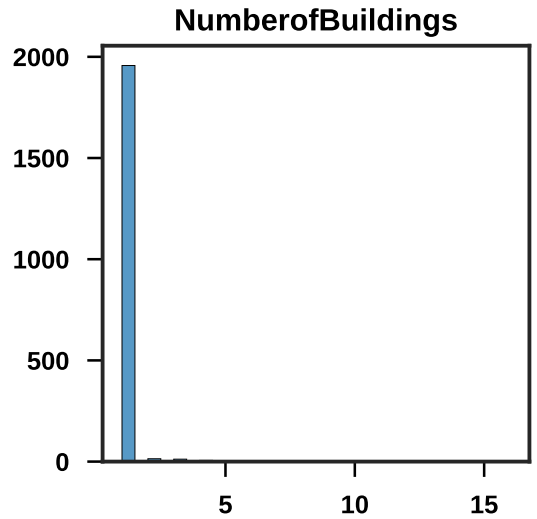


DataFrame: dfDATA2

DataFrame: dfDATA2

Outliers

1. 'NumberofFloors' +1
2. 'YearsENERGYSTARCertified', 'YearBuilt' > 1900
3. 'Zip Codes' > 80000
4. **'NumberofBuildings' = 1**
5. Decade classification



DataFrame: dfDATA2

Dimensions : 1957 rows x 21 columns

Outliers

Seattle

1. 'NumberofFloors' +1
2. 'YearsENERGYSTARCertified', 'YearBuilt' > 1900
3. 'Zip Codes' > 80000
4. 'NumberofBuildings' = 1
5. **Decade classification**

'YearBuilt_10y'

'YearsENERGYSTARCertified_10y'

OSEBuildingID	YearBuilt_10y	YearsENERGYSTARCertified_10y
21652	1920.0	NaN
482	2000.0	2010.0
422	1980.0	2000.0
26677	1990.0	NaN
21537	1960.0	NaN

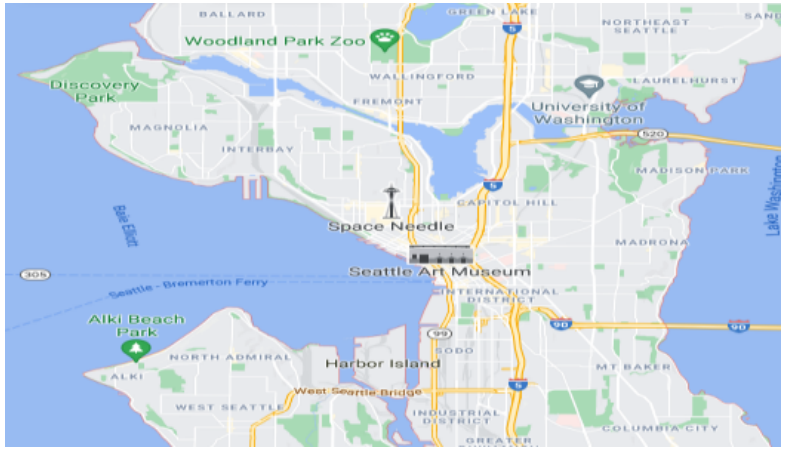
DataFrame: dfDATA2

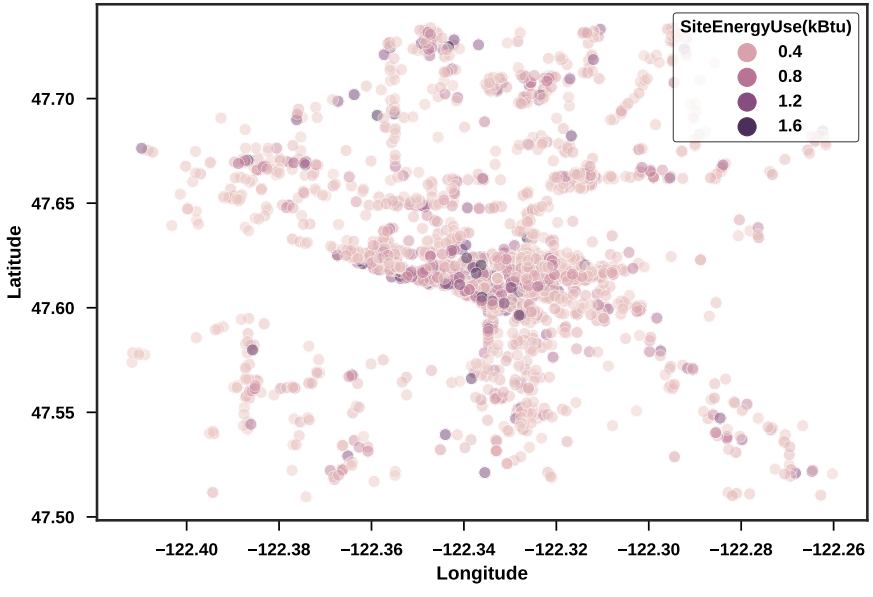
Dimensions : 1957 rows x 23 columns

Index

Analyse graphique

1. **EnergySiteUse - Map**
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global



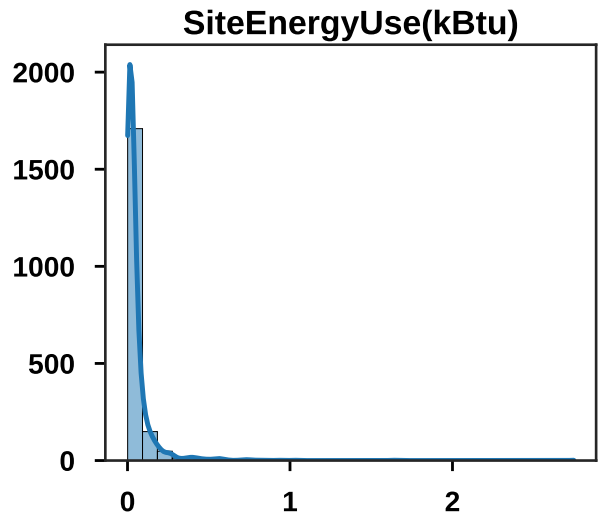


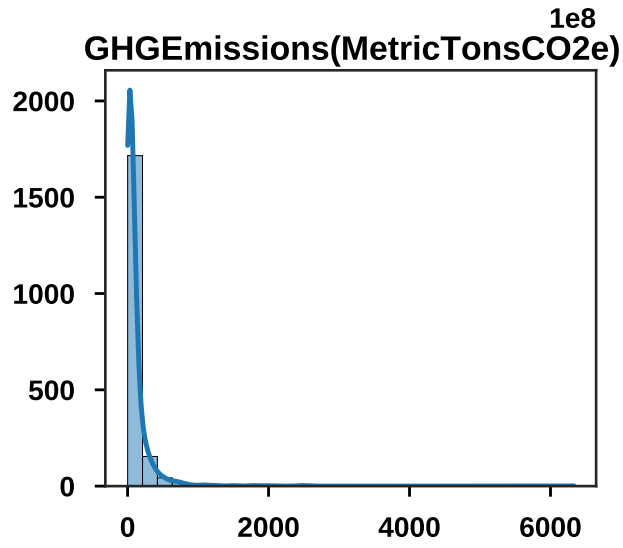
DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. **Distribution Univariée**
   - **Cible**
   - **Energy**
   - **Other**
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

- **Privilegier log scale**



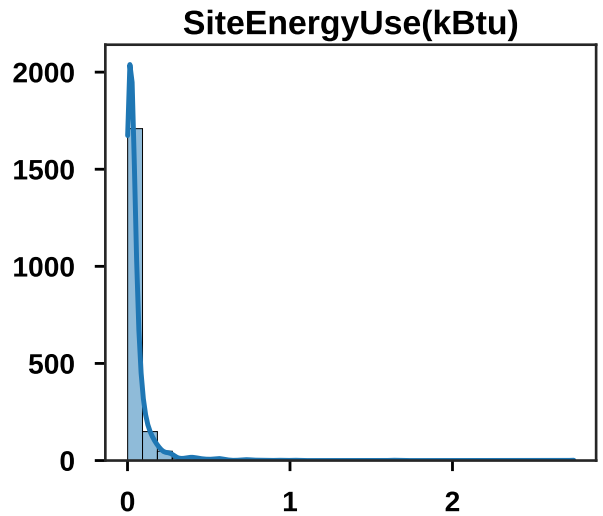


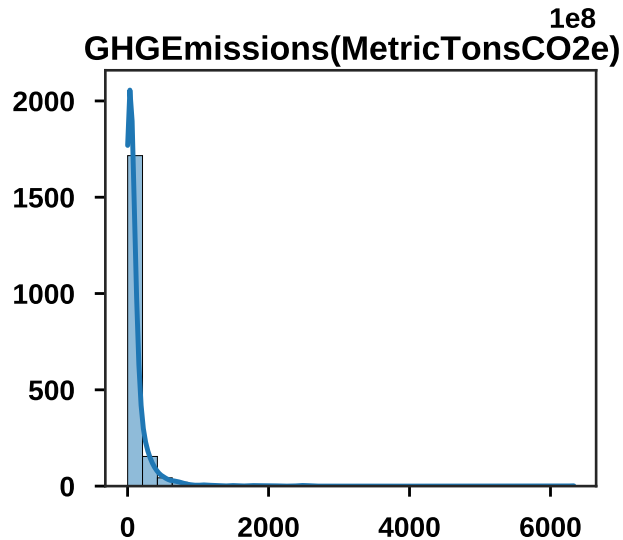
DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - **Cible**
   - Energy
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

- **Privilegier log scale**



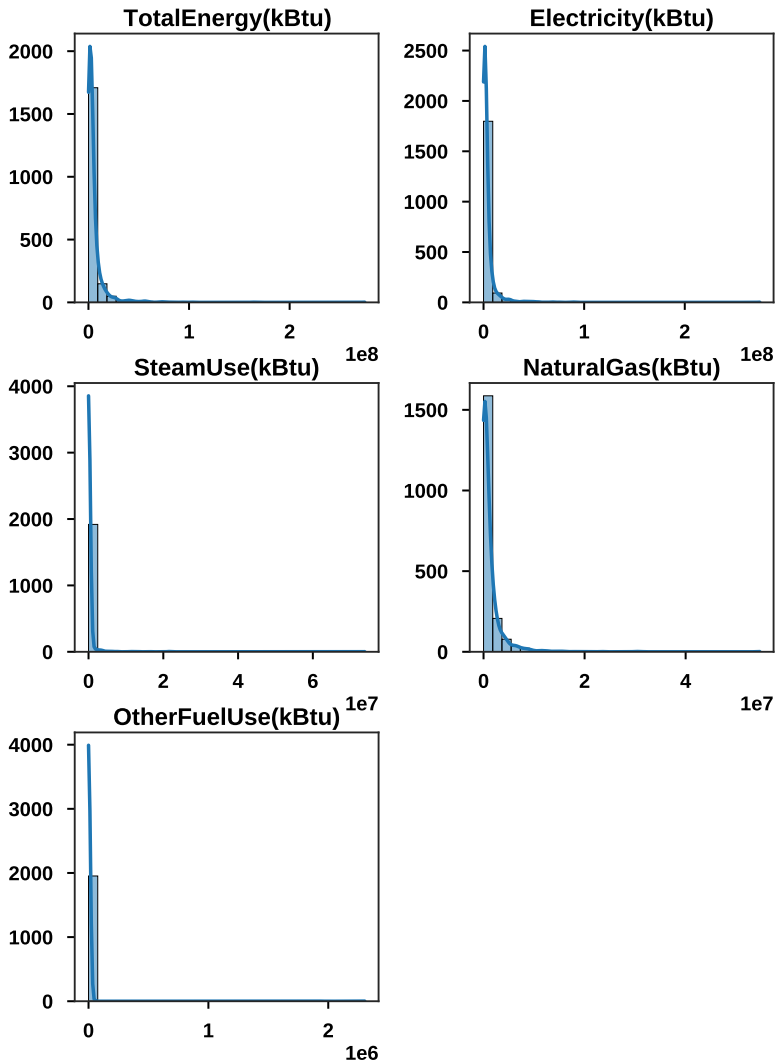


DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - **Energy**
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

- **Variables continues**

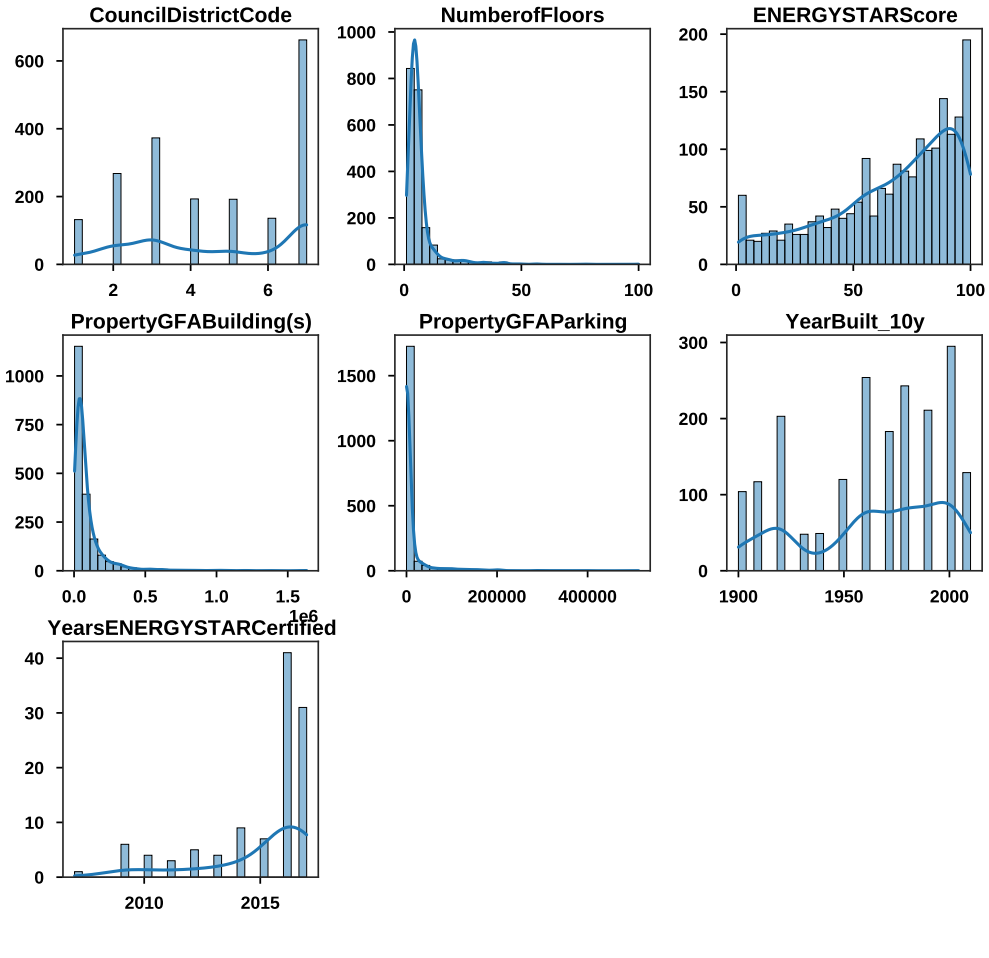


DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - **Other**
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

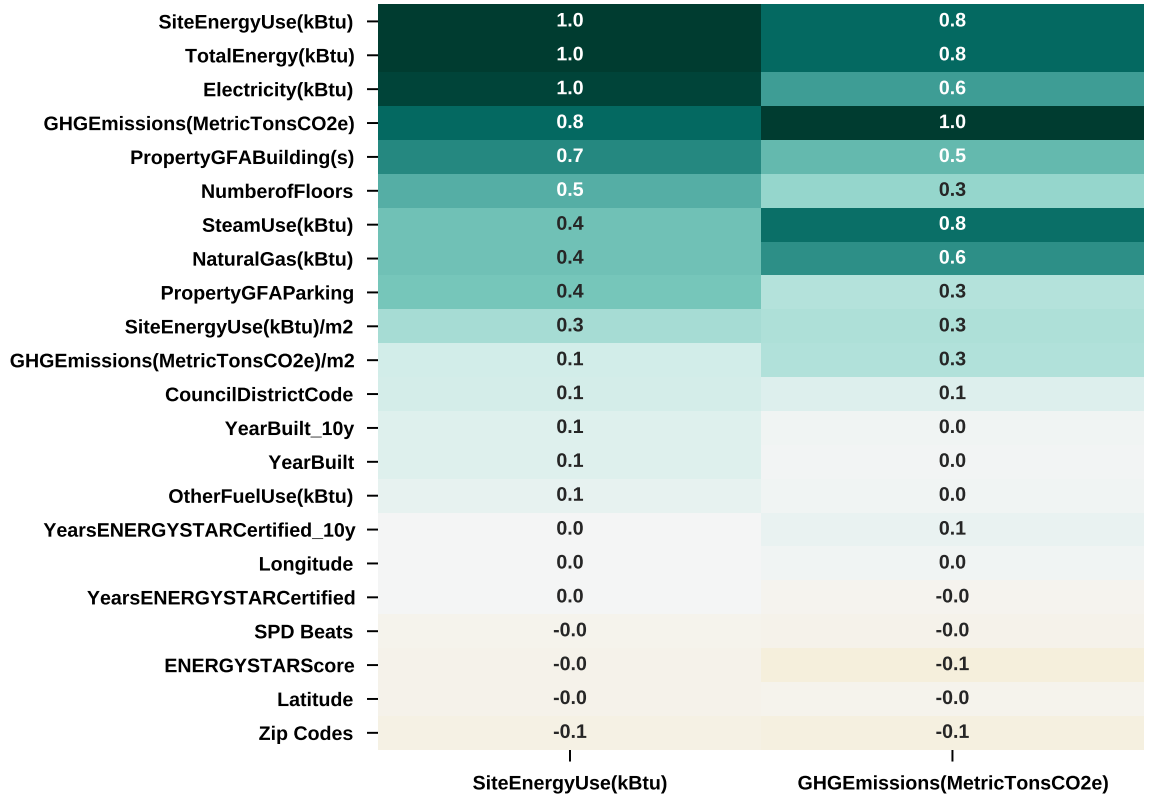
- **Variables continues:** PropertyGFABuildings, PropertyGFAParking
- **Variables categories:** CouncilDistrictCode, NumberofFloors, ZipCode, ENERGYSTARTscore, YearBuilt



DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. **Distribution Bivariée**
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

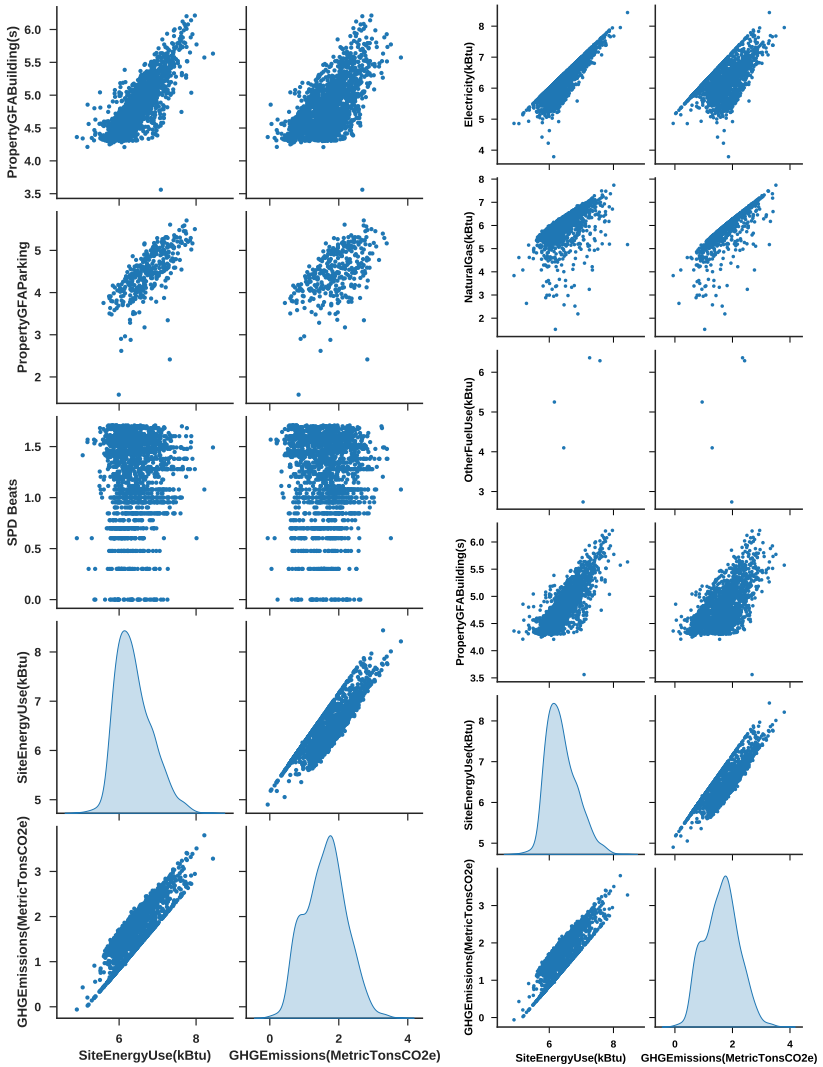


DataFrame: dfDATA2

1. **SiteEnergyUse :** TotalEnergy(Electricity), GHGEmissions / PropertyGFABuildings(s), NumberofFloors
2. **GHGEmissions :** SiteEnergyUse, TotalEnergy(SteamUse, Electricity) / NaturalGas, PropertyGFABuildings(s)

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. Distribution Bivariée
4. **Variables continue**
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

