ML Engineer

CentraleSupélec

Project 3

Prediction of buildings energy consumption

Victor Benard

MISSION STATEMENT

City of Seattle

Seattle city's objective is to reach carbon neutral emissions by 2050. As energy metering reports are costly, there is an incentive to predict CO2 emissions and total energy consumption based on selected data.

TABLE OF CONTENTS

01

OVERVIEW & TARGETS

OVERVIEW & TARGETS

Seattle buildings data is from their open Data program
https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy

Objective:

From previous metering reports, build a ML model to predict energy consumption and CO2 emission, based solely on commercial data.

Key information:

- Data available for 2 years: 2015 and 2016 = 2 dataframes
- Dataframe size is only (3340, 47) and (3376, 46) resp.
- Target features -to be predicted
  - Energy consumption -> 'SiteEnergyUse(kBtu)'
  - CO2 emissions -> 'TotalGHGEmissions'

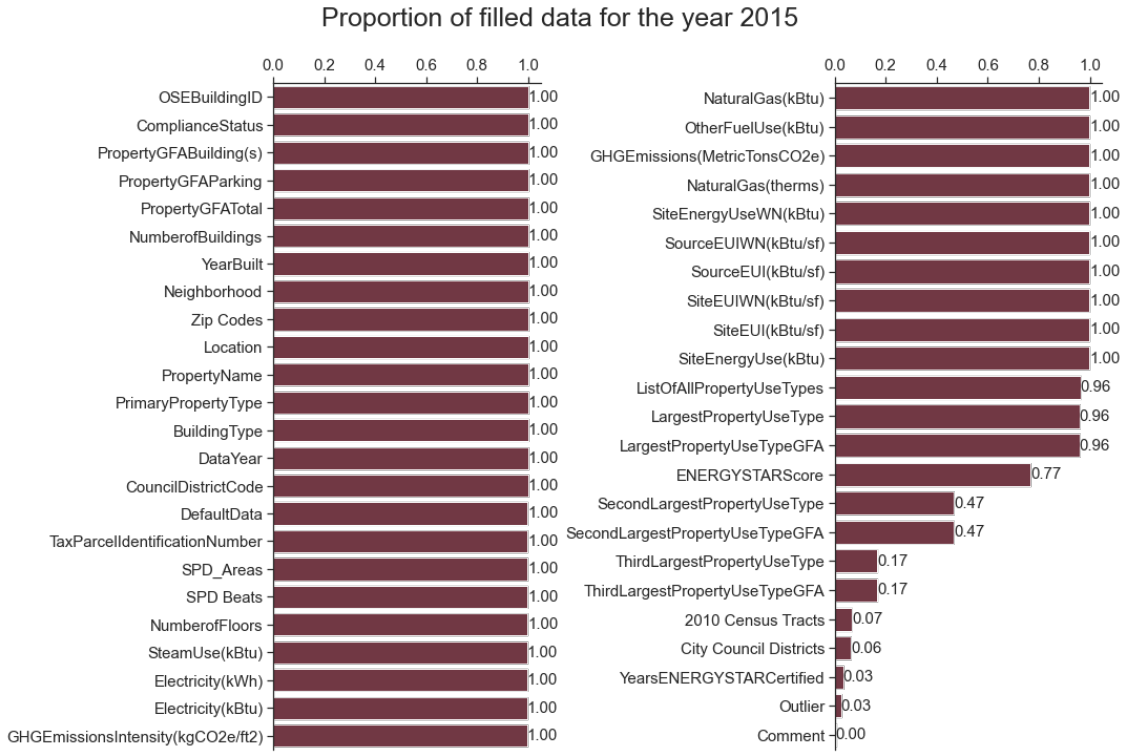
OVERVIEW & TARGETS

Features overview – can be categorized

General information	Location	Energies	Type and size
OSEBuildingID	Address	YearsENERGYSTARCertified	BuildingType
PropertyName	City	ENERGYSTARScore	PrimaryPropertyType
TaxParcelIdentificationNumber	State	SiteEUI(kBtu/sf)	YearBuilt
DataYear	ZipCode	SiteEUIWN(kBtu/sf)	NumberofBuildings
DefaultData	CouncilDistrictCode	SourceEUI(kBtu/sf)	NumberofFloors
Comments	Neighborhood	SourceEUIWN(kBtu/sf)	PropertyGFATotal
ComplianceStatus	Latitude	SiteEnergyUse(kBtu)	PropertyGFAParking
Outlier	Longitude	SiteEnergyUseWN(kBtu)	PropertyGFABuilding(s)
		SteamUse(kBtu)	ListOfAllPropertyUseTypes
		Electricity(kWh)	LargestPropertyUseType
		Electricity(kBtu)	SecondLargestPropertyUseType
		NaturalGas(therms)	ThirdLargestPropertyUseType
		NaturalGas(kBtu)	LargestPropertyUseTypeGFA
		TotalGHGEmissions	SecondLargestPropertyUseTypeGFA
		GHGEmissionsIntensity	ThirdLargestPropertyUseTypeGFA