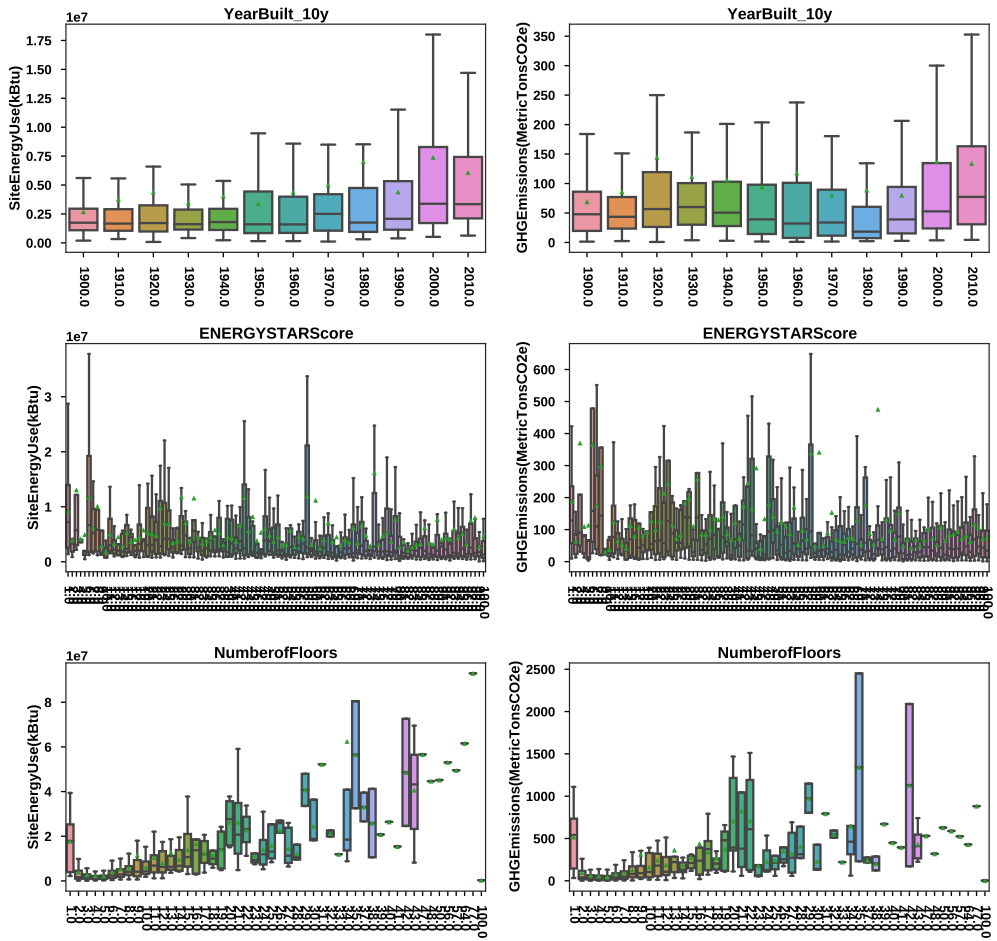


DataFrame: dfDATA2

Analyse graphique

- ↗YearBuilt $\Longrightarrow$↑SiteEnergyUse, ↗GHGEmissions

- ↗NumberofFloors$\Longrightarrow$↑↑SiteEnergyUse, ↑↑GHGEmissions



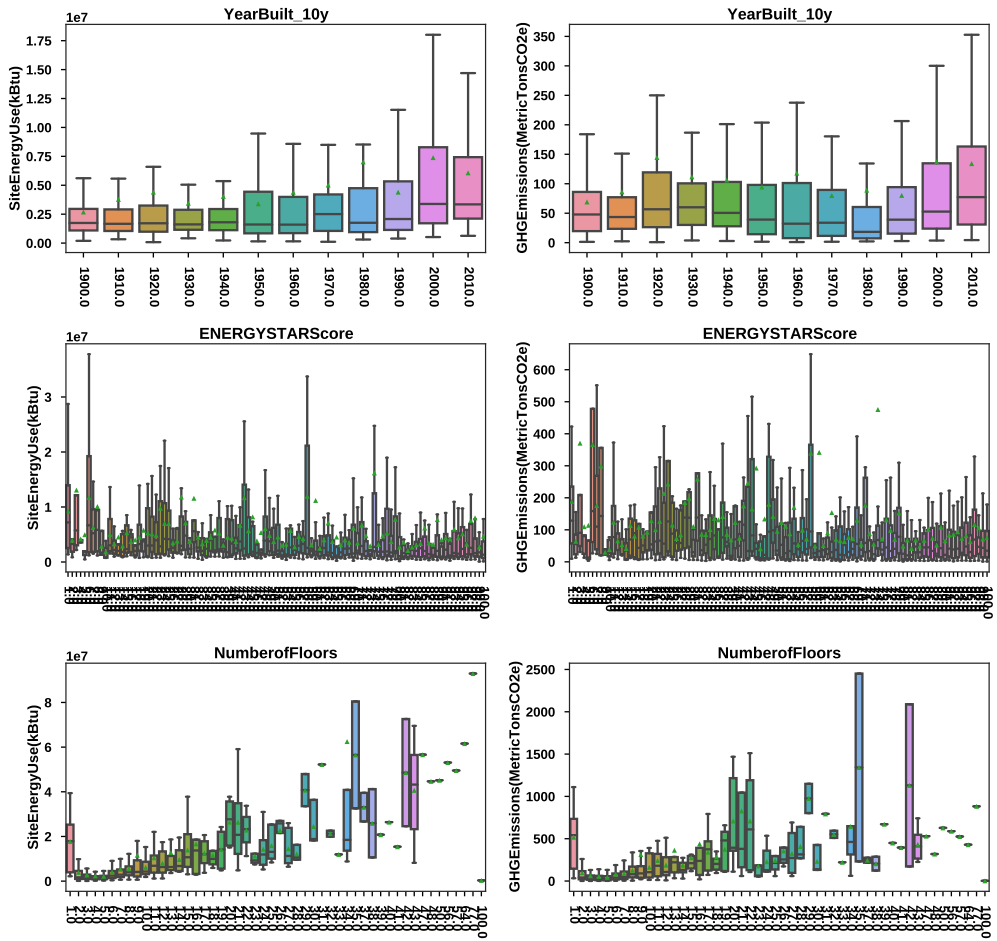
DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - **YearBuilt, ENERGYSTARTscore, NumberofFloors**
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - Global

- ↗YearBuilt $\Longrightarrow$↑SiteEnergyUse, ↗GHGEmissions

- ↗NumberofFloors$\Longrightarrow$↑↑SiteEnergyUse, ↑↑GHGEmissions



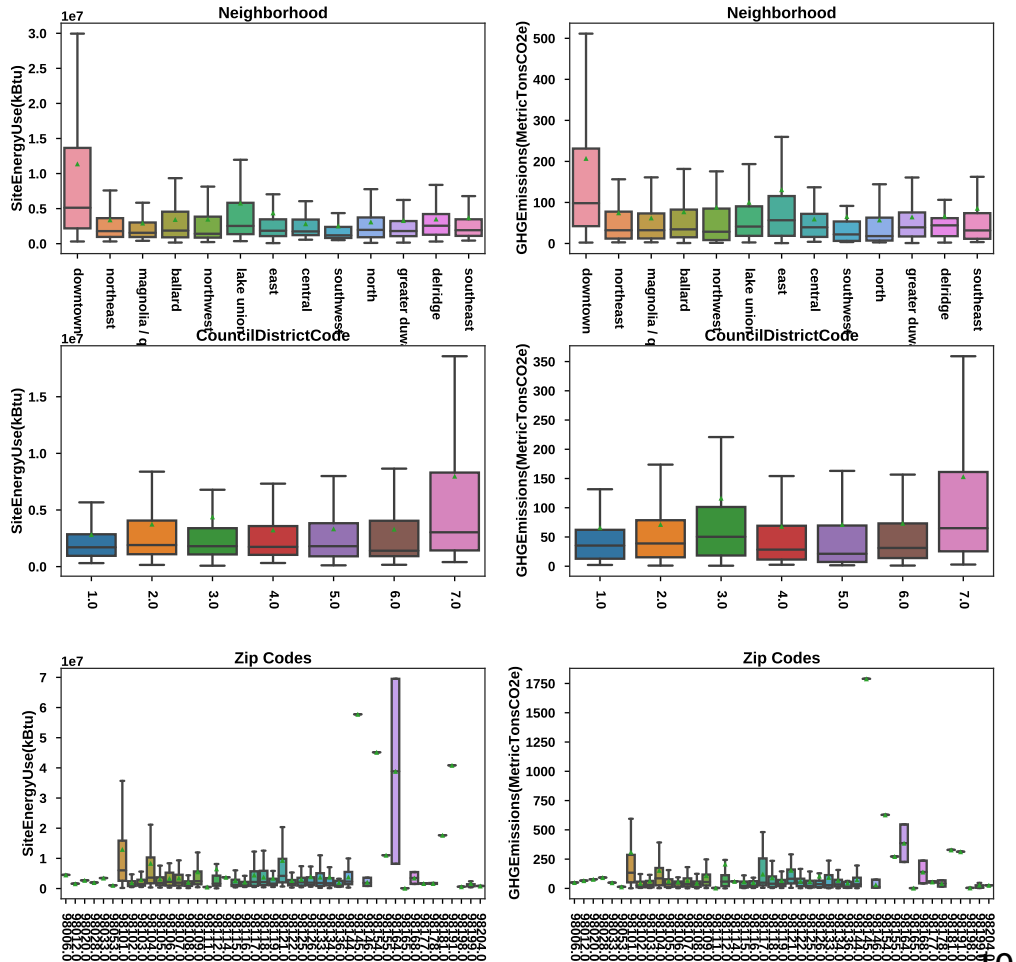
DataFrame: dfDATA2

Analyse graphique

- Neighborhood(downtown) $\Longrightarrow$↑SiteEnergyUse, ↑GHGEmissions

- CouncilDistrictCode (7) $\Longrightarrow$↑SiteEnergyUse, ↑GHGEmissions

- ZipCode(98145-98164) $\Longrightarrow$↑SiteEnergyUse, ↑GHGEmissions

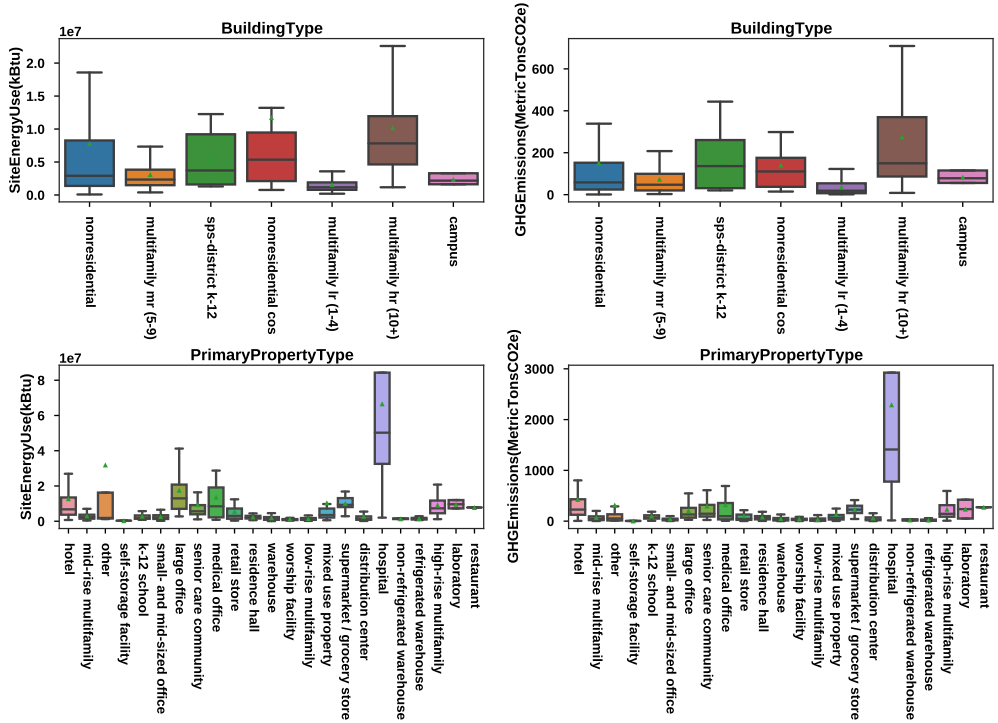


Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - **BuildingType, PrimaryPropertyType**
   - Global

- BuildingType(multifamily,campus) $\Longrightarrow \downarrow$SiteEnergyUse, $\downarrow$GHGEmissions

- PrimaryPropertyType (hospital) $\Longrightarrow \uparrow$SiteEnergyUse, $\uparrow$GHGEmissions

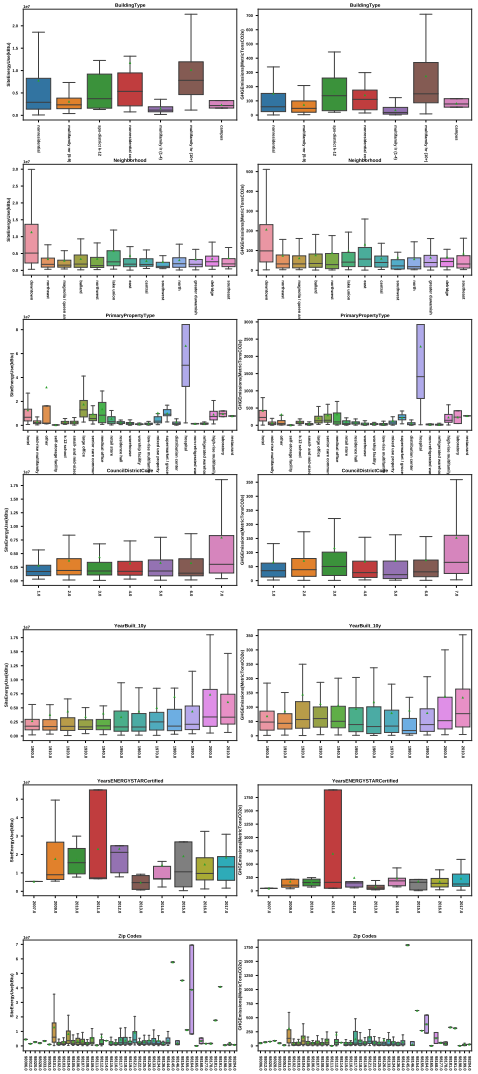


DataFrame: dfDATA2

Analyse graphique

1. EnergySiteUse - Map
2. Distribution Univariée
   - Cible
   - Energy
   - Other
3. Distribution Bivariée
4. Variables continue
5. Variables categorique
   - YearBuilt, ENERGYSTARTscore, NumberofFloors
   - Neighborhood, Council, ZipCode
   - BuildingType, PrimaryPropertyType
   - **Global**

Comportaiment similaire entre SiteEnergyUse et GHGEmissions



DataFrame: dfDATA2

Test de correlation

Seattle

Appliquer l'ANOVA -> Vérifier la normalité de la distribution et l'homoscedasticité

1. **Test de normalité**
   - **Shapiro–Wilk's test**
   - **Visualisation (qqplot)**
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - pval
   - chi2
   - Correlations
5. ANOVA test*

L'hypothèse nulle : La population est normalement distribuée.
Si pval $< \alpha$ choisi, alors l'hypothèse nulle est rejetée et il est prouvé que les données testées ne sont pas normalement distribuées.

	W	pval	normal
CouncilDistrictCode	0.86	0.00	False
ENERGYSTARScore	0.91	0.00	False
Electricity(kBtu)	0.30	0.00	False
GHGEmissions(MetricTonsCO2e)	0.34	0.00	False
Latitude	0.97	0.00	False
Longitude	0.98	0.00	False
NaturalGas(kBtu)	0.45	0.00	False
NumberofFloors	0.52	0.00	False
OtherFuelUse(kBtu)	0.01	0.00	False
PropertyGFABuilding(s)	0.50	0.00	False
PropertyGFAParking	0.30	0.00	False
SPD Beats	0.94	0.00	False
SiteEnergyUse(kBtu)	0.36	0.00	False
SteamUse(kBtu)	0.07	0.00	False
YearBuilt	0.92	0.00	False
YearsENERGYSTARCertified	0.77	0.00	False
Zip Codes	0.74	0.00	False
YearBuilt_10y	0.92	0.00	False
YearsENERGYSTARCertified_10y	0.26	0.00	False

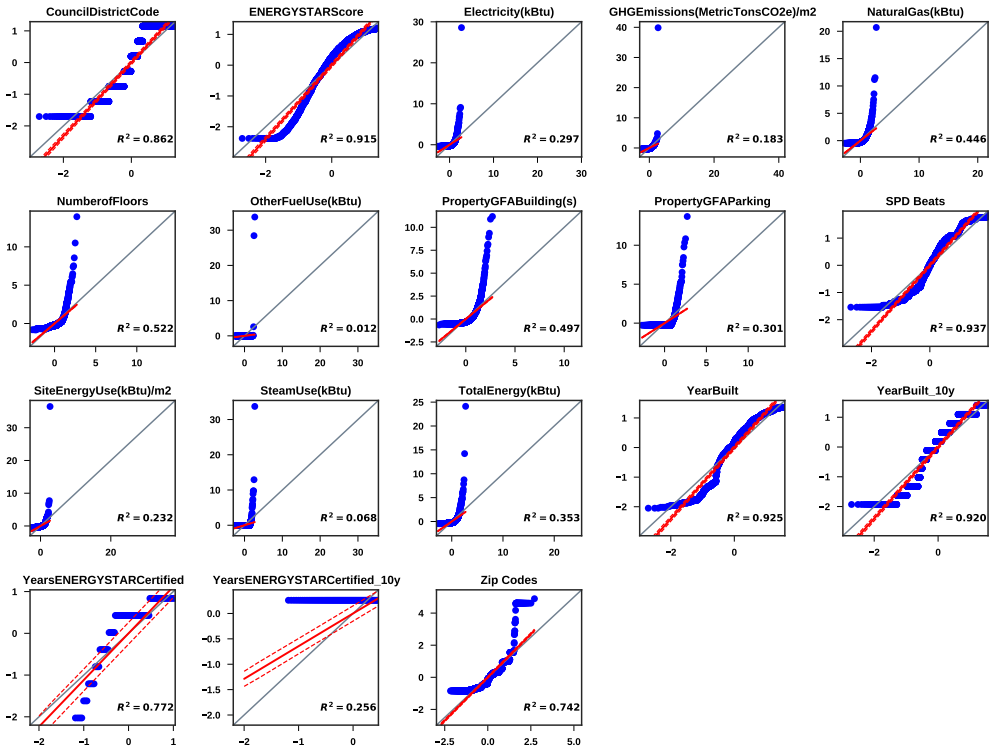
DataFrame: dfDATA2

Dans notre cas, la probabilité que les variables soint normalement dis-

Test de correlation

1. Test de normalité
   - Shapiro–Wilk's test
   - **Visualisation (qqplot)**
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - pval
   - chi2
   - Correlations
5. ANOVA test*

Appliquer l'ANOVA -> Vérifier la normalité de la distribution et l'homoscedasticité

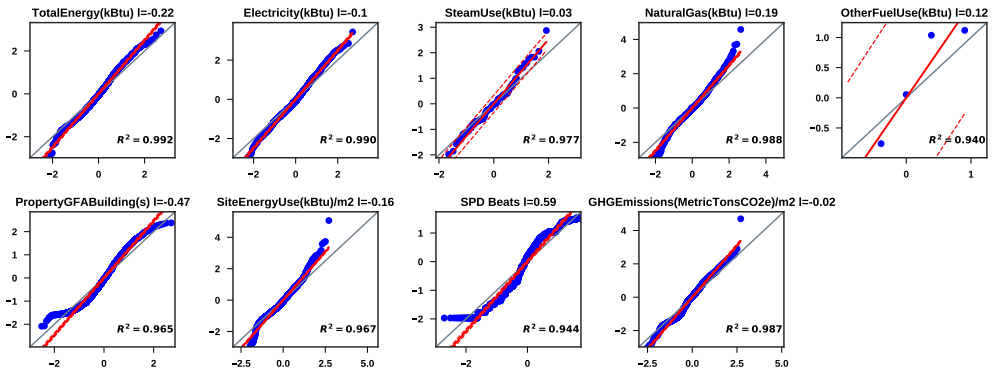


DataFrame: dfDATA2

Test de correlation

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. **Transformation de Box-Cox**
3. Kruskal-Wallis test
4. Chi2
   - pval
   - chi2
   - Correlations
5. ANOVA test*

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$



DataFrame: dfDATA2

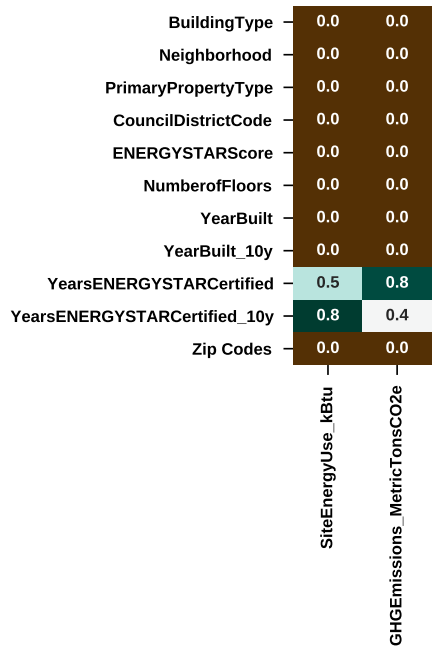
Pas de distribution normal (après Box-Cox)-> Le test ANOVA n'est pas applicable

Test de correlation

Seattle

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. **Kruskal-Wallis test**
4. Chi2
   - pval
   - chi2
   - Correlations
5. ANOVA test*

L'hypothèse nulle : La médiane de population de tous les groupes est égale (version non paramétrique de l'ANOVA).
Dans notre cas, la probabilité que les variables soint independants est de 0.00% -> (la probabilité que les variables soint dependants est de 100%) -> Correlation.



DataFrame: dfDATA2'

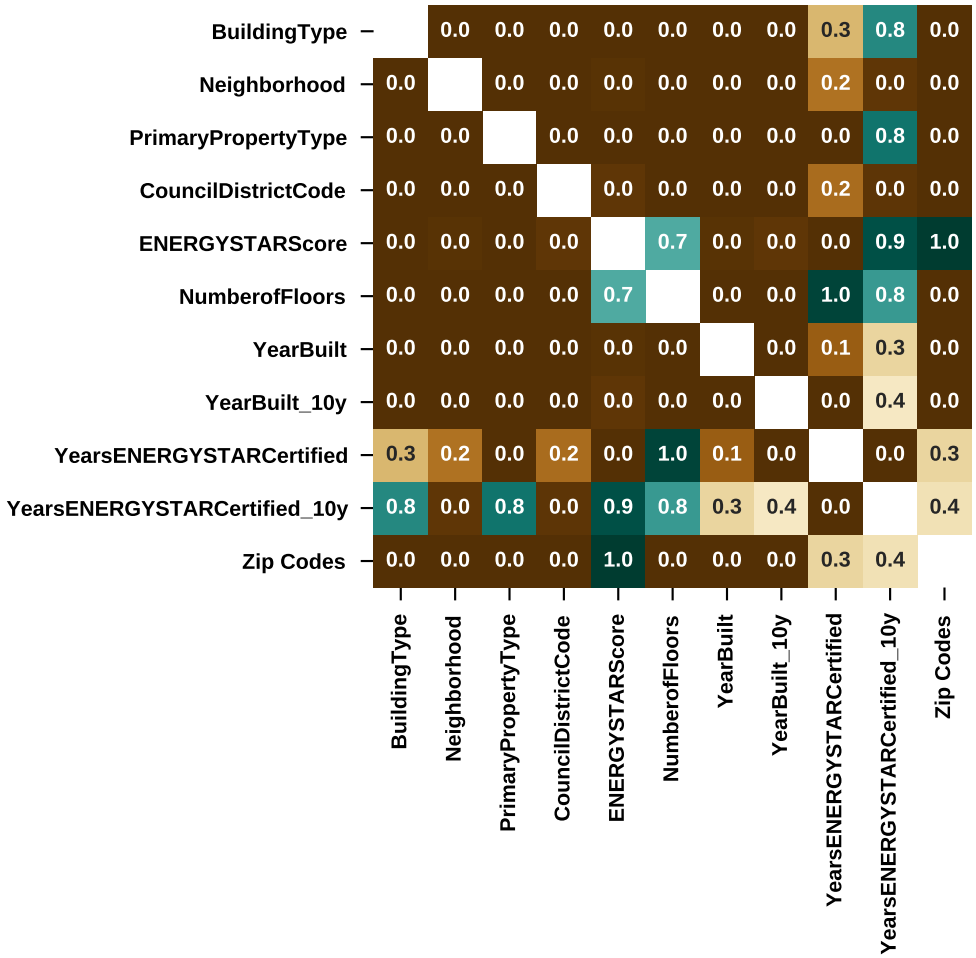
Correlation entre les variables objectifs et les variables explicatives. La seule exception est entre YearsENERGYSTARTCertified et YearsENERGYSTARTCertified_10y

Test de correlation

Seattle

L'hypothèse nulle : Les 2 variables sont indépendant.
ex: La probabilite que la variable 'YearBuilt' et 'YearBuilt_10y' soient independant est du 0.0% -> (la probabilité que les variables soint dependants est de 100%) -> Correlation.

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. **Chi2**
   - **pval**
   - **chi2**
   - **Correlations**
5. ANOVA test*



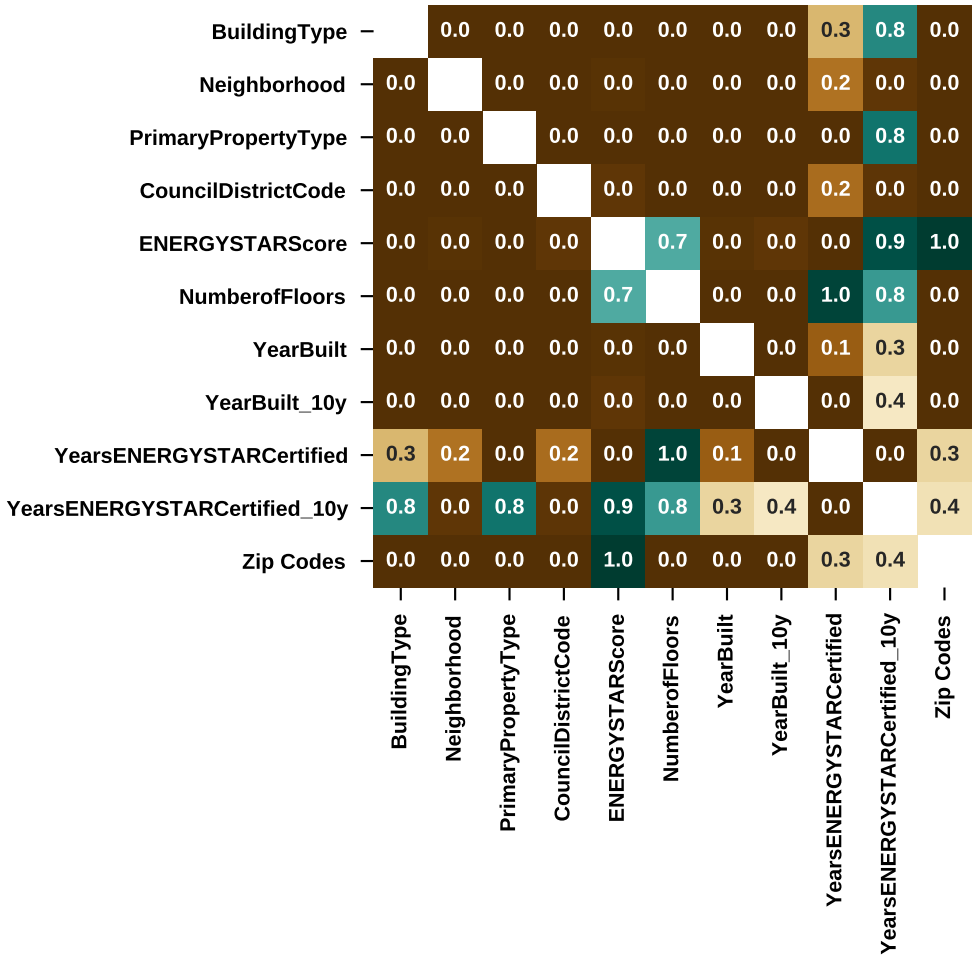
pval

Test de correlation

Seattle

L'hypothèse nulle : Les 2 variables sont indépendant.
ex: La probabilite que la variable 'YearBuilt' et 'YearBuilt_10y' soient independant est du 0.0% -> (la probabilité que les variables soint dependants est de 100%) -> Correlation.

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - **pval**
   - chi2
   - Correlations
5. ANOVA test*



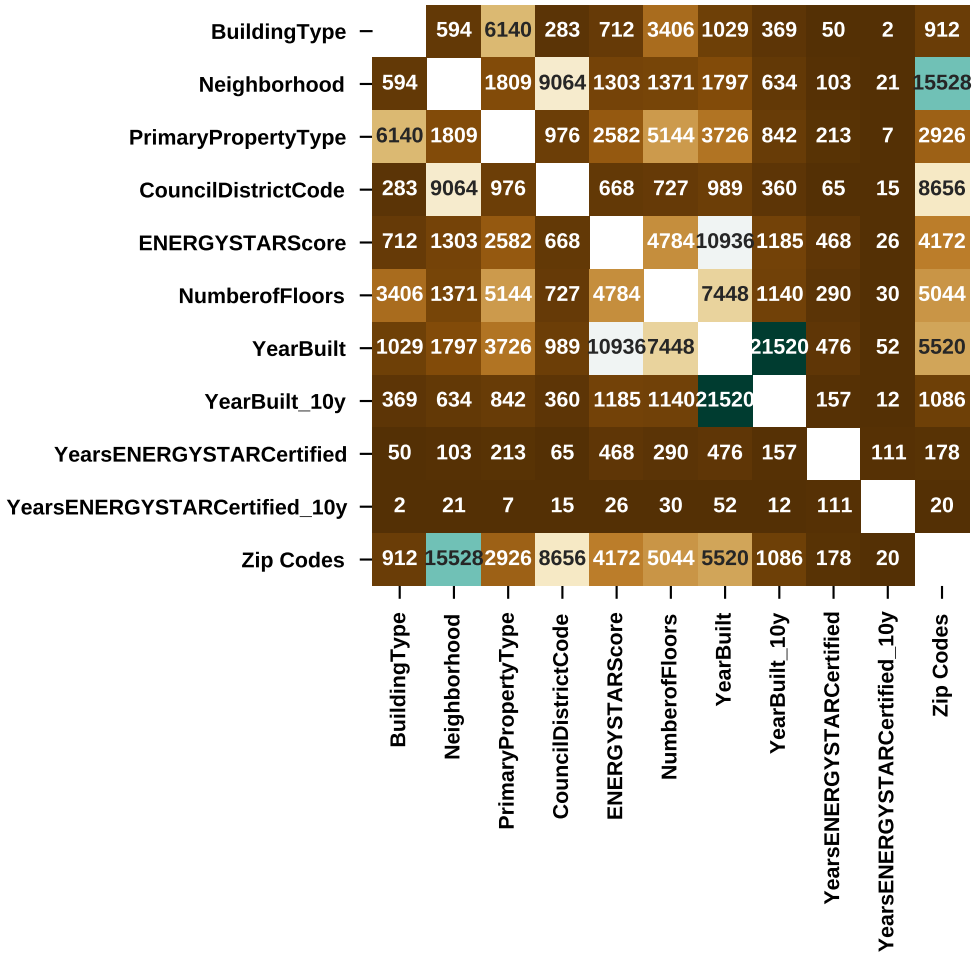
pval

Test de correlation

Seattle

L'hypothèse nulle : Les 2 variables sont indépendant.
ex: La probabilite que la variable 'YearBuilt' et 'YearBuilt_10y' soient independant est du 0.0% -> (la probabilité que les variables soint dependants est de 100%) -> Correlation.

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - pval
   - **chi2**
   - Correlations
5. ANOVA test*



Test de correlation

Seattle

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - pval
   - chi2
   - **Correlations**
5. ANOVA test*

L'hypothèse nulle : Les 2 variables sont indépendant.
ex: La probabilite que la variable 'YearBuilt' et 'YearBuilt_10y' soient independant est du 0.0% -> (la probabilité que les variables soint dependants est de 100%) -> Correlation.

1. YearBuilt - YearBuilt_10y
2. YearBuilt - ENERGYSTARTscore
3. YearBuilt - NumberofFloors

4. Neighborhood - CouncilDistrictCode
5. Neighborhood - ZipCode
6. ZipCode - CouncilDistrictCode

7. PrimaryPropertyType - BuildingType

Test de correlation

Seattle

1. Test de normalité
   - Shapiro–Wilk's test
   - Visualisation (qqplot)
2. Transformation de Box-Cox
3. Kruskal-Wallis test
4. Chi2
   - pval
   - chi2
   - Correlations
5. **ANOVA test***

L'hypothèse nulle 1 : La population est normalement distribuée.
L'hypothèse nulle 2 : Homogénéité des variances.
L'hypothèse nulle 3 : Chaque échantillon analysé est indépendant des autres échantillons.
ex: La variable 'SiteEnergyUse' **aurait eu** une probabilite du 73% d'être indépendant (27% dépendant) de la variable 'Neighborhood' et elle a une contribution du 0%.

	BuildingType	Neighborhood	PrimaryPropertyType
SiteEnergyUse_kBtu	0.04	0.73	0.99
GHGEmissions_MetricTonsCO2e	1.00	0.23	1.00

pval

	BuildingType	Neighborhood	PrimaryPropertyType
SiteEnergyUse_kBtu	0.00	0.00	0.00
GHGEmissions_MetricTonsCO2e	-0.00	0.01	-0.02

np2

Features selection

1. **Energetic features**
2. Geographical features
3. Building classification
4. Other features
5. Export 6 features selections

~~'Electricity(kBtu)', 'NaturalGas(kBtu)', 'OtherFuelUse(kBtu)', 'PropertyGFABuilding(s)', 'SPD Beats', 'SteamUse(kBtu)', 'TotalEnergy(kBtu)'~~

They are more difficult to find it

Features selection

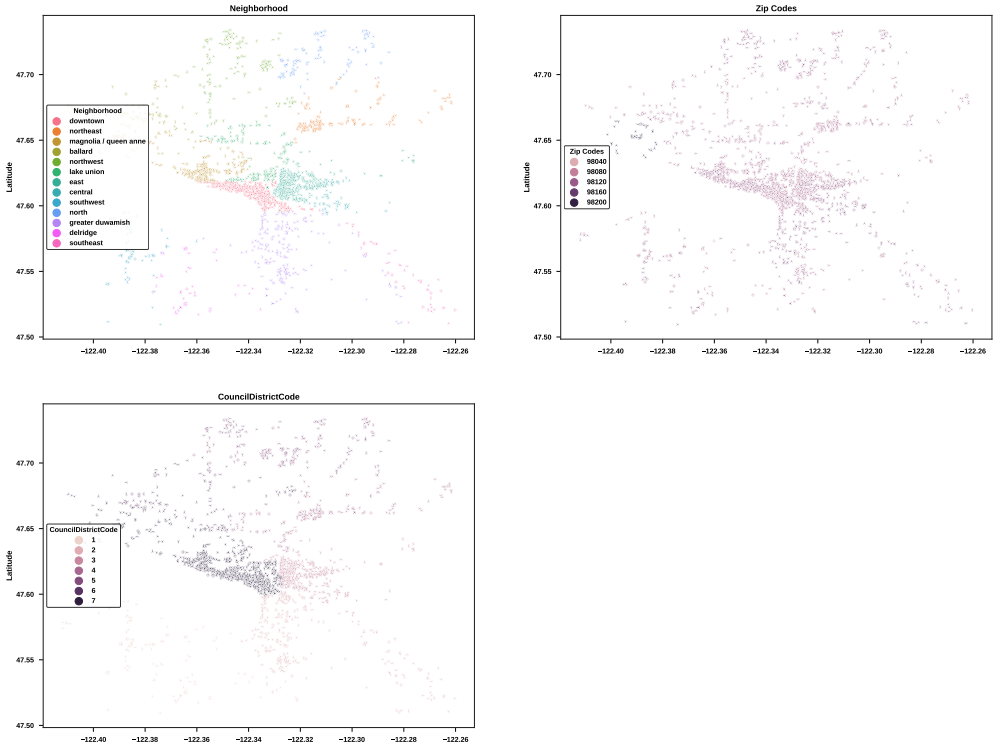
Seattle

1. Energetic features
2. **Geographical features**
3. Building classification
4. Other features
5. Export 6 features selections

'Neighborhood': 19 categories

~~'CouncilDistrictCode'~~: 7 categories

~~'ZipCode'~~: 55 categories



Same information

Neighborhood is more representative

Features selection

Seattle

1. Energetic features
2. Geographical features
3. **Building classification**
4. Other features
5. Export 6 features selections

'PrimaryPropertyType ': 26 categories
~~'BuildingType'~~: 8 categories

Features selection

1. Energetic features
2. Geographical features
3. Building classification
4. **Other features**
5. Export 6 features selections

'YearENERGYSTARCertified', 'PropertyGFAParking'

Not representative

Features selection

Seattle

1. Energetic features
2. Geographical features
3. Building classification
4. Other features
5. **Export 6 features selections**

'PropertyGFABuilding(s)', 'NumberofFloors', 'PrimaryPropertyType', 'Neighborhood', 'YearBuilt_10y', 'ENERGYSTARScore' (CO2)

Easy to implement

Easy to maintenance (low update)

Index

Preparation data

1. **Training and Test Set split**
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - Concat
   - Scaling

Xtrain, ytrain_Energy $\Rightarrow$ 1564 rows (80%)

Xtest, ytest_Energy $\Rightarrow$ 392 rows (20%)

Preparation data

Seattle

1. Training and Test Set split
2. **Encoding**
   - **TargetEncoder**
   - **OrdinalEncoder**
3. Scaling features
   - Concat
   - Scaling

'PrimaryPropertyType', 'Neighborhood'

OSEBuildingID	PrimaryPropertyType	Neighborhood
329	1.710437e+07	1.031354e+07
20011	1.497206e+06	3.735103e+06
19889	1.497206e+06	1.031354e+07
498	6.956444e+07	5.819853e+06
21299	1.497206e+06	3.103828e+06
...	...	...
19685	1.497206e+06	1.031354e+07
763	2.723755e+06	1.031354e+07
20094	9.523515e+06	1.031354e+07
26838	1.254663e+07	1.031354e+07
26056	1.497206e+06	4.686974e+06

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

OSEBuildingID	PrimaryPropertyType	Neighborhood
329	16.372506	15.483854
20011	14.030716	14.746871
19889	14.030716	15.483854
498	17.004108	14.957166
21299	14.030716	14.462867
...	...	...
19685	14.030716	15.483854
763	14.579895	15.483854
20094	15.812853	15.483854
26838	15.858895	15.483854
26056	14.030716	14.595277

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

Seattle

1. Training and Test Set split
2. Encoding
   - **TargetEncoder**
   - OrdinalEncoder
3. Scaling features
   - Concat
   - Scaling

'PrimaryPropertyType', 'Neighborhood'

OSEBuildingID	PrimaryPropertyType	Neighborhood
329	1.710437e+07	1.031354e+07
20011	1.497206e+06	3.735103e+06
19889	1.497206e+06	1.031354e+07
498	6.956444e+07	5.819853e+06
21299	1.497206e+06	3.103828e+06
...	...	...
19685	1.497206e+06	1.031354e+07
763	2.723755e+06	1.031354e+07
20094	9.523515e+06	1.031354e+07
26838	1.254663e+07	1.031354e+07
26056	1.497206e+06	4.686974e+06

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

OSEBuildingID	PrimaryPropertyType	Neighborhood
329	16.372506	15.483854
20011	14.030716	14.746871
19889	14.030716	15.483854
498	17.004108	14.957166
21299	14.030716	14.462867
...	...	...
19685	14.030716	15.483854
763	14.579895	15.483854
20094	15.812853	15.483854
26838	15.858895	15.483854
26056	14.030716	14.595277

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - **OrdinalEncoder**
3. Scaling features
   - Concat
   - Scaling

'YearBuilt_10y'

OSEBuildingID	PrimaryPropertyType	Neighborhood	YearBuilt_10y
329	1.710437e+07	1.031354e+07	10.0
20011	1.497206e+06	3.735103e+06	9.0
19889	1.497206e+06	1.031354e+07	1.0
498	6.956444e+07	5.819853e+06	10.0
21299	1.497206e+06	3.103828e+06	0.0
...	...	...	...
19685	1.497206e+06	1.031354e+07	2.0
763	2.723755e+06	1.031354e+07	0.0
20094	9.523515e+06	1.031354e+07	7.0
26838	1.254663e+07	1.031354e+07	1.0
26056	1.497206e+06	4.686974e+06	1.0

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

OSEBuildingID	PrimaryPropertyType	Neighborhood	YearBuilt_10y
329	16.372506	15.483854	10.0
20011	14.030716	14.746871	9.0
19889	14.030716	15.483854	1.0
498	17.004108	14.957166	10.0
21299	14.030716	14.462867	0.0
...	...	...	...
19685	14.030716	15.483854	2.0
763	14.579895	15.483854	0.0
20094	15.812853	15.483854	7.0
26838	15.858895	15.483854	1.0
26056	14.030716	14.595277	1.0

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. **Scaling features**
   - **Concat**
   - **Scaling**

OSEBuildingID	PropertyGFABuilding(s)	NumberofFloors	PrimaryPropertyType	Neighborhood	YearBuilt_10y
329	542671	24	16.372506	15.483854	10.0
20011	23533	4	14.030716	14.746871	9.0
19889	21284	4	14.030716	15.483854	1.0
498	201075	8	17.004108	14.957166	10.0
21299	25800	4	14.030716	14.462867	0.0
...	...	...	...	...	...
19685	33400	4	14.030716	15.483854	2.0
763	51218	7	14.579895	15.483854	0.0
20094	140241	24	15.812853	15.483854	7.0
26838	40552	6	15.858895	15.483854	1.0
26056	30040	4	14.030716	14.595277	1.0

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - **Concat**
   - Scaling

OSEBuildingID	PropertyGFABuilding(s)	NumberofFloors	PrimaryPropertyType	Neighborhood	YearBuilt_10y
329	542671	24	16.372506	15.483854	10.0
20011	23533	4	14.030716	14.746871	9.0
19889	21284	4	14.030716	15.483854	1.0
498	201075	8	17.004108	14.957166	10.0
21299	25800	4	14.030716	14.462867	0.0
...	...	...	...	...	...
19685	33400	4	14.030716	15.483854	2.0
763	51218	7	14.579895	15.483854	0.0
20094	140241	24	15.812853	15.483854	7.0
26838	40552	6	15.858895	15.483854	1.0
26056	30040	4	14.030716	14.595277	1.0

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Preparation data

Seattle

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - Concat
   - **Scaling**

OSEBuildingID	PropertyGFABuilding(s)	NumberofFloors	PrimaryPropertyType	Neighborhood	YearBuilt_10y
329	3.313513	2.685662	-0.889351	-1.862969	1.087098
20011	-0.518523	-0.368157	-0.889351	-1.862969	0.784783
19889	-0.535124	-0.368157	-0.889351	-1.862969	-1.633740
498	0.792010	0.242607	-0.889350	-1.862969	1.087098
21299	-0.501789	-0.368157	-0.889351	-1.862969	-1.936055
...	...	...	...	...	...
19685	-0.445689	-0.368157	-0.889351	-1.862969	-1.331425
763	-0.314165	0.089916	-0.889351	-1.862969	-1.936055
20094	0.342962	2.685662	-0.889351	-1.862969	0.180152
26838	-0.392896	-0.062775	-0.889351	-1.862969	-1.633740
26056	-0.470491	-0.368157	-0.889351	-1.862969	-1.633740

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Promising models

Seattle

1. **RandomizedSearchCV**
   - **Parameters**
   - **Training**
2. Results
   - r2, MSE
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	
alpha	0,01	150	rnd