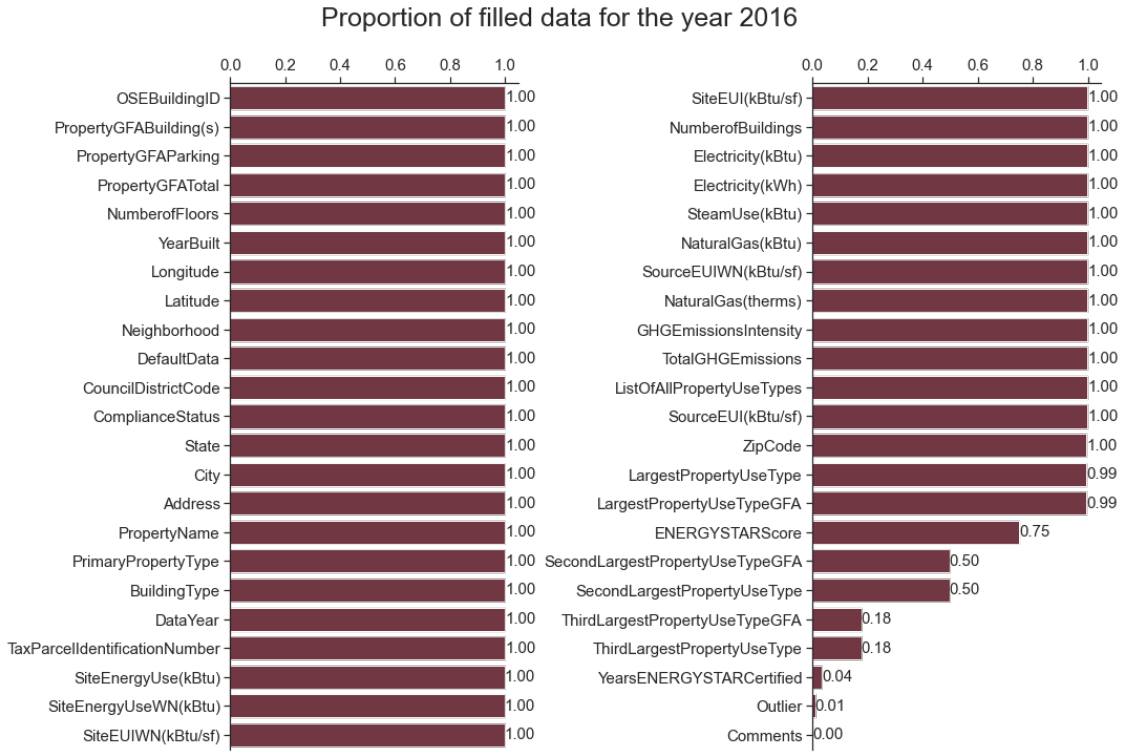
OVERVIEW & TARGETS

Non missing data



OVERVIEW & TARGETS

Non missing data -> Missing data insignificant for most features of both years



02

DATA CLEANING & ANALYSIS

DATA CLEANING & ANALYSIS

Cleaning

Combine data ▸▸▸ Duplicates ▸▸▸ Features selection ▸▸▸ Features cleaning

Data exploratory analysis

Univariate

Multivariate
- ▸▸▸ Correlation matrix
- ▸▸▸ PCA
- ▸▸▸ ANOVA

DATA CLEANING

Combination of 2015 and 2016 data

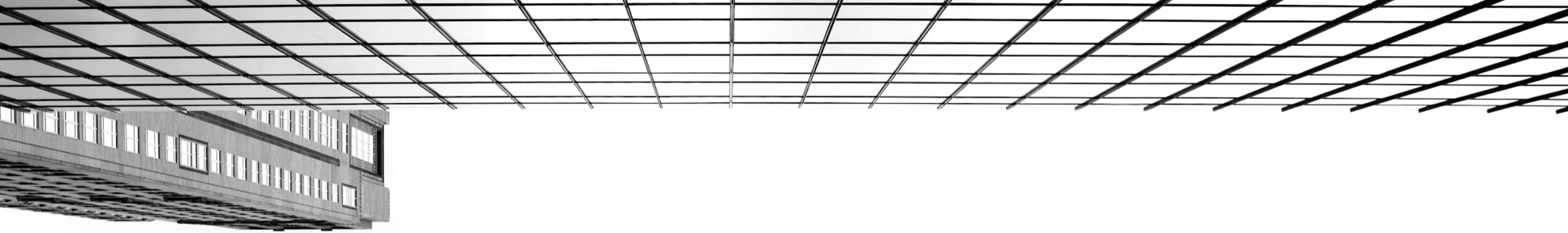
Comparison:

- Features
  - 9 features names differ between 2015 and 2016, mostly because of location features `diff = [i for i in df15 if i not in df16]`
- Building ID comparison
  - 56 new buildings between 2015 and 2016
  - 92 buildings missing in 2016 compared to 2015

Actions:

- Rename 2015 features as per 2016 model
- Consider 2016 as main, and complete missing data with 2015

```
df0 = df16.combine_first(df15_tidy)
```



0

Duplicates

```
df0.duplicated('OSEBuildingID').sum()
```

DATA CLEANING

Features selection

Criteria:

- Less than 50% missing values