Ridge, Lasso

Parameter	min	max	
alpha	0,01	150	rnd
l1_ratio	0,25	0,75	3 val

ElasticNet

Parameter	min	max	
max_features	2	5	rnd
max_leaf_nodes	2	80	rnd
n_estimators	1	250	rnd

RandomForest

Parameter	min	max	
alpha	0,01	50	rnd
Colsample bytree	0	1	rnd
max_depth	1	6	rnd
min_child weight	0	1,5	rnd
n_estimators	1	250	rnd

XGBoost

Promising models

Seattle

1. RandomizedSearchCV
   - **Parameters**
   - Training
2. Results
   - r2, MSE
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	
alpha	0,01	150	rnd

Ridge, Lasso

Parameter	min	max	
alpha	0,01	150	rnd
l1_ratio	0,25	0,75	3 val

ElasticNet

Parameter	min	max	
max_features	2	5	rnd
max_leaf_nodes	2	80	rnd
n_estimators	1	250	rnd

RandomForest

Parameter	min	max	
alpha	0,01	50	rnd
Colsample bytree	0	1	rnd
max_depth	1	6	rnd
min_child weight	0	1,5	rnd
n_estimators	1	250	rnd

XGBoost

Promising models

1. RandomizedSearchCV
   - Parameters
   - **Training**
2. Results
   - r2, MSE
3. BestModel
   - Comparison
   - Features contribution

100 iterations number

4 number of cross-validation splits

Multi-metric evaluation : r2 and mean_squared_error

Export results (.pkl)

Promising models

Seattle

1. RandomizedSearchCV
   - Parameters
   - Training
2. **Results**
   - **r2, MSE**
3. BestModel
   - Comparison
   - Features contribution

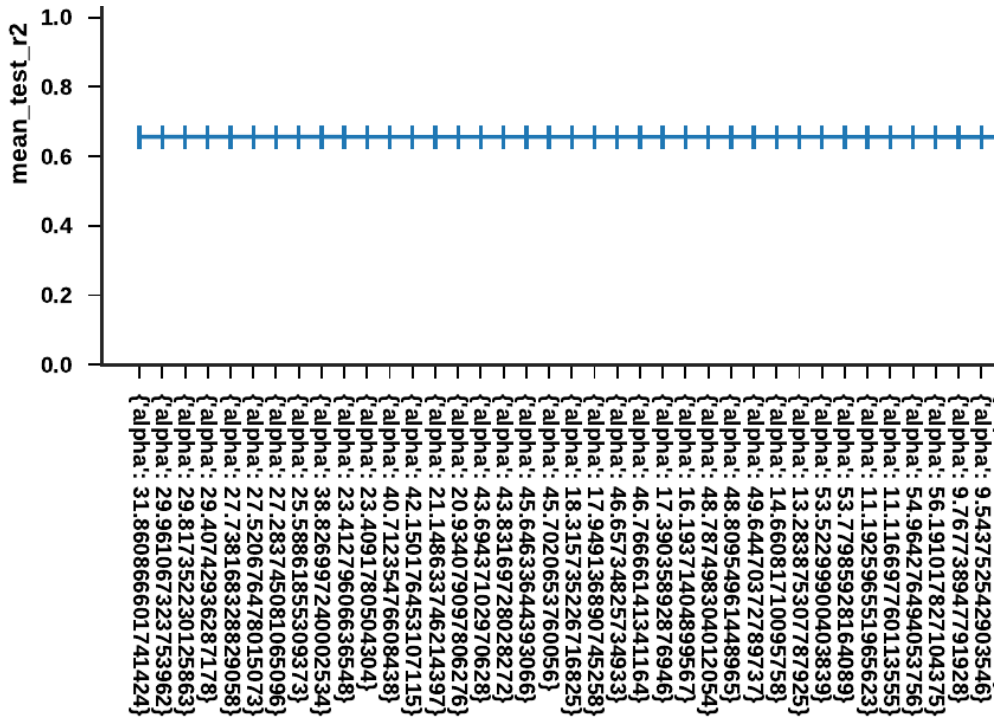
MODELS		params	mean_test_r2	mean_test_neg_mean_squared_error
Ridge()	13	{'alpha': 31.860866601741424}	0.655882	-0.392948
	26	{'alpha': 29.961067323753962}	0.655879	-0.392961
	71	{'alpha': 29.817352230125863}	0.655878	-0.392962
Lasso()	0	{'alpha': 56.191017827104375}	-0.006143	-1.156154
	72	{'alpha': 0.8383175685403599}	-0.006143	-1.156154
	71	{'alpha': 29.817352230125863}	-0.006143	-1.156154
ElasticNet()	64	{'alpha': 1.0528195796786055, 'l1_ratio': 0.6330604769331742}	0.033699	-1.112619
	36	{'alpha': 0.8383175685403599, 'l1_ratio': 0.8615960713411256}	0.010637	-1.138518
	0	{'alpha': 56.191017827104375, 'l1_ratio': 0.9630357298074371}	-0.006143	-1.156154
RandomForestRegressor(random_state=42, bootstrap=[False])	26	{'max_features': 3, 'max_leaf_nodes': 63, 'n_estimators': 228}	0.790652	-0.238611
	59	{'max_features': 3, 'max_leaf_nodes': 63, 'n_estimators': 225}	0.790589	-0.238685
	82	{'max_features': 3, 'max_leaf_nodes': 68, 'n_estimators': 147}	0.790367	-0.239048
xgboost.XGBRegressor(objective="reg:linear")	91	{'alpha': 2.684264837368445, 'colsample_bytree': 0.9585414968831983, 'max_depth': 3, 'min_child_weight': 0.13693001517920034, 'n_estimators': 199}	0.788576	-0.240868
	76	{'alpha': 4.901708032550074, 'colsample_bytree': 0.49161587511683236, 'max_depth': 4, 'min_child_weight': 0.2598028048650228, 'n_estimators': 222}	0.786144	-0.243532
	33	{'alpha': 1.8543473677266398, 'colsample_bytree': 0.6095643339798968, 'max_depth': 2, 'min_child_weight': 0.6165555199773469, 'n_estimators': 87}	0.785703	-0.243549

Selection : Ridge, RandomForest & XGBoost

Promising models

Seattle

1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - **r2, MSE**
3. BestModel
   - Comparison
   - Features contribution



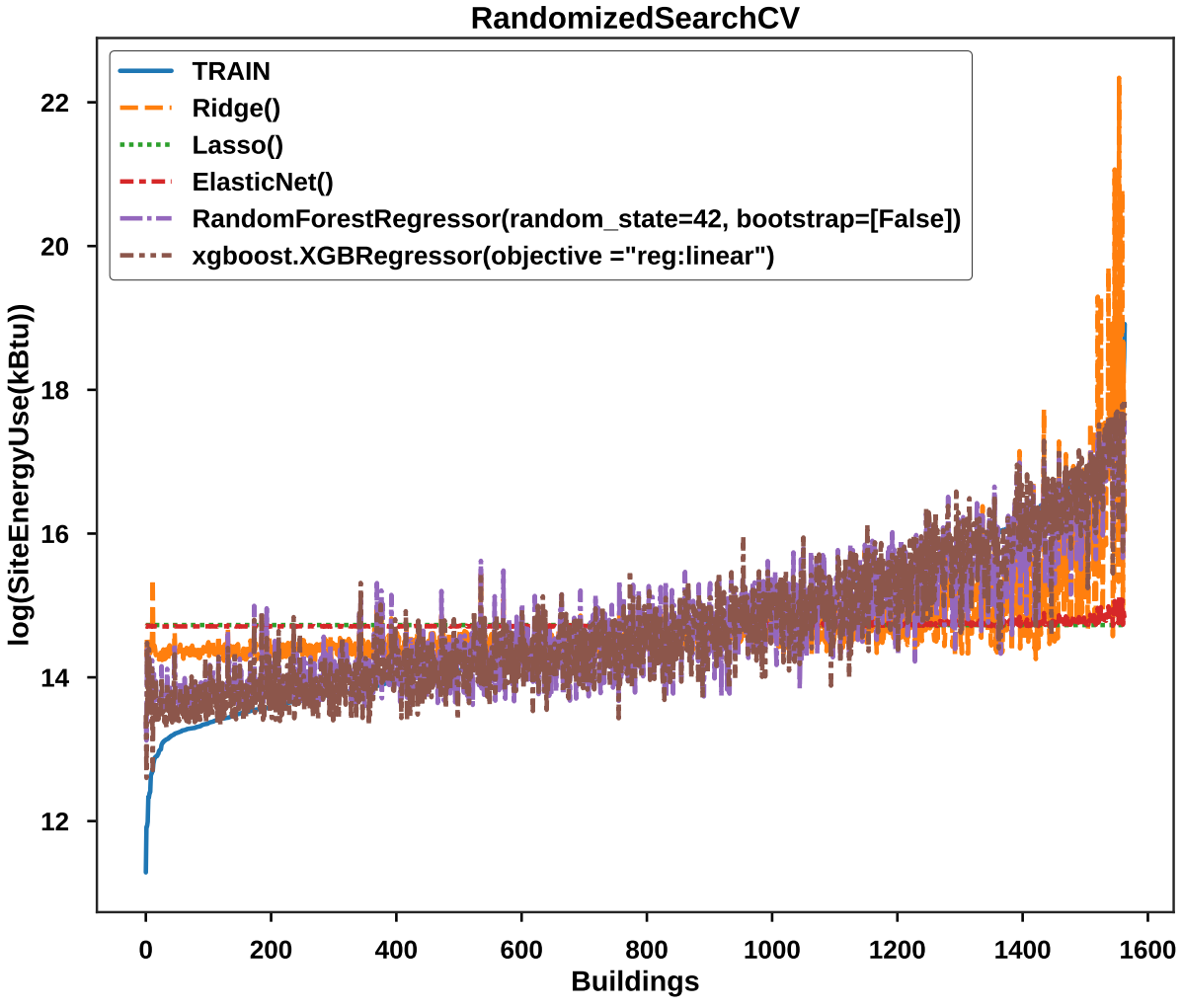
Alpha between: 20 - 50

Promising models

1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - **r2, MSE**
3. BestModel
   - Comparison
   - Features contribution

Promising models

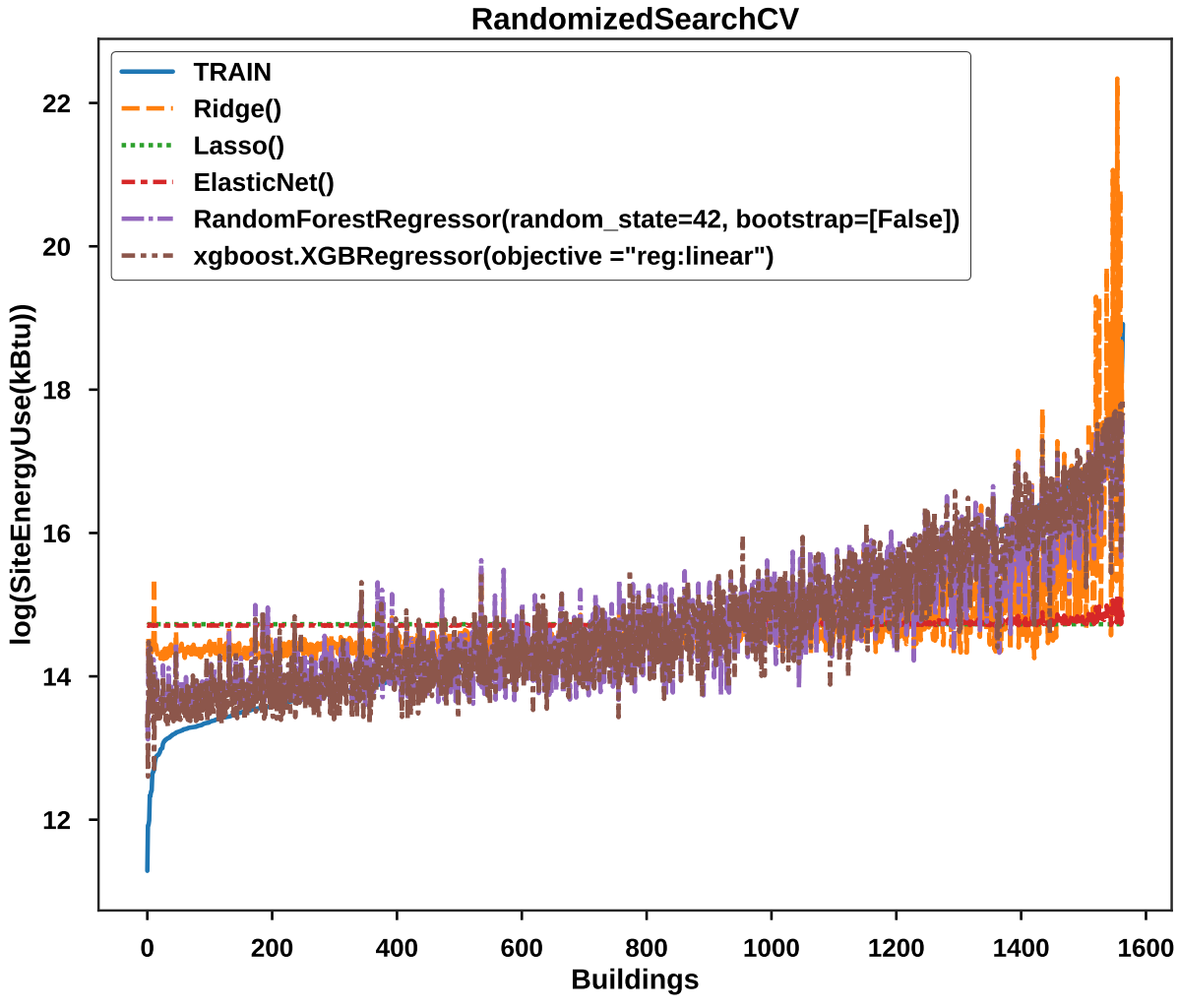
1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - r2, MSE
3. **BestModel**
   - **Comparison**
   - **Features contribution**



Selection : Ridge, RandomForest & XGBoost

Promising models

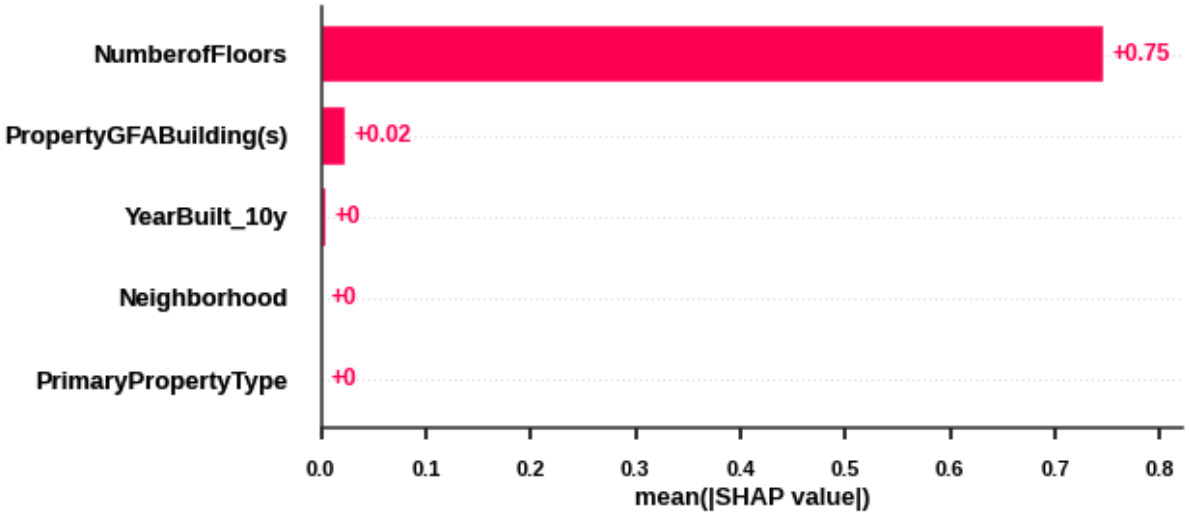
1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - r2, MSE
3. BestModel
   - **Comparison**
   - Features contribution

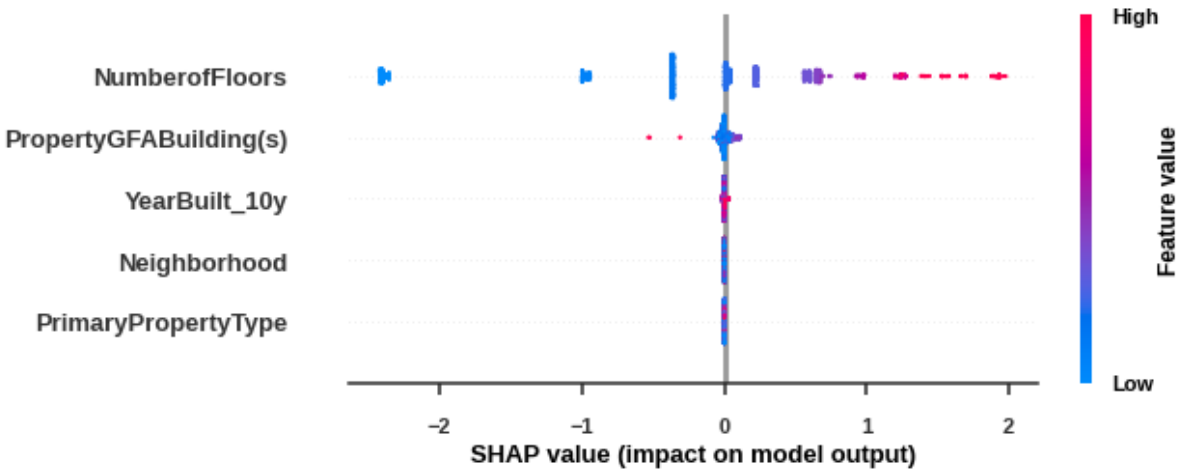


Selection : Ridge, RandomForest & XGBoost

Promising models

1. RandomizedSearchC\
   - Parameters
   - Training
2. Results
   - r2, MSE
3. BestModel
   - Comparison
   - **Features contribution**





The highest building spent much more energy than the lowest building

Fine-tune

Seattle

1. **GridSearchCV**
   - **Parameters**
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	nbr	
alpha	10	70	20	lin

Ridge

Parameter	min	max	nbr	
max_features	3			
max_leaf_nodes	80	150	4	lin
n_estimators	50	500	5	lin

RandomForest

Parameter	min	max	nbr	
alpha	0,01	5	7	log
Colsample bytree	0,75			
max_depth	3			
min_child weight	0,75			
n_estimators	20	150	3	log

XGBoost

Fine-tune

Seattle

1. GridSearchCV
   - **Parameters**
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	nbr	
alpha	10	70	20	lin

Ridge

Parameter	min	max	nbr	
max_features	3			
max_leaf_nodes	80	150	4	lin
n_estimators	50	500	5	lin

RandomForest

Parameter	min	max	nbr	
alpha	0,01	5	7	log
Colsample bytree	0,75			
max_depth	3			
min_child weight	0,75			
n_estimators	20	150	3	log

XGBoost

Fine-tune

1. GridSearchCV
   - Parameters
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

20 iterations number per model

10 number of cross-validation splits

Multi-metric evaluation : r2 and mean_squared_error

Export results (.pkl)

Fine-tune

Seattle

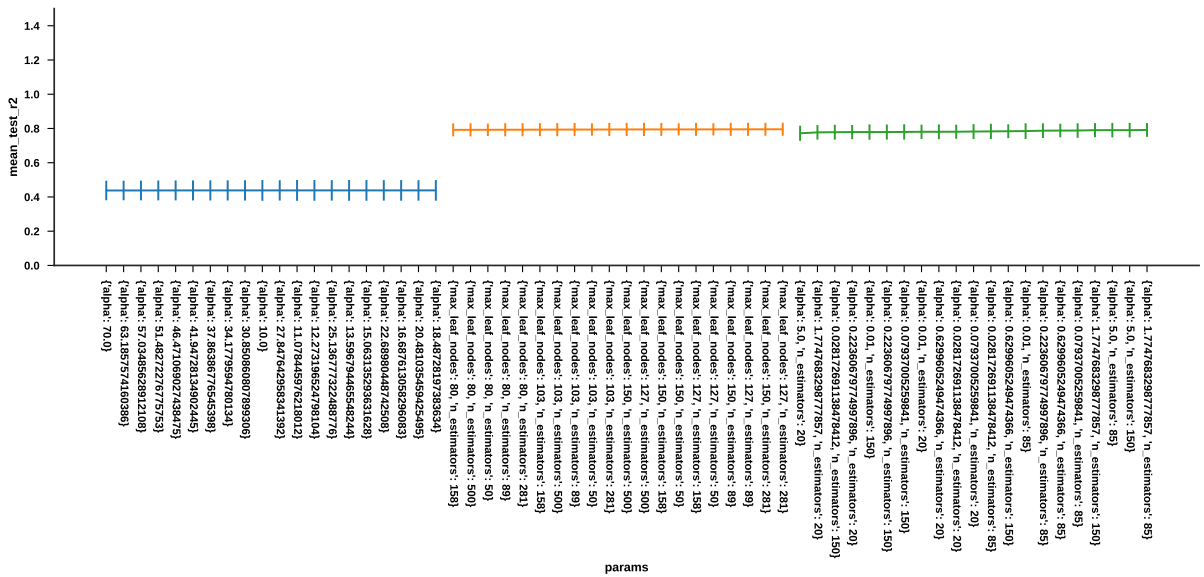
1. GridSearchCV
   - Parameters
   - Training
2. **Results**
   - **r2**
3. BestModel
   - Comparison
   - Features contribution

MODELS				
Ridge()	**19**	{'alpha': 70.0}	0.437702	-0.642923
RandomForestRegressor(random_state=42, bootstrap=[False], max_features=3)	**2**	{'max_leaf_nodes': 80, 'n_estimators': 158}	0.791797	-0.237118
xgboost.XGBRegressor(objective ="reg:linear", colsample_bytree=0.75, max_depth=3, min_child_weight=0.75)	**18**	{'alpha': 5.0, 'n_estimators': 20}	0.771817	-0.259736

Selection : RandomForest

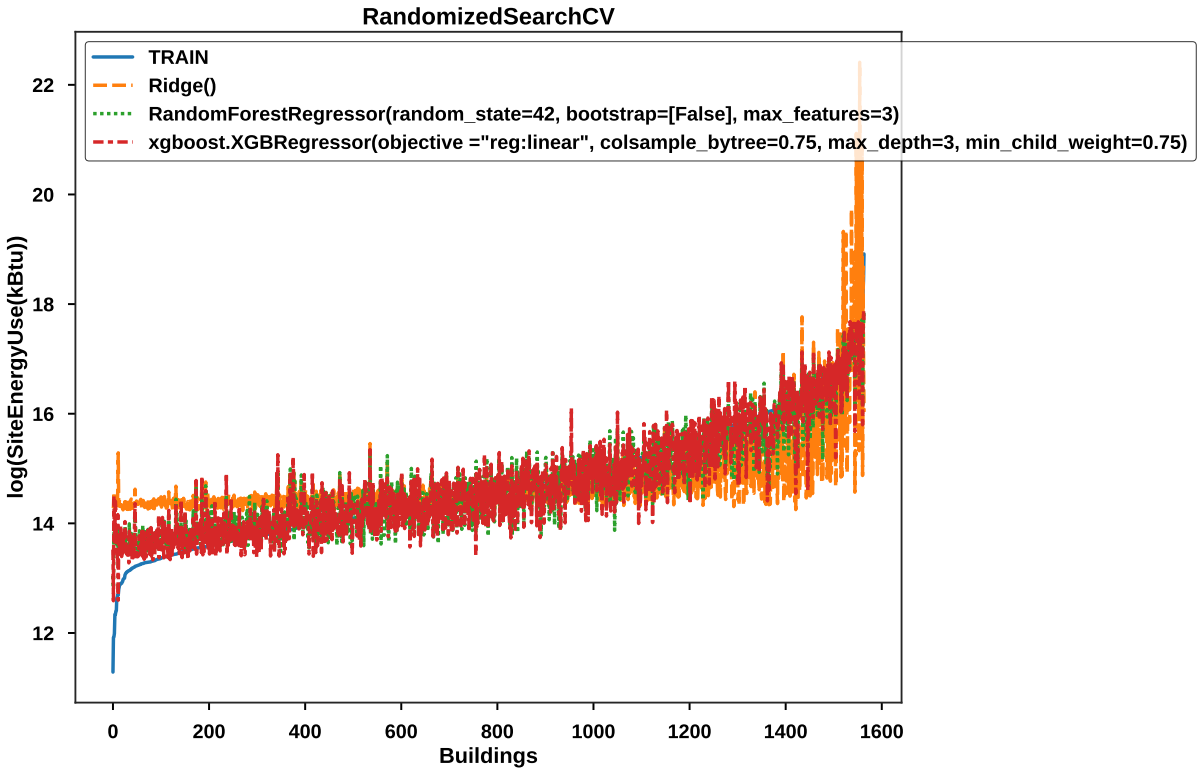
Fine-tune

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - **r2**
3. BestModel
   - Comparison
   - Features contribution



Fine-tune

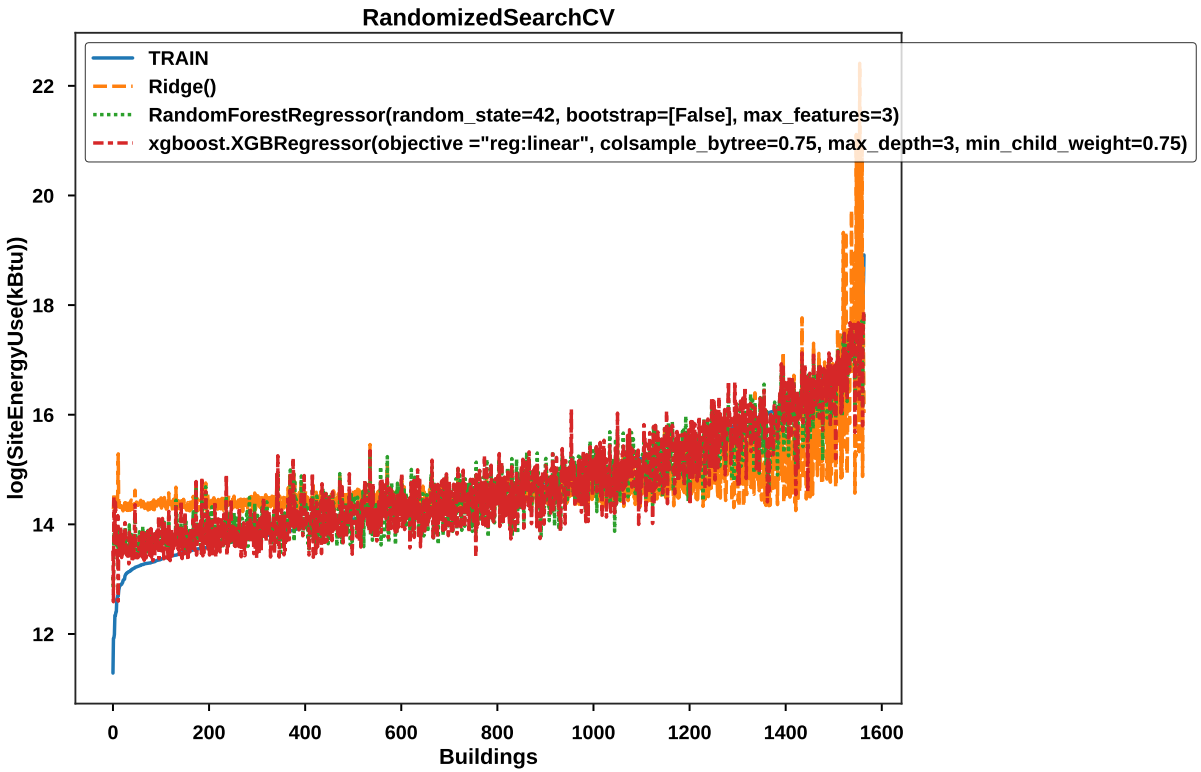
1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. **BestModel**
   - **Comparison**
   - **Features contribution**



Selection : RandomForest

Fine-tune

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - **Comparison**
   - Features contribution

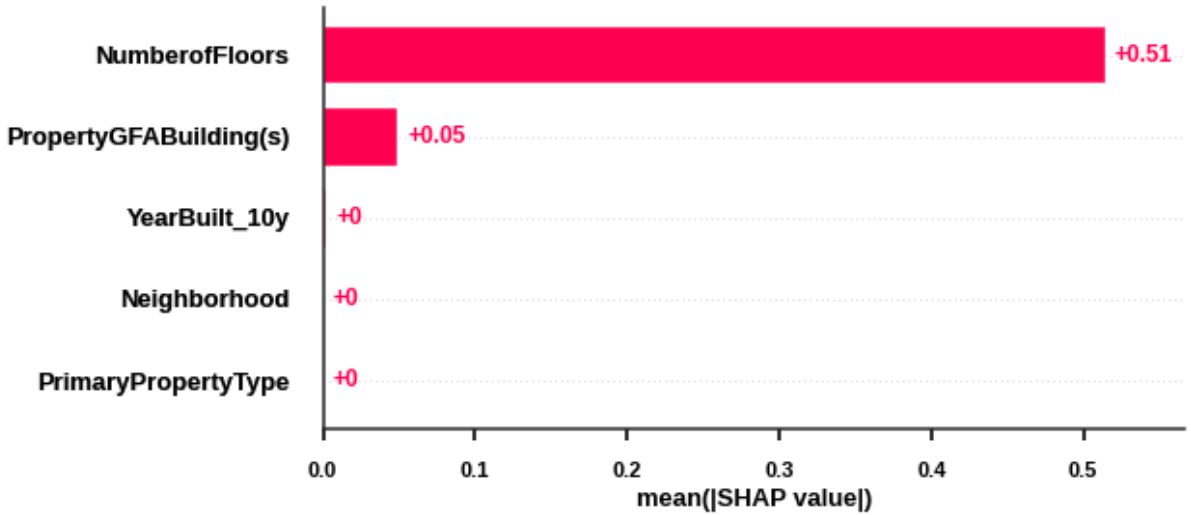


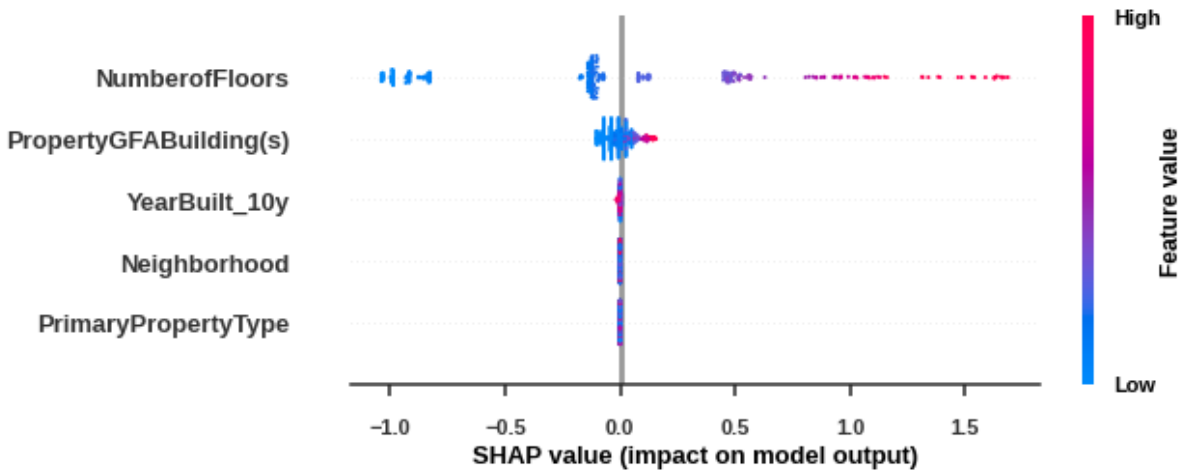
Selection : RandomForest

Fine-tune

Seattle

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - **Features contribution**





The highest building spent much more energy than the lowest building

Validation

1. **BestModel**
   - **Results**
   - **Selection**
2. Features contribution
3. Export

	params	mean_test_r2	mean_test_neg_mean_squared_error
RandomForestRegressor(random_state=42, bootstrap=[False], max_features=3)	{'max_leaf_nodes': 80, 'n_estimators': 158}	0.945845	0.070093

Validation

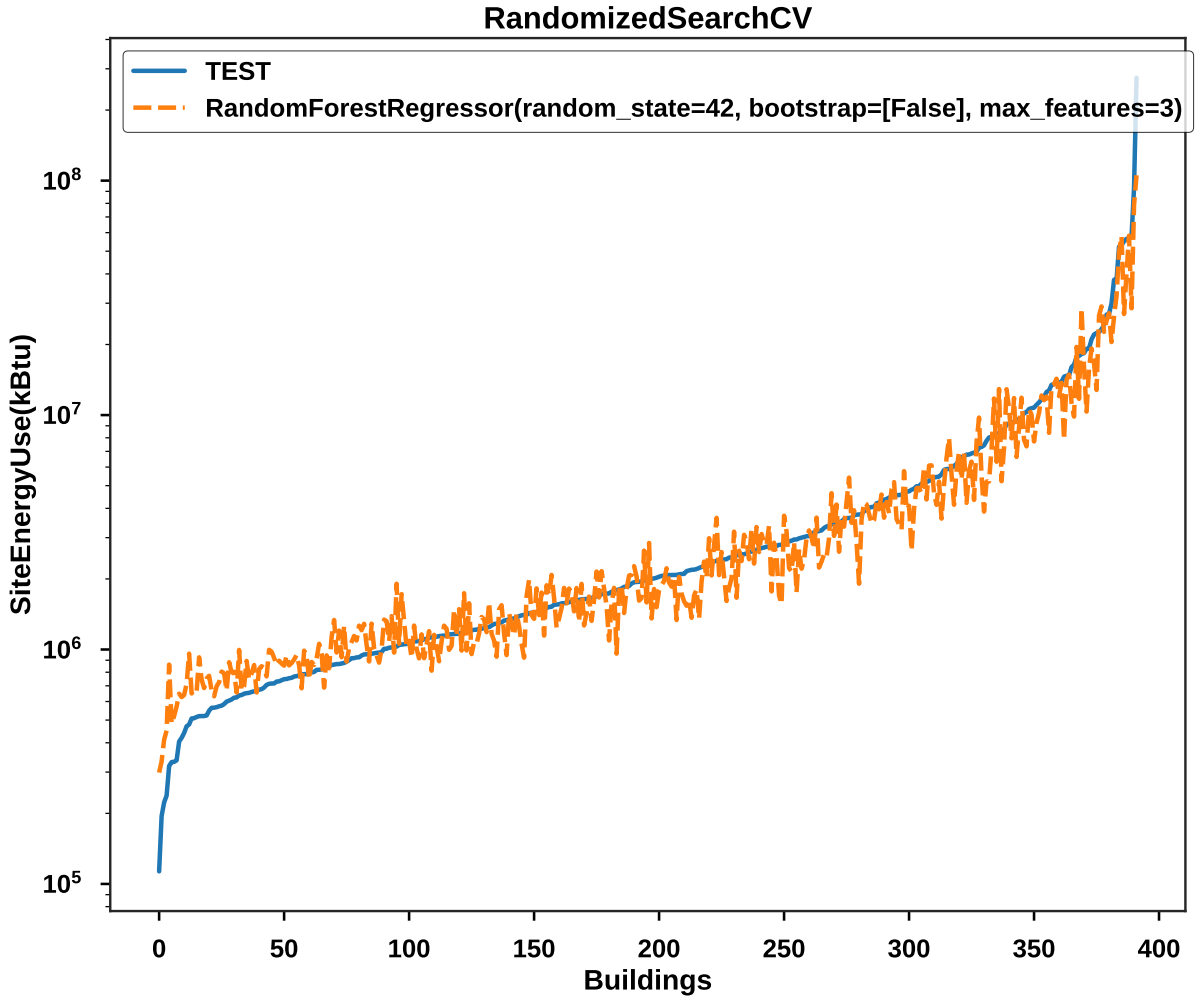
Seattle

1. BestModel
   - **Results**
   - Selection
2. Features contribution
3. Export

	params	mean_test_r2	mean_test_neg_mean_squared_error
RandomForestRegressor(random_state=42, bootstrap=[False], max_features=3)	{'max_leaf_nodes': 80, 'n_estimators': 158}	0.945845	0.070093

Validation

1. BestModel
   - Results
   - **Selection**
2. Features contribution
3. Export

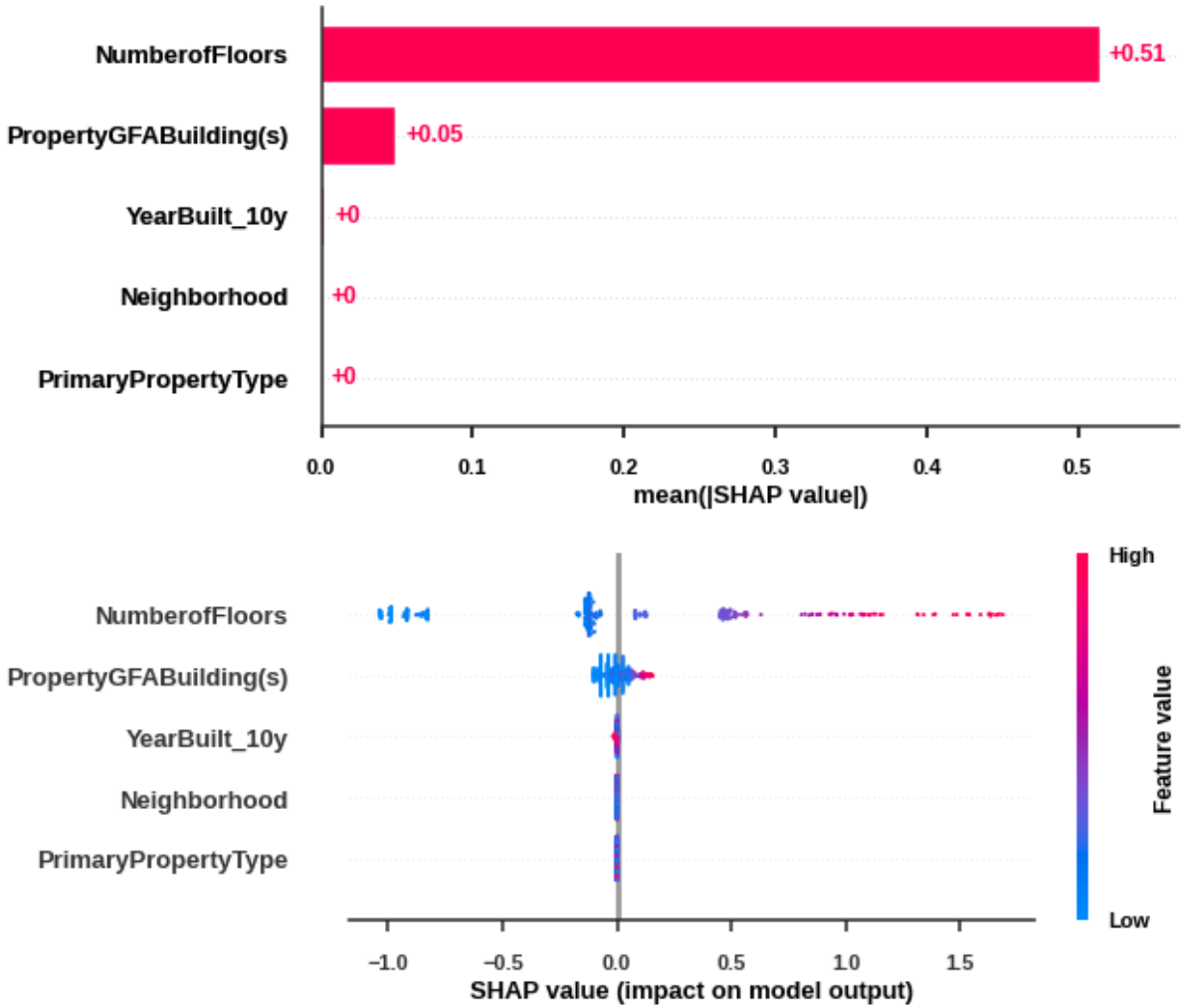


The RandomForest model is validated

Validation

Seattle

1. BestModel
   - Results
   - Selection
2. **Features contribution**
3. Export



The highest building spent much more energy than the lowest building

Validation

1. BestModel
   - Results
   - Selection
2. Features contribution
3. **Export**

SiteEnergyUse_prediced -> Added to original DF -> Export to CSV

Index

Preparation data

1. **Training and Test Set split**
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - Concat
   - Scaling

Integration of 'ENERGYSTARScore' & 'SiteEnergyUse_prediced'

Xtrain, ytrain $\Rightarrow$ 1564 rows (80%)

Xtest, ytest $\Rightarrow$ 392 rows (20%)

Preparation data

Seattle

1. Training and Test Set split
2. **Encoding**
   - **TargetEncoder**
   - **OrdinalEncoder**
3. Scaling features
   - Concat
   - Scaling

'PrimaryPropertyType', 'Neighborhood'

	PrimaryPropertyType	Neighborhood
1911	208.867710	92.530150
42	45.602979	95.028842
475	23.656667	64.096938
1874	298.060000	110.801853
580	72.309343	110.801853
...	...	...
1190	43.864785	110.801853
1233	33.109889	64.757789
1600	119.776316	202.928522
1167	51.116552	64.096938
485	72.309343	64.757789

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

	PrimaryPropertyType	Neighborhood
1911	4.930853	3.759940
42	3.491767	3.514679
475	2.530243	3.547074
1874	5.092281	3.802733
580	3.740817	3.802733
...	...	...
1190	3.346121	3.802733
1233	2.925881	3.452801
1600	4.085347	4.564862
1167	3.405488	3.547074
485	3.740817	3.452801

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

Seattle

1. Training and Test Set split
2. Encoding
   - **TargetEncoder**
   - OrdinalEncoder
3. Scaling features
   - Concat
   - Scaling

'PrimaryPropertyType', 'Neighborhood'

	PrimaryPropertyType	Neighborhood
1911	208.867710	92.530150
42	45.602979	95.028842
475	23.656667	64.096938
1874	298.060000	110.801853
580	72.309343	110.801853
...	...	...
1190	43.864785	110.801853
1233	33.109889	64.757789
1600	119.776316	202.928522
1167	51.116552	64.096938
485	72.309343	64.757789

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

	PrimaryPropertyType	Neighborhood
1911	4.930853	3.759940
42	3.491767	3.514679
475	2.530243	3.547074
1874	5.092281	3.802733
580	3.740817	3.802733
...	...	...
1190	3.346121	3.802733
1233	2.925881	3.452801
1600	4.085347	4.564862
1167	3.405488	3.547074
485	3.740817	3.452801

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - **OrdinalEncoder**
3. Scaling features
   - Concat
   - Scaling

'YearBuilt_10y', 'ENERGYSTARScore'

	PrimaryPropertyType	Neighborhood	YearBuilt_10y	ENERGYSTARScore
1911	208.867710	92.530150	10.0	70.0
42	45.602979	95.028842	1.0	87.0
475	23.656667	64.096938	5.0	44.0
1874	298.060000	110.801853	6.0	50.0
580	72.309343	110.801853	8.0	75.0
...	...	...	...	...
1190	43.864785	110.801853	3.0	90.0
1233	33.109889	64.757789	10.0	91.0
1600	119.776316	202.928522	1.0	68.0
1167	51.116552	64.096938	1.0	14.0
485	72.309343	64.757789	11.0	95.0

DataFrame: 'Xtrain_cat_encoded', 'Xtest_cat_encoded'