- No energy-related features except targets
- Removal of irrelevant informative features

General information	Location	Energies	Type and size
OSEBuildingID	Address	YearsENERGYSTARCertified	BuildingType
PropertyName	City	ENERGYSTARScore	PrimaryPropertyType
TaxParcelIdentificationNumber	State	SiteEUI(kBtu/sf)	YearBuilt
DataYear	ZipCode	SiteEUIWN(kBtu/sf)	NumberofBuildings
DefaultData	CouncilDistrictCode	SourceEUI(kBtu/sf)	NumberofFloors
Comments	Neighborhood	SourceEUIWN(kBtu/sf)	PropertyGFATotal
ComplianceStatus	Latitude	Energy	PropertyGFAParking
Outlier	Longitude	SiteEnergyUseWN(kBtu)	PropertyGFABuilding(s)
		SteamUse(kBtu)	ListOfAllPropertyUseTypes
		Electricity(kWh)	LargestPropertyUseType
		Electricity(kBtu)	SecondLargestPropertyUseType
		NaturalGas(therms)	ThirdLargestPropertyUseType
		NaturalGas(kBtu)	LargestPropertyUseTypeGFA
		Emissions	SecondLargestPropertyUseTypeGFA
		GHGEmissionsIntensity	ThirdLargestPropertyUseTypeGFA

DATA CLEANING

Target features cleaning

A `.describe()` shows negative values for CO2 emissions. They are corrected to 0.