	PrimaryPropertyType	Neighborhood	YearBuilt_10y	ENERGYSTARScore
1911	4.930853	3.759940	10.0	70.0
42	3.491767	3.514679	1.0	87.0
475	2.530243	3.547074	5.0	44.0
1874	5.092281	3.802733	6.0	50.0
580	3.740817	3.802733	8.0	75.0
...	...	...	...	...
1190	3.346121	3.802733	3.0	90.0
1233	2.925881	3.452801	10.0	91.0
1600	4.085347	4.564862	1.0	68.0
1167	3.405488	3.547074	1.0	14.0
485	3.740817	3.452801	11.0	95.0

DataFrame: 'Xtrain_cat_encoded_ylog', 'Xtest_cat_encoded_ylog'

Preparation data

Seattle

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. **Scaling features**
   - **Concat**
   - **Scaling**

	SiteEnergyUse(kBtu)_predicted	NumberofFloors	PropertyGFABuilding(s)	PrimaryPropertyType	Neighborhood	YearBuilt_10y	ENERGYSTARScore
1911	0.042646	-0.081862	1.832127	1.039735	-0.230292	1.089675	0.153244
42	-0.803071	-0.511019	-0.502124	-0.580065	-0.180425	-1.622662	0.772460
475	-0.803071	-0.511019	-0.516703	-0.797801	-0.797734	-0.417179	-0.793793
1874	1.920086	1.491716	0.413084	1.924639	0.134357	-0.115808	-0.575246
580	-0.170678	-0.081862	-0.371967	-0.315103	0.134357	0.486933	0.335366
...	...	...	...	...	...	...	...
1190	-0.294116	-0.224914	-0.129387	-0.597310	0.134357	-1.019920	0.881733
1233	-0.362615	-0.367967	0.020795	-0.704013	-0.784545	1.089675	0.918158
1600	-0.719874	-0.511019	0.372290	0.155832	1.972931	-1.622662	0.080395
1167	0.260035	0.061191	-0.213027	-0.525363	-0.797734	-1.622662	-1.886527
485	0.339694	0.061191	-0.517467	-0.315103	-0.784545	1.391046	1.063856

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - **Concat**
   - Scaling

	SiteEnergyUse(kBtu)_predicted	NumberofFloors	PropertyGFABuilding(s)	PrimaryPropertyType	Neighborhood	YearBuilt_10y	ENERGYSTARScore
1911	0.042646	-0.081862	1.832127	1.039735	-0.230292	1.089675	0.153244
42	-0.803071	-0.511019	-0.502124	-0.580065	-0.180425	-1.622662	0.772460
475	-0.803071	-0.511019	-0.516703	-0.797801	-0.797734	-0.417179	-0.793793
1874	1.920086	1.491716	0.413084	1.924639	0.134357	-0.115808	-0.575246
580	-0.170678	-0.081862	-0.371967	-0.315103	0.134357	0.486933	0.335366
...	...	...	...	...	...	...	...
1190	-0.294116	-0.224914	-0.129387	-0.597310	0.134357	-1.019920	0.881733
1233	-0.362615	-0.367967	0.020795	-0.704013	-0.784545	1.089675	0.918158
1600	-0.719874	-0.511019	0.372290	0.155832	1.972931	-1.622662	0.080395
1167	0.260035	0.061191	-0.213027	-0.525363	-0.797734	-1.622662	-1.886527
485	0.339694	0.061191	-0.517467	-0.315103	-0.784545	1.391046	1.063856

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Preparation data

1. Training and Test Set split
2. Encoding
   - TargetEncoder
   - OrdinalEncoder
3. Scaling features
   - Concat
   - **Scaling**

	SiteEnergyUse(kBtu)_predicted	NumberofFloors	PropertyGFABuilding(s)	PrimaryPropertyType	Neighborhood	YearBuilt_10y	ENERGYSTARScore
1911	0.042646	-0.081862	1.832127	-0.983586	-2.001881	1.089675	0.153244
42	-0.803071	-0.511019	-0.502124	-0.997863	-2.006775	-1.622662	0.772460
475	-0.803071	-0.511019	-0.516703	-1.007403	-2.006129	-0.417179	-0.793793
1874	1.920086	1.491716	0.413084	-0.981984	-2.001027	-0.115808	-0.575246
580	-0.170678	-0.081862	-0.371967	-0.995392	-2.001027	0.486933	0.335366
...	...	...	...	...	...	...	...
1190	-0.294116	-0.224914	-0.129387	-0.999308	-2.001027	-1.019920	0.881733
1233	-0.362615	-0.367967	0.020795	-1.003478	-2.008010	1.089675	0.918158
1600	-0.719874	-0.511019	0.372290	-0.991974	-1.985817	-1.622662	0.080395
1167	0.260035	0.061191	-0.213027	-0.998719	-2.006129	-1.622662	-1.886527
485	0.339694	0.061191	-0.517467	-0.995392	-2.008010	1.391046	1.063856

DataFrame: 'Xtrain_S_E_ylog', 'Xtest_S_E_ylog'

Promising models

Seattle

1. **RandomizedSearchCV**
   - **Parameters**
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	
alpha	0,01	150	rnd

Ridge, Lasso

Parameter	min	max	
alpha	0,01	150	rnd
l1_ratio	0,25	0,75	3 val

ElasticNet

Parameter	min	max	
max_features	2	5	rnd
max_leaf_nodes	2	80	rnd
n_estimators	1	250	rnd

RandomForest

Parameter	min	max	
alpha	0,01	50	rnd
Colsample bytree	0	1	rnd
max_depth	1	6	rnd
min_child weight	0	1,5	rnd
n_estimators	1	250	rnd

XGBoost

Promising models

Seattle

1. RandomizedSearchCV
   - **Parameters**
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	
alpha	0,01	150	rnd

Ridge, Lasso

Parameter	min	max	
alpha	0,01	150	rnd
l1_ratio	0,25	0,75	3 val

ElasticNet

Parameter	min	max	
max_features	2	5	rnd
max_leaf_nodes	2	80	rnd
n_estimators	1	250	rnd

RandomForest

Parameter	min	max	
alpha	0,01	50	rnd
Colsample bytree	0	1	rnd
max_depth	1	6	rnd
min_child weight	0	1,5	rnd
n_estimators	1	250	rnd

XGBoost

Promising models

1. RandomizedSearchCV
   - Parameters
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

100 iterations number

4 number of cross-validation splits

Multi-metric evaluation : r2 and mean_squared_error

Export results (.pkl)

Promising models

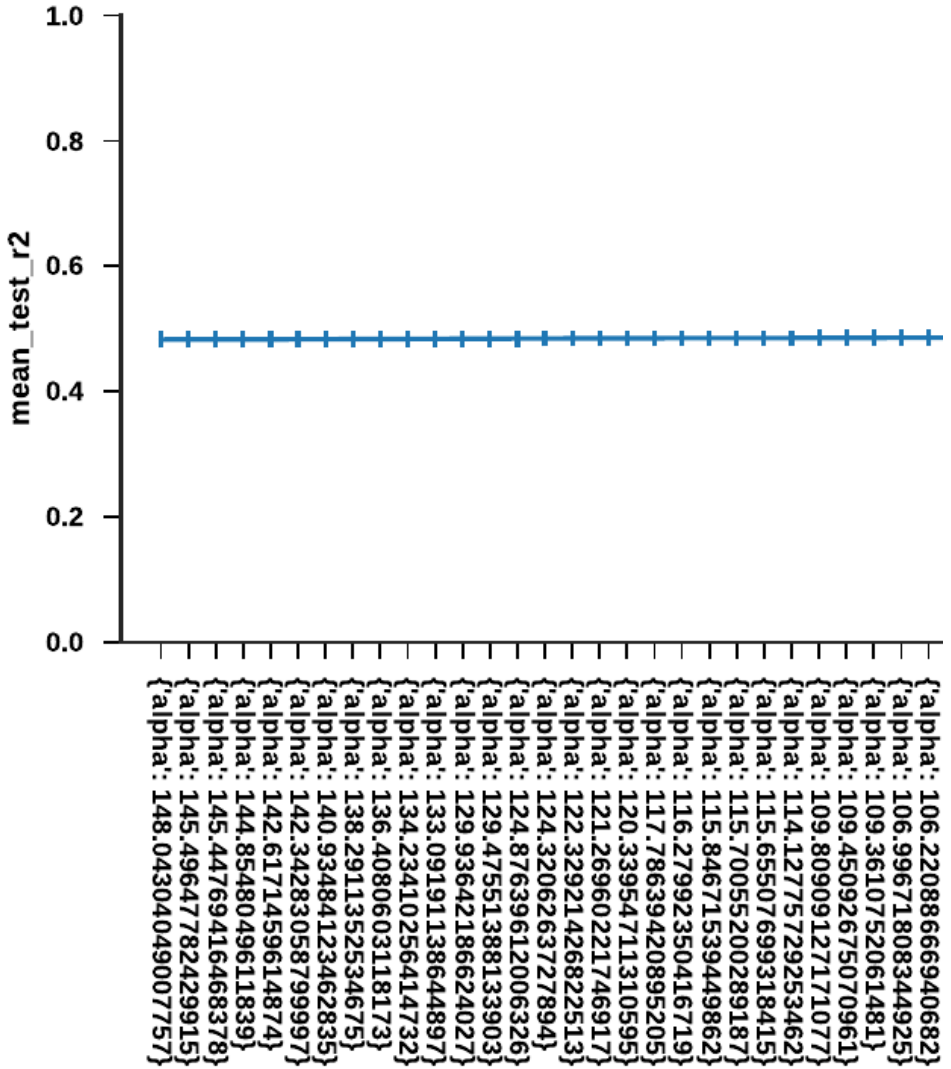
1. RandomizedSearchCV
   - Parameters
   - Training
2. **Results**
   - **r2**
3. BestModel
   - Comparison
   - Features contribution

MODELS		params	mean_test_r2	mean_test_neg_mean_squared_error
Ridge()	56	{'alpha': 13.283875307787925}	0.489799	-0.917921
	37	{'alpha': 14.66081710095758}	0.489799	-0.917918
	99	{'alpha': 16.19371404899567}	0.489796	-0.917920
Lasso()	72	{'alpha': 0.8383175685403599}	0.015184	-1.773428
	10	{'alpha': 3.097674144370367}	0.004783	-1.791773
	98	{'alpha': 3.8228690116142783}	0.001041	-1.798441
ElasticNet()	64	{'alpha': 1.0528195796786055, 'l1_ratio': 0.6330604769331742}	0.102136	-1.617531
	36	{'alpha': 0.8383175685403599, 'l1_ratio': 0.8615960713411256}	0.082113	-1.654087
	49	{'alpha': 3.8228690116142783, 'l1_ratio': 0.33091857024497834}	0.010762	-1.781177
RandomForestRegressor(random_state=42, bootstrap=[False])	62	{'max_features': 4, 'max_leaf_nodes': 71, 'n_estimators': 200}	0.654660	-0.621596
	1	{'max_features': 4, 'max_leaf_nodes': 73, 'n_estimators': 189}	0.654483	-0.621913
	74	{'max_features': 4, 'max_leaf_nodes': 69, 'n_estimators': 183}	0.654268	-0.622306
xgboost.XGBRegressor(objective="reg:linear")	81	{'alpha': 6.435017661695509, 'colsample_bytree': 0.811204176736003, 'max_depth': 3, 'min_child_weight': 1.4121972131647877, 'n_estimators': 67}	0.673004	-0.589017
	36	{'alpha': 11.891877199619984, 'colsample_bytree': 0.7282163486118596, 'max_depth': 5, 'min_child_weight': 0.9484587458903693, 'n_estimators': 198}	0.663917	-0.605746
	58	{'alpha': 4.868824688538426, 'colsample_bytree': 0.6150072266991697, 'max_depth': 2, 'min_child_weight': 0.38124547360460814, 'n_estimators': 234}	0.663811	-0.605531

Selection : Ridge, RandomForest & XGBoost

Promising models

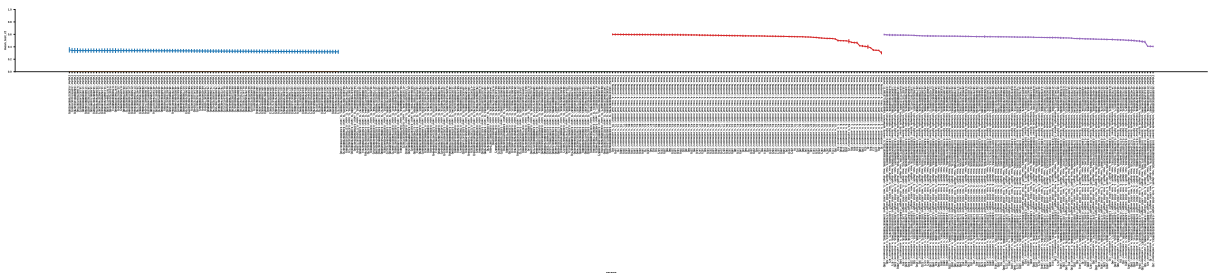
1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - **r2**
3. BestModel
   - Comparison
   - Features contribution



Alpha between: 100 - 170

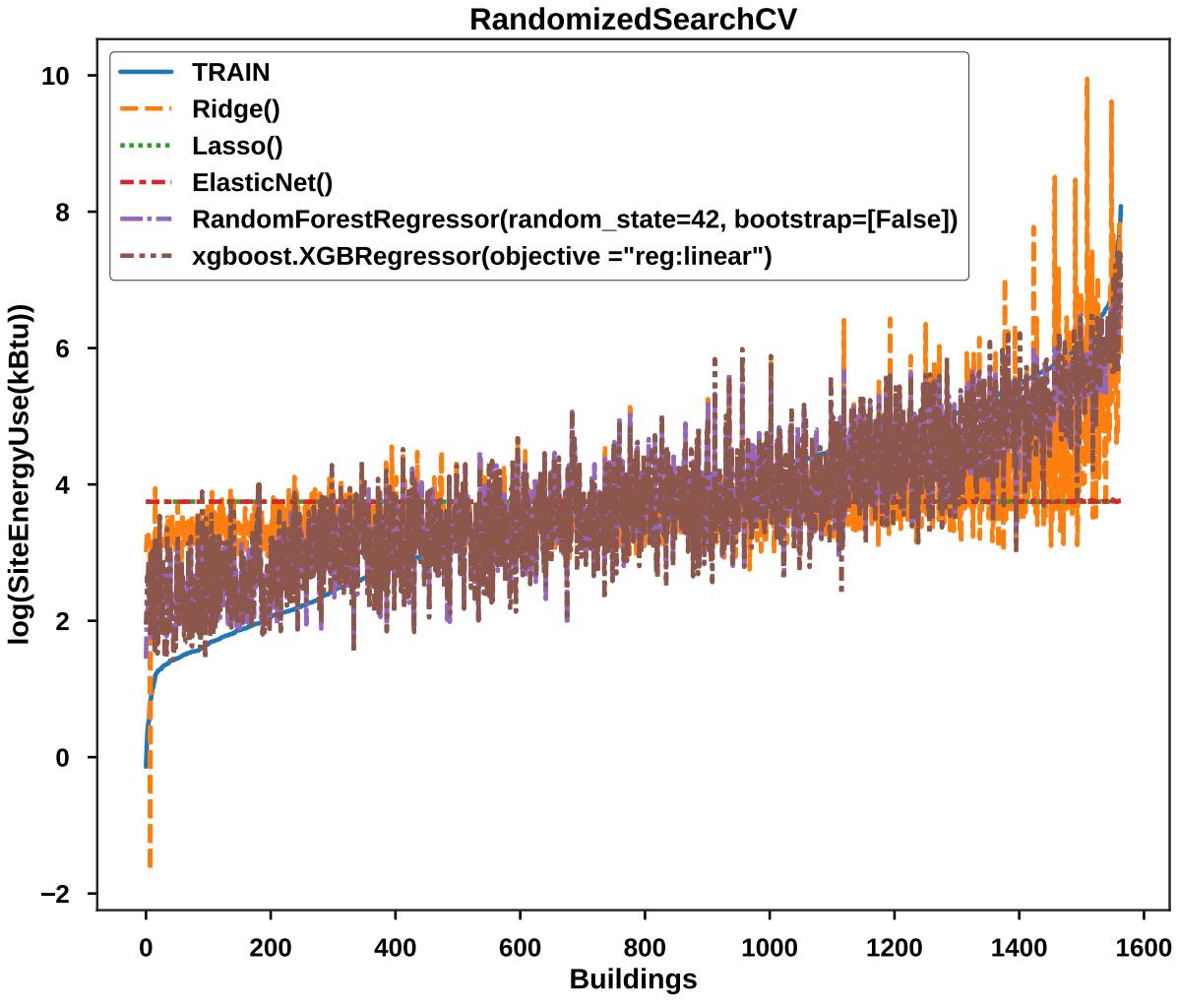
Promising models

1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - **r2**
3. BestModel
   - Comparison
   - Features contribution



Promising models

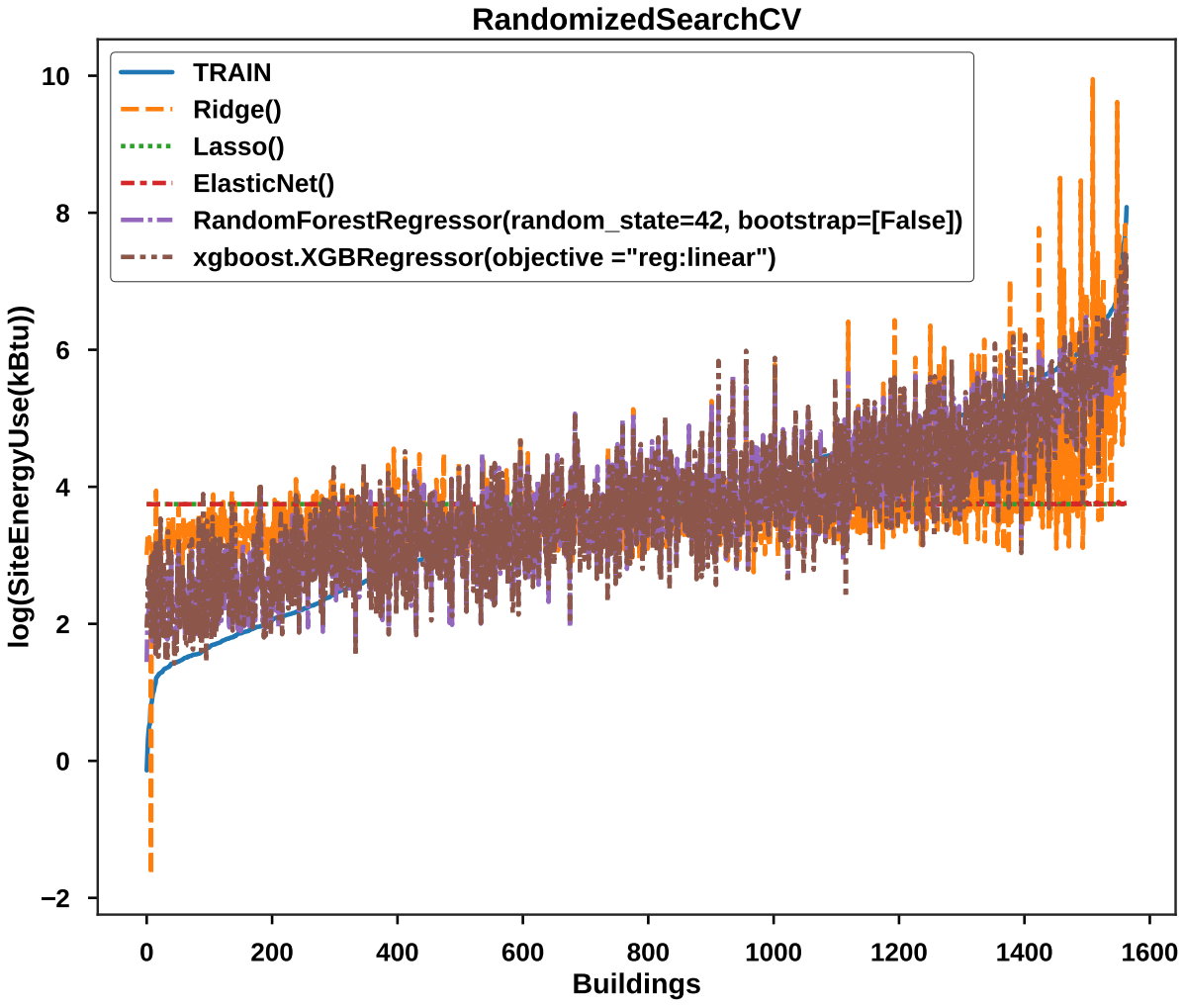
1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. **BestModel**
   - **Comparison**
   - **Features contribution**



Selection : Ridge, RandomForest & XGBoost

Promising models

1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - **Comparison**
   - Features contribution

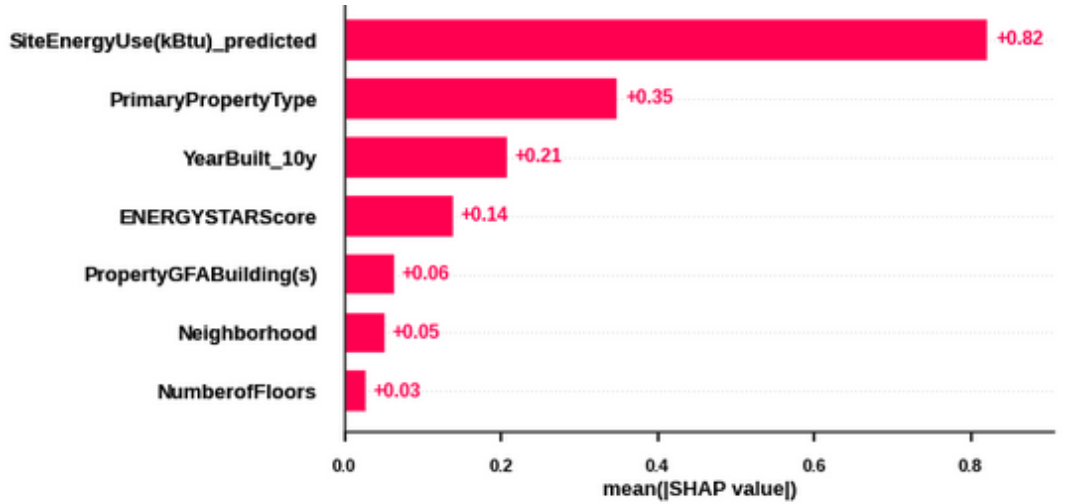


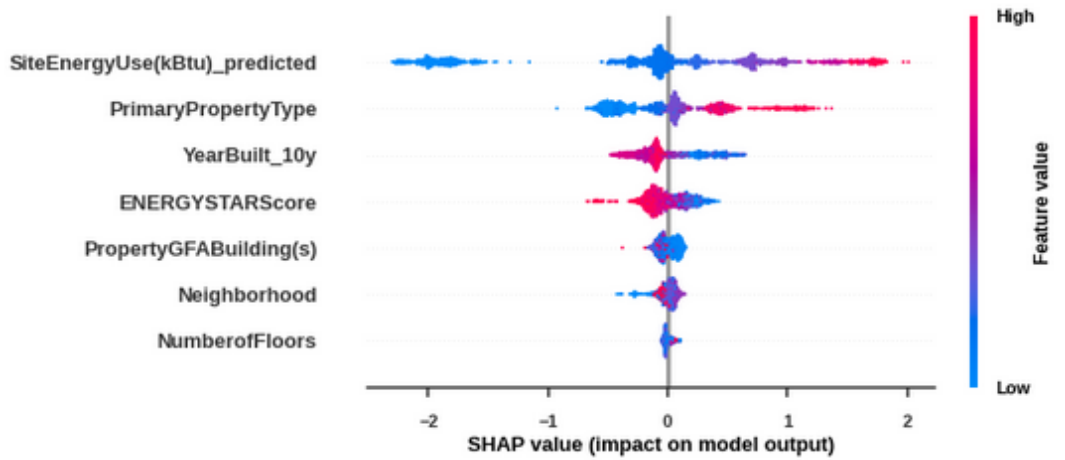
Selection : Ridge, RandomForest & XGBoost

Promising models

Seattle

1. RandomizedSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - **Features contribution**





↑ SiteEnergyUse(NumberofFloors) ⟹ ↑ CO2

↑ YearBuilt_10, ENERGYSTARTscore ⟹ ↓ CO2

Fine-tune

Seattle

1. **GridSearchCV**
   - **Parameters**
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	nbr	
alpha	0,1	100	20	log

Ridge

Parameter	min	max		
max_features	20	100	6	
max_leaf_nodes	80	150	4	lin

RandomForest

Parameter	min	max		
alpha	0,01	10	10	log
Colsample bytree	0,5	2	2	lin
max_depth	3	4	2	lin
min_child weight	0,5	1,5	2	lin
n_estimators	50	150	3	log

XGBoost

Fine-tune

1. GridSearchCV
   - **Parameters**
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Parameter	min	max	nbr	
alpha	0,1	100	20	log

Ridge

Parameter	min	max		
max_features	20	100	6	
max_leaf_nodes	80	150	4	lin

RandomForest

Parameter	min	max		
alpha	0,01	10	10	log
Colsample bytree	0,5	2	2	lin
max_depth	3	4	2	lin
min_child weight	0,5	1,5	2	lin
n_estimators	50	150	3	log

XGBoost

Fine-tune

1. GridSearchCV
   - Parameters
   - **Training**
2. Results
   - r2
3. BestModel
   - Comparison
   - Features contribution

Ridge : 20 iterations number

RandomForest : 20 iterations number

XGBoost : 360 iterations number

10 number of cross-validation splits

Multi-metric evaluation : r2 and mean_squared_error

Export results (.pkl)

Fine-tune

Seattle

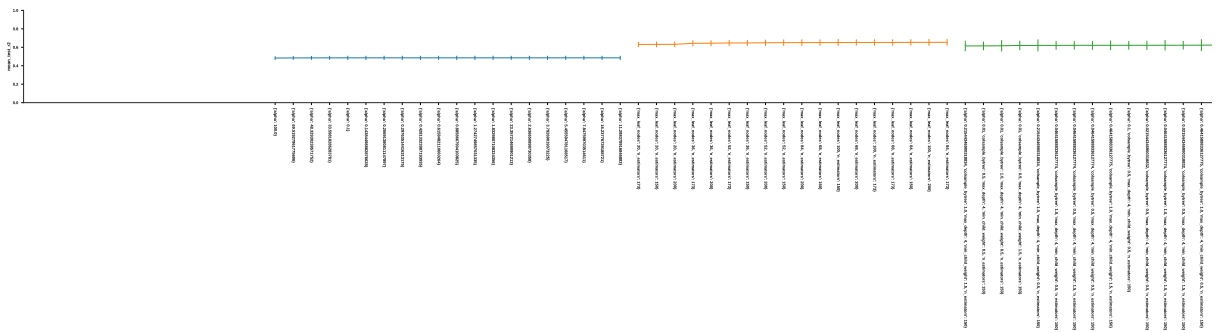
1. GridSearchCV
   - Parameters
   - Training
2. **Results**
   - **r2**
3. BestModel
   - Comparison
   - Features contribution

MODELS		params	mean_test_r2	mean_test_neg_mean_squared_error
Ridge()	13	{'alpha': 11.288378916846883}	0.485813	-0.920649
RandomForestRegressor()	13	{'max_leaf_nodes': 84, 'n_estimators': 173}	0.654044	-0.617417
xgboost.XGBRegressor(objective="reg:linear")	210	{'alpha': 4.6415888336127775, 'colsample_bytree': 1.0, 'max_depth': 4, 'min_child_weight': 0.5, 'n_estimators': 50}	0.668262	-0.591426

Selection : XGBoost

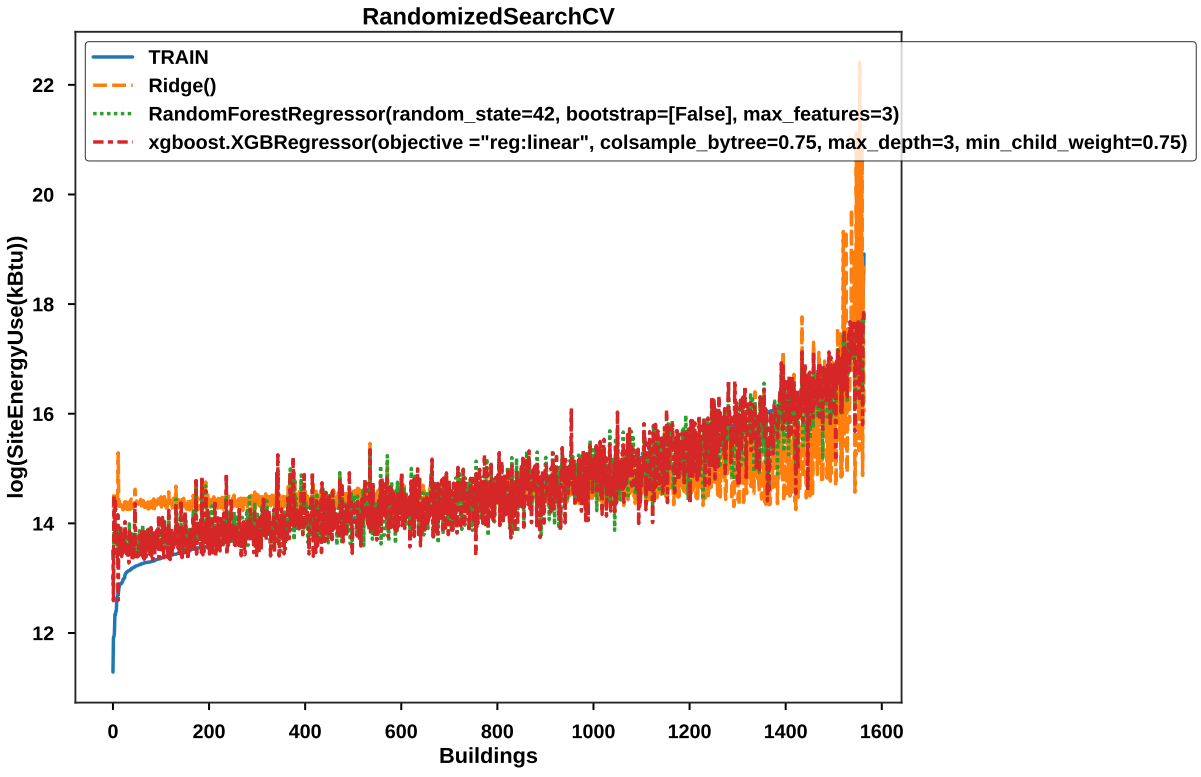
Fine-tune

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - **r2**
3. BestModel
   - Comparison
   - Features contribution



Fine-tune

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. **BestModel**
   - **Comparison**
   - **Features contribution**

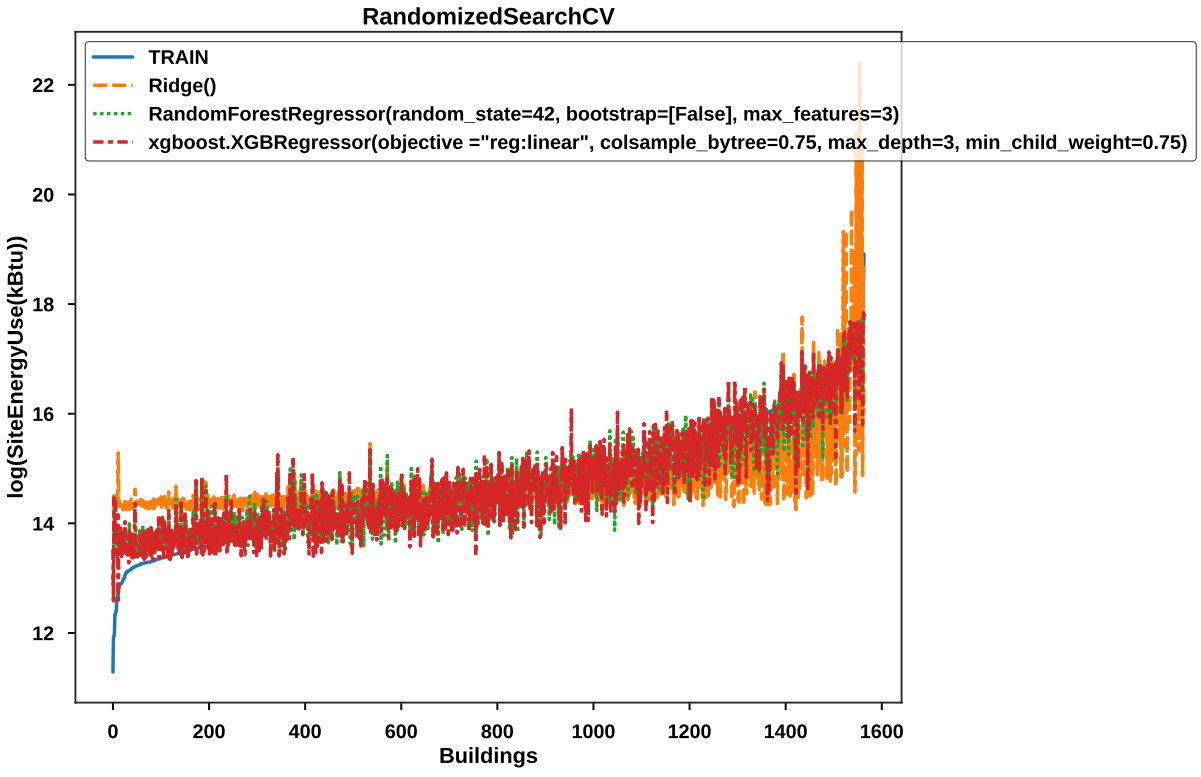


Selection : XGBoost

Fine-tune

Seattle

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - **Comparison**
   - Features contribution

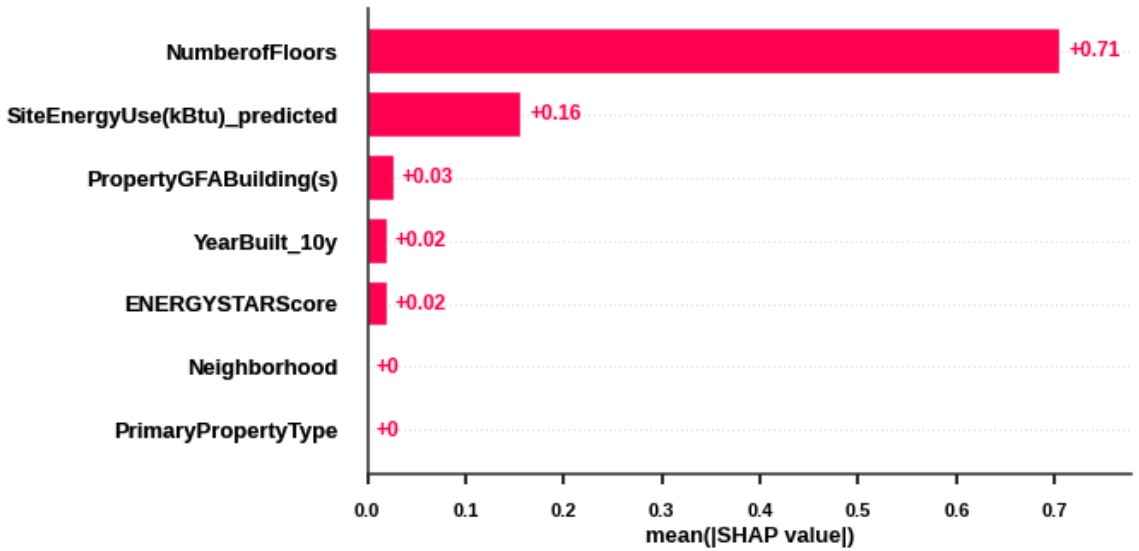


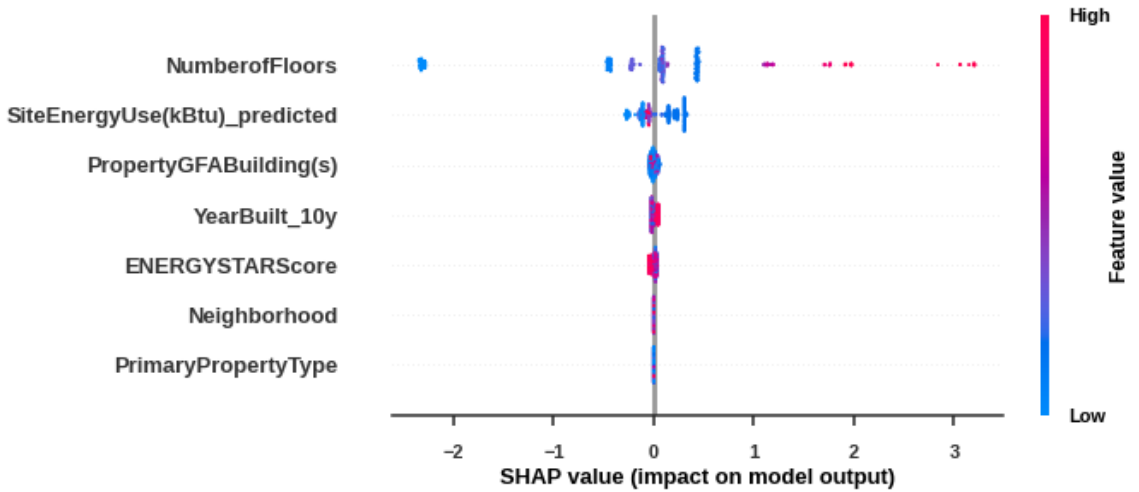
Selection : XGBoost

Fine-tune

Seattle

1. GridSearchCV
   - Parameters
   - Training
2. Results
   - r2
3. BestModel
   - Comparison
   - **Features contribution**





The highest building spent much more CO2 than the smallest building

Validation

1. **BestModel**
   - **Results**
   - **Selection**
2. Features contribution
3. RMSE-PrimaryPropertyType
4. Recursive Feature Elimination (RFE)

	params	mean_test_r2	mean_test_neg_mean_squared_error
xgboost.XGBRegressor(objective="reg:linear")	{'alpha': 4.6415888336127775, 'colsample_bytre...	0.831823	0.340938

Validation

Seattle

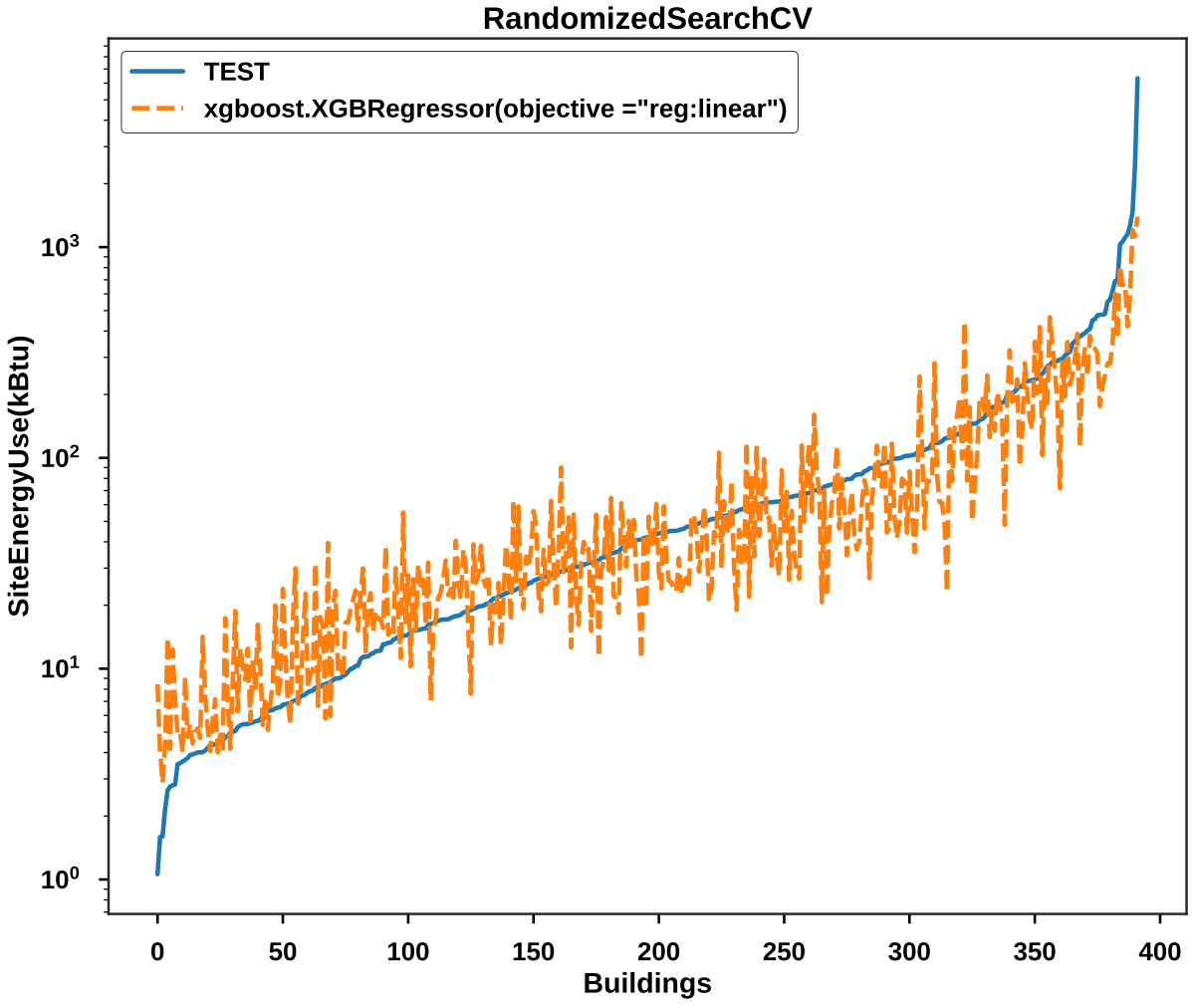
1. BestModel
   - **Results**
   - Selection
2. Features contribution
3. RMSE-PrimaryPropertyType
4. Recursive Feature Elimination (RFE)

	params	mean_test_r2	mean_test_neg_mean_squared_error
xgboost.XGBRegressor(objective="reg:linear")	{'alpha': 4.6415888336127775, 'colsample_bytre...	0.831823	0.340938

Validation

Seattle

1. BestModel
   - Results
   - **Selection**
2. Features contribution
3. RMSE-PrimaryPropertyType
4. Recursive Feature Elimination (RFE)