Input features cleaning

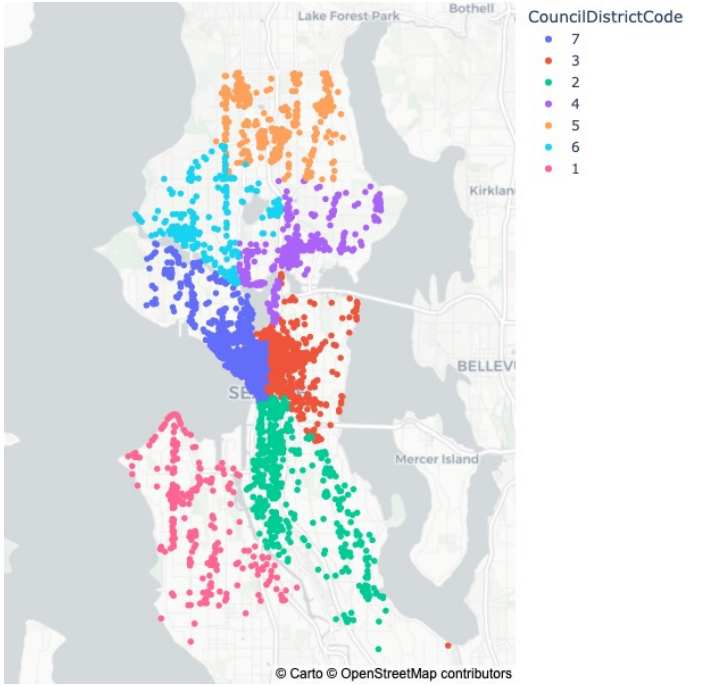
Five categories:

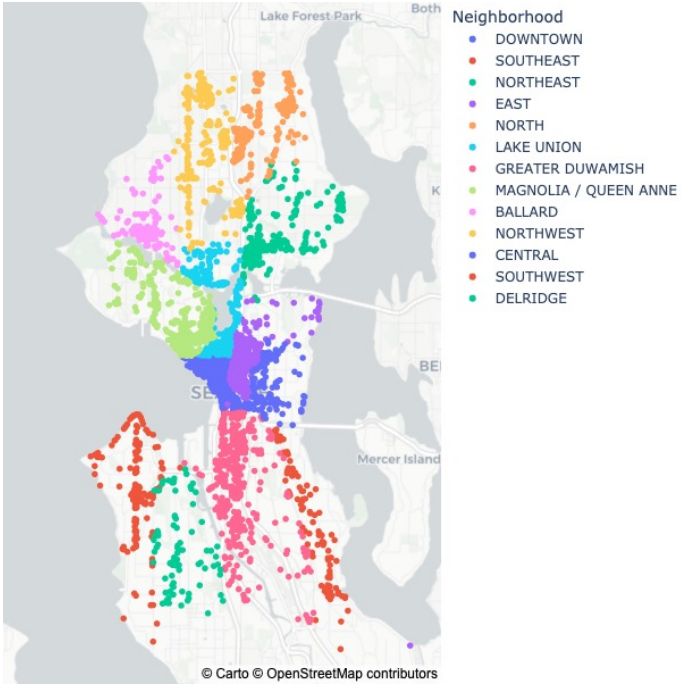
- Location
- Building type
- Building size
- Construction date
- Energy star score

DATA CLEANING

Location features

- ‘Neighborhood’ bins are treated -> 19 categories.
- ‘CouncilDistrictCode’ contains 7 categories.
- ‘ZipCode’ has too many categories -> discarded