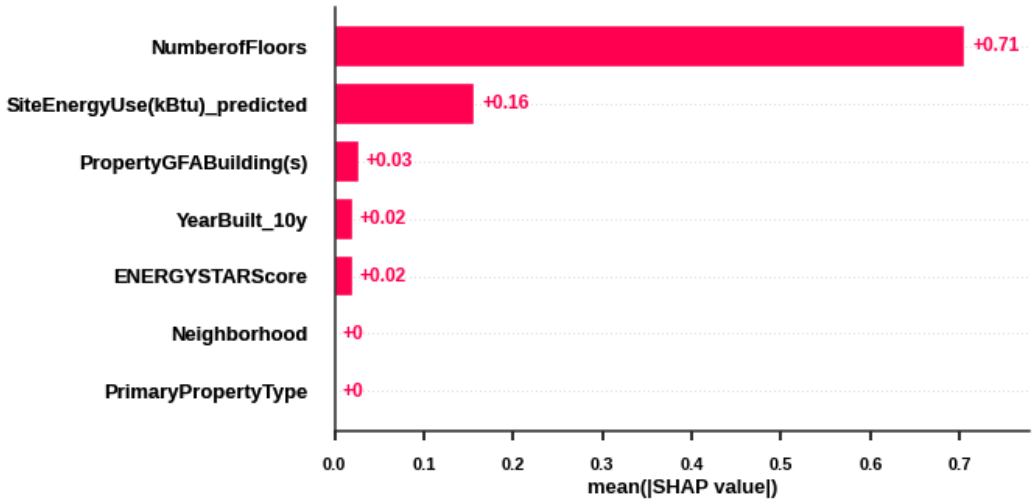


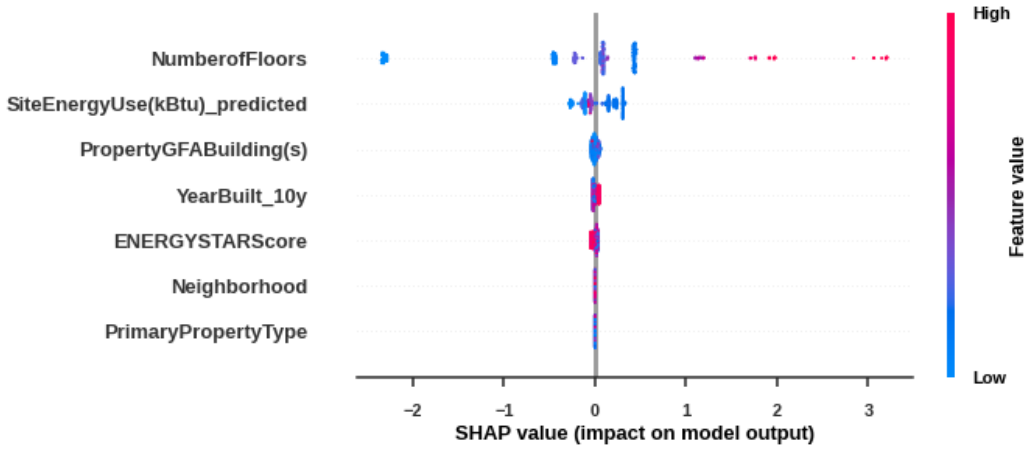
The XGBoost model is validated

Validation

Seattle

1. BestModel
   - Results
   - Selection
2. **Features contribution**
3. RMSE-PrimaryPropertyType
4. Recursive Feature Elimination (RFE)



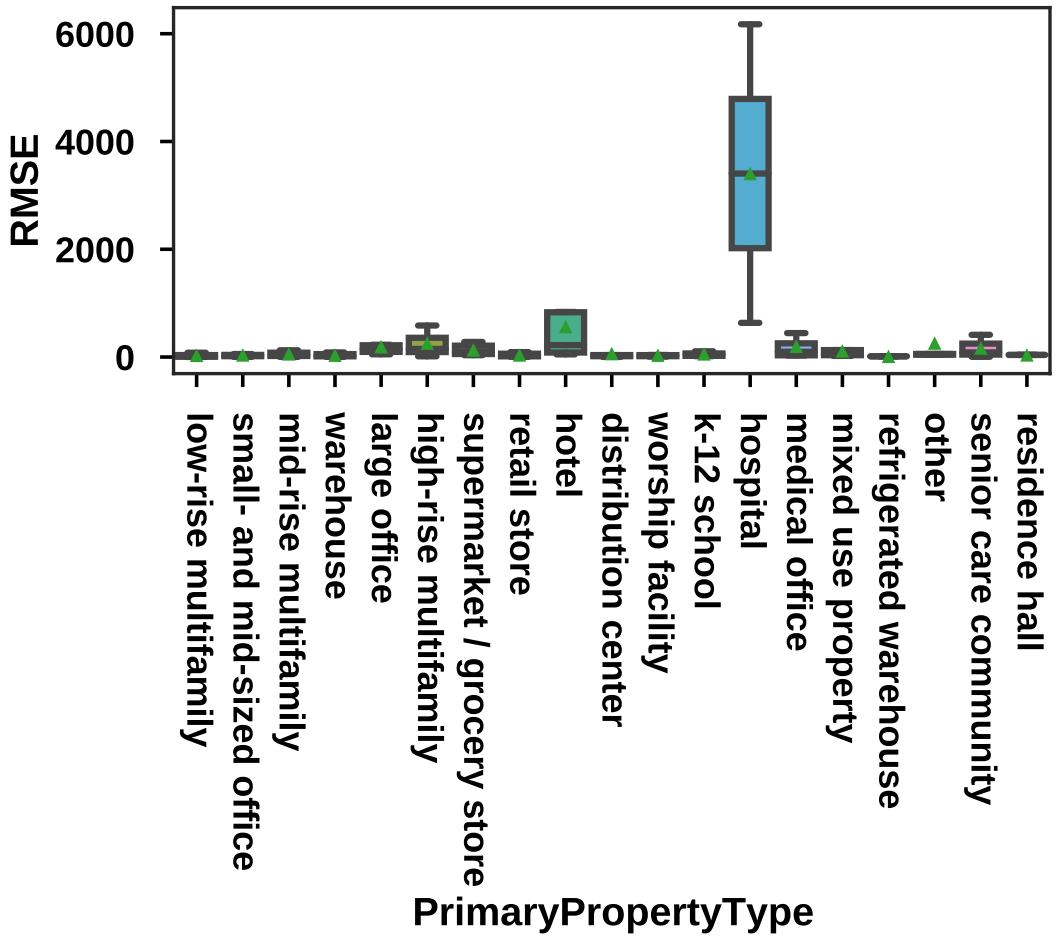


The highest building spent much more CO2 than the smallest building

Validation

Seattle

1. BestModel
   - Results
   - Selection
2. Features contribution
3. **RMSE-PrimaryPropertyType**
4. Recursive Feature Elimination (RFE)

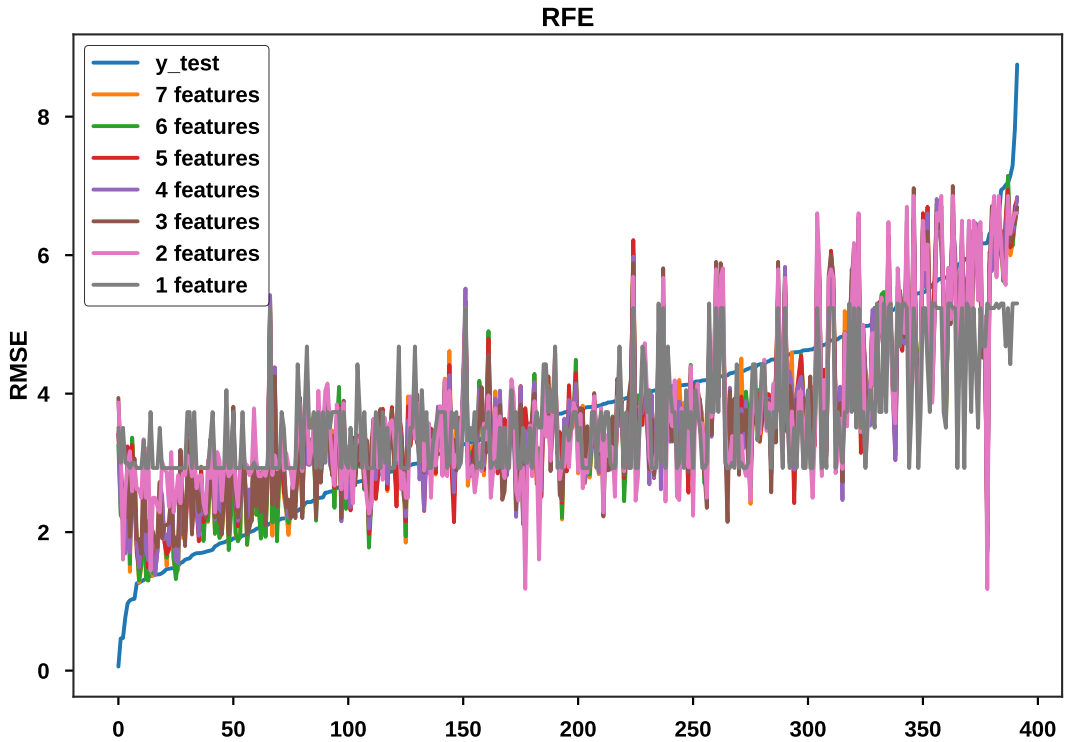


⇑⇑ **RMSE from Hostipal**

Validation

Seattle

1. BestModel
   - Results
   - Selection
2. Features contribution
3. RMSE-PrimaryPropertyType
4. **Recursive Feature Elimination (RFE)**



Index

Conclusions - SiteEnergyUse

1. Features selection:

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
2. Results:

	r2	MSE
Train	0.79	-0.23
Test	0.94	0.071

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
2. Results:
   - Best model (r2, MSE) during training : RandomForest, ~~XGBoost & Ridge~~

	r2	MSE
Train	0.79	-0.23
Test	0.94	0.071

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
2. Results:
   - Best model (r2, MSE) during training : RandomForest, ~~XGBoost & Ridge~~

	r2	MSE
Train	0.79	-0.23
Test	0.94	0.071

3. Features contribution:

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
2. Results:
   - Best model (r2, MSE) during training : RandomForest, ~~XGBoost & Ridge~~

	r2	MSE
Train	0.79	-0.23
Test	0.94	0.071

3. Features contribution:
   - ↑↑ NumberofFloors, ↗ PropertyGFABuildings

Conclusions - SiteEnergyUse

1. Features selection:
   - PropertyGFABuildings(s), ~~TotalEnergy(Electricity)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
2. Results:
   - Best model (r2, MSE) during training : RandomForest, ~~XGBoost & Ridge~~

	r2	MSE
Train	0.79	-0.23
Test	0.94	0.071

3. Features contribution:
   - ↑↑ NumberofFloors, ↗ PropertyGFABuildings
   - The 'biggest' building spent much more energy than the 'smallest' building

Conclusions - GHGEmissions

1. Features selection:

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore
2. Results:

	r2	MSE
Train	0.67	-0.59
Test	0.83	0.34

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore
2. Results:
   - Best model (r2, MSE) during training : XGBoost, ~~RandomForest & Ridge~~

	r2	MSE
Train	0.67	-0.59
Test	0.83	0.34

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore
2. Results:
   - Best model (r2, MSE) during training : XGBoost, ~~RandomForest & Ridge~~

	r2	MSE
Train	0.67	-0.59
Test	0.83	0.34

3. Features contribution:

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore
2. Results:
   - Best model (r2, MSE) during training : XGBoost, ~~RandomForest & Ridge~~

	r2	MSE
Train	0.67	-0.59
Test	0.83	0.34

3. Features contribution:
   - ↑↑ NumberofFloors, ↑ SiteEnergyUse_predicted

Conclusions - GHGEmissions

1. Features selection:
   - PropertyGFABuildings(s) ~~TotalEnergy(Electricity, SteamUse, NaturalGas)~~
   - 'YearBuilt', 'NumberofFloors'
   - Neighborhood(downtown), ~~'CouncilDistrictCode', 'ZipCode'~~
   - PrimaryPropertyType (hospital), ~~'BuildingType'~~
   - +SiteEnergyUse_predicted & ENERGYSTARTscore
2. Results:
   - Best model (r2, MSE) during training : XGBoost, ~~RandomForest & Ridge~~

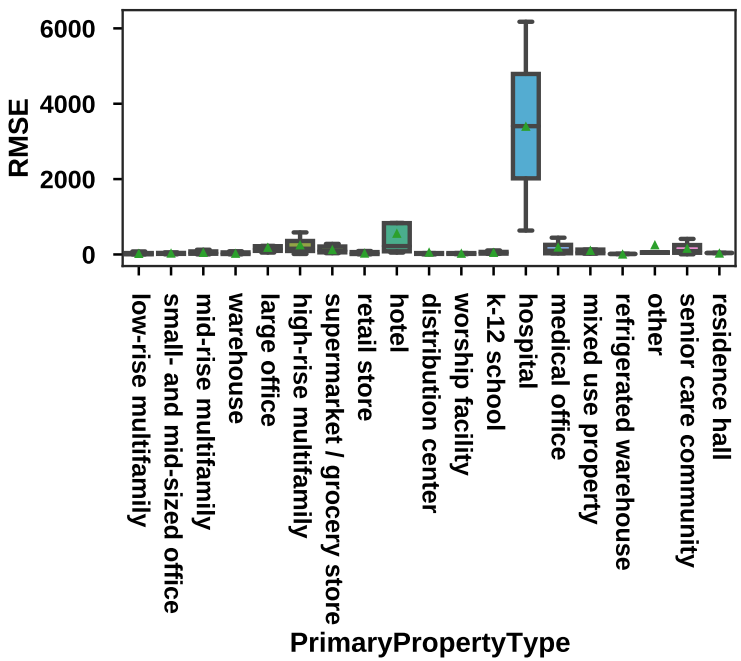
	r2	MSE
Train	0.67	-0.59
Test	0.83	0.34

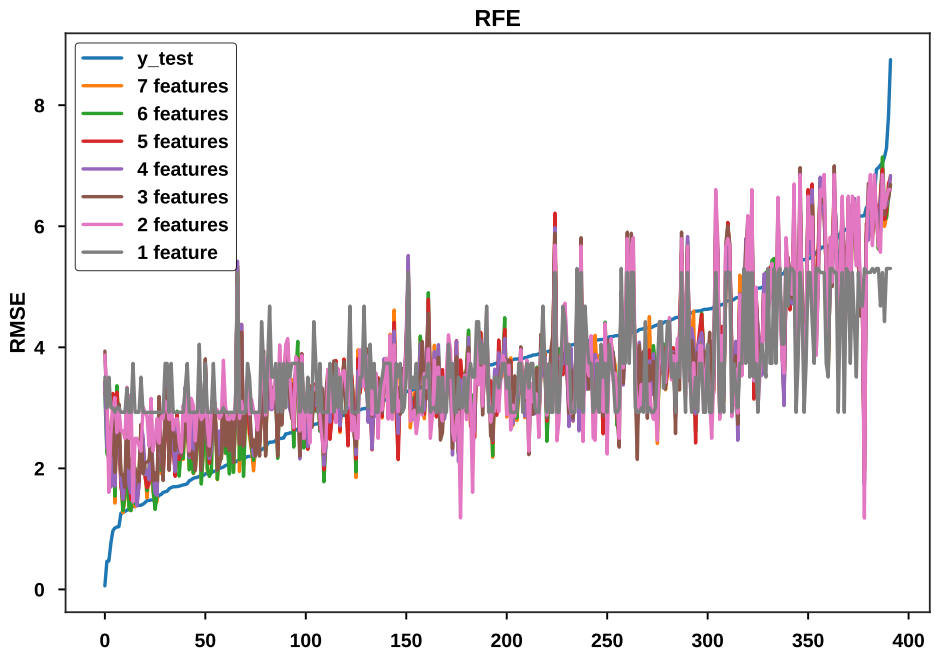
3. Features contribution:
   - ↑↑ NumberofFloors, ↑ SiteEnergyUse_predicted
   - The highest building spent much more CO2 than the lowest building

Perspectives

Seattle

1. Improve the RMSE from Hostipal / Hotel

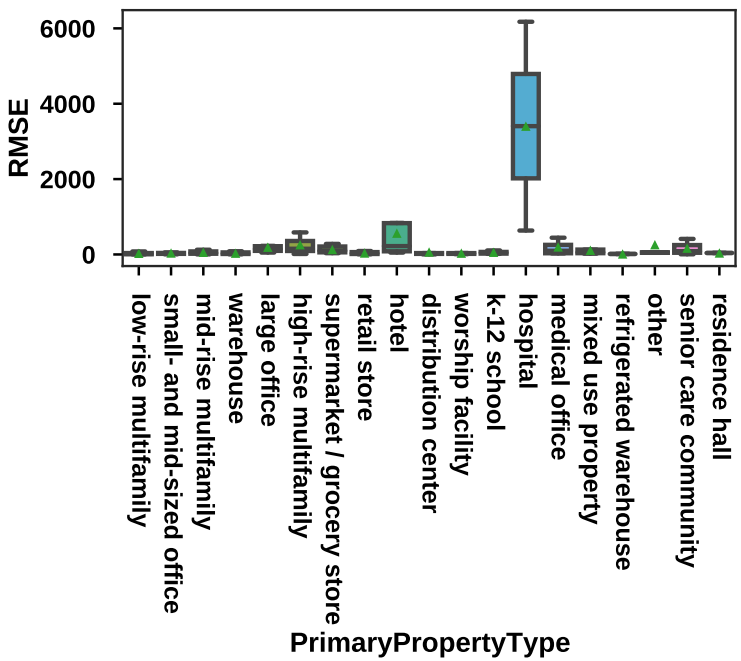


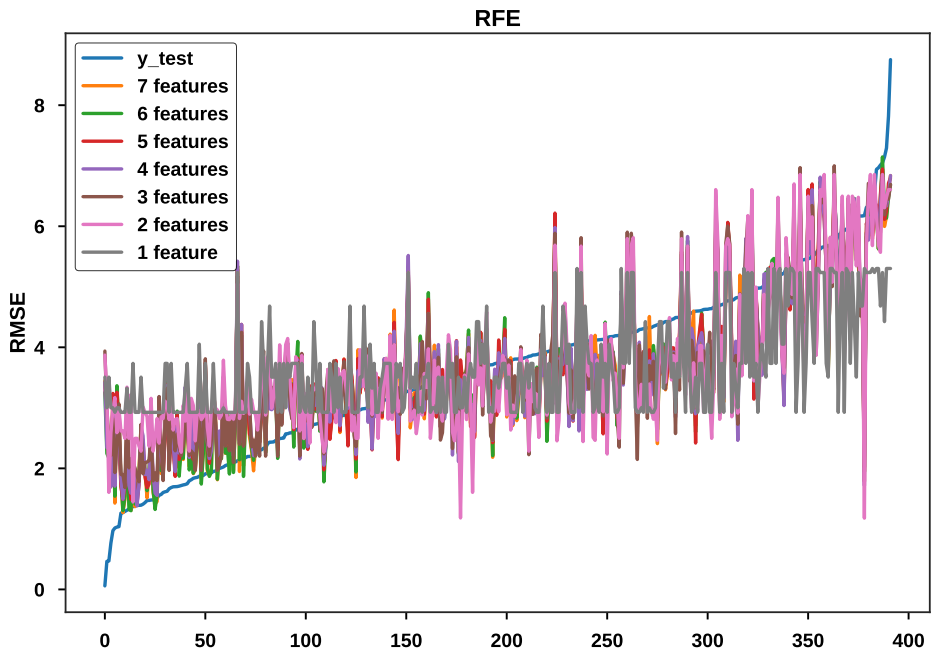


Perspectives

Seattle

1. Improve the RMSE from Hostipal / Hotel
   - New data

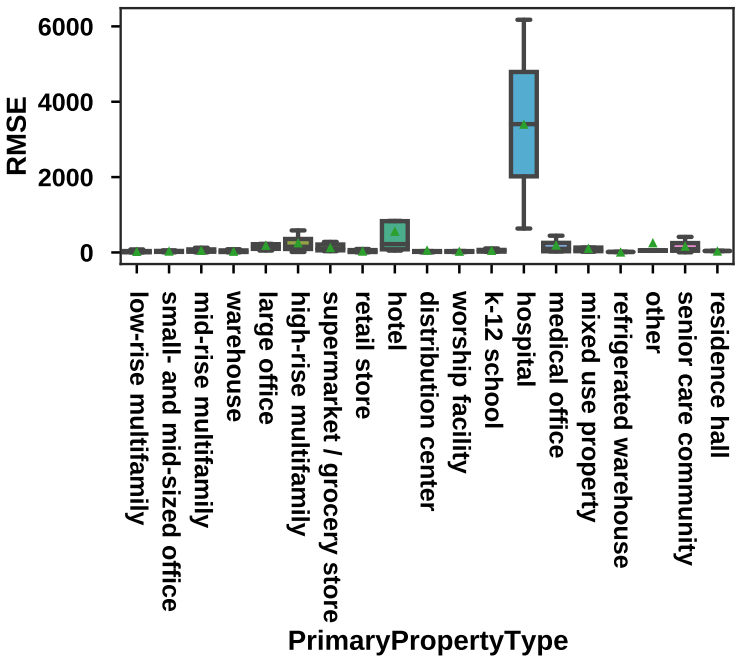


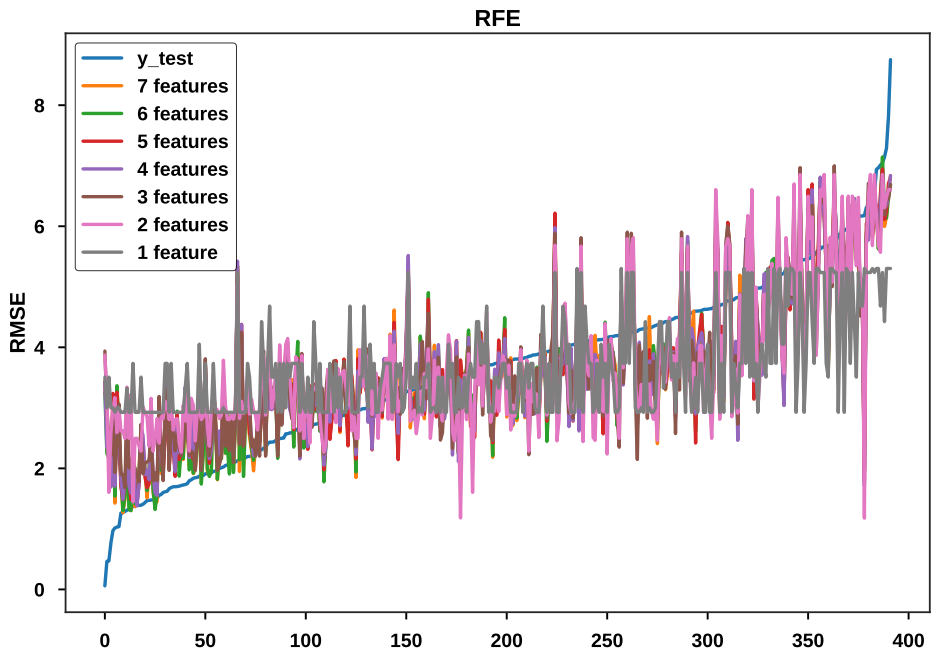


Perspectives

Seattle

1. Improve the RMSE from Hostipal / Hotel
   - New data
   - Specific model





Perspectives

Seattle

1. Improve the RMSE from Hostipal / Hotel
   - New data
   - Specific model
2. Recursive feature elimination