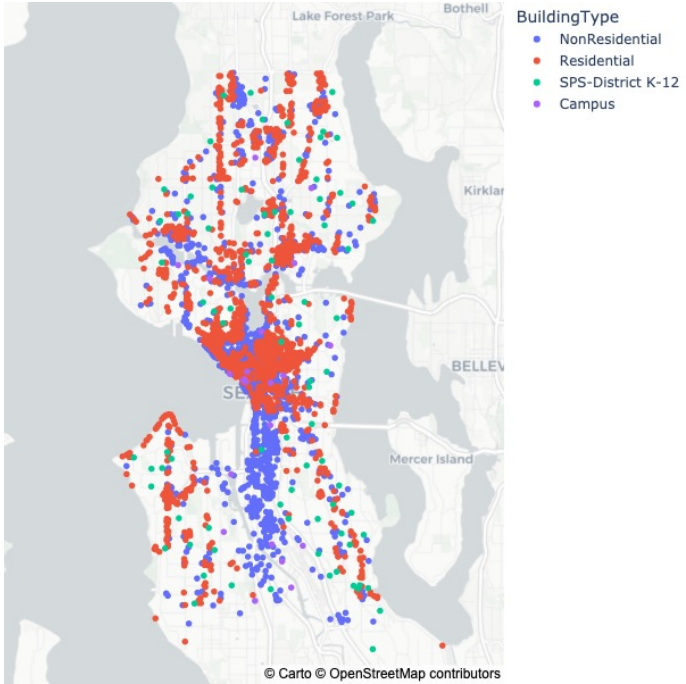


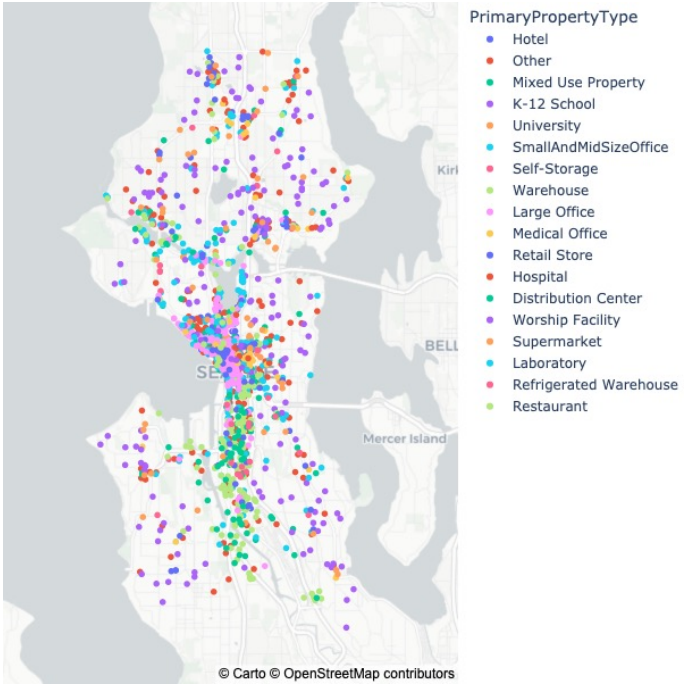


DATA CLEANING

Type features

- Deletion of residential properties
- Regrouped few ‘PrimaryPropertyType’ categories

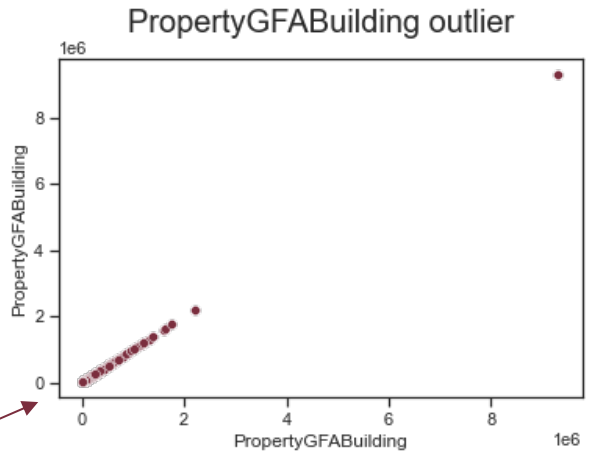




DATA CLEANING

Construction features

- Number of buildings
  - Binned into categories: 1 ; and more than 1
  - Values at 0 corrected to 1
- Number of floors
  - Row where Nb of floors over 76 removed
  - Values at 0 corrected to 1
- Building size
  - ‘PropertyGFABuilding’ contains 1 outlier
  - Creation of a new feature: Total inner surface = GFA*Nb of floors
- Construction date is binned per decade
- Energy star score is checked, no outlier



DATA CLEANING & ANALYSIS

Cleaning

Combine data ▸▸▸ Duplicates ▸▸▸ Features selection ▸▸▸ Features cleaning

Data exploratory analysis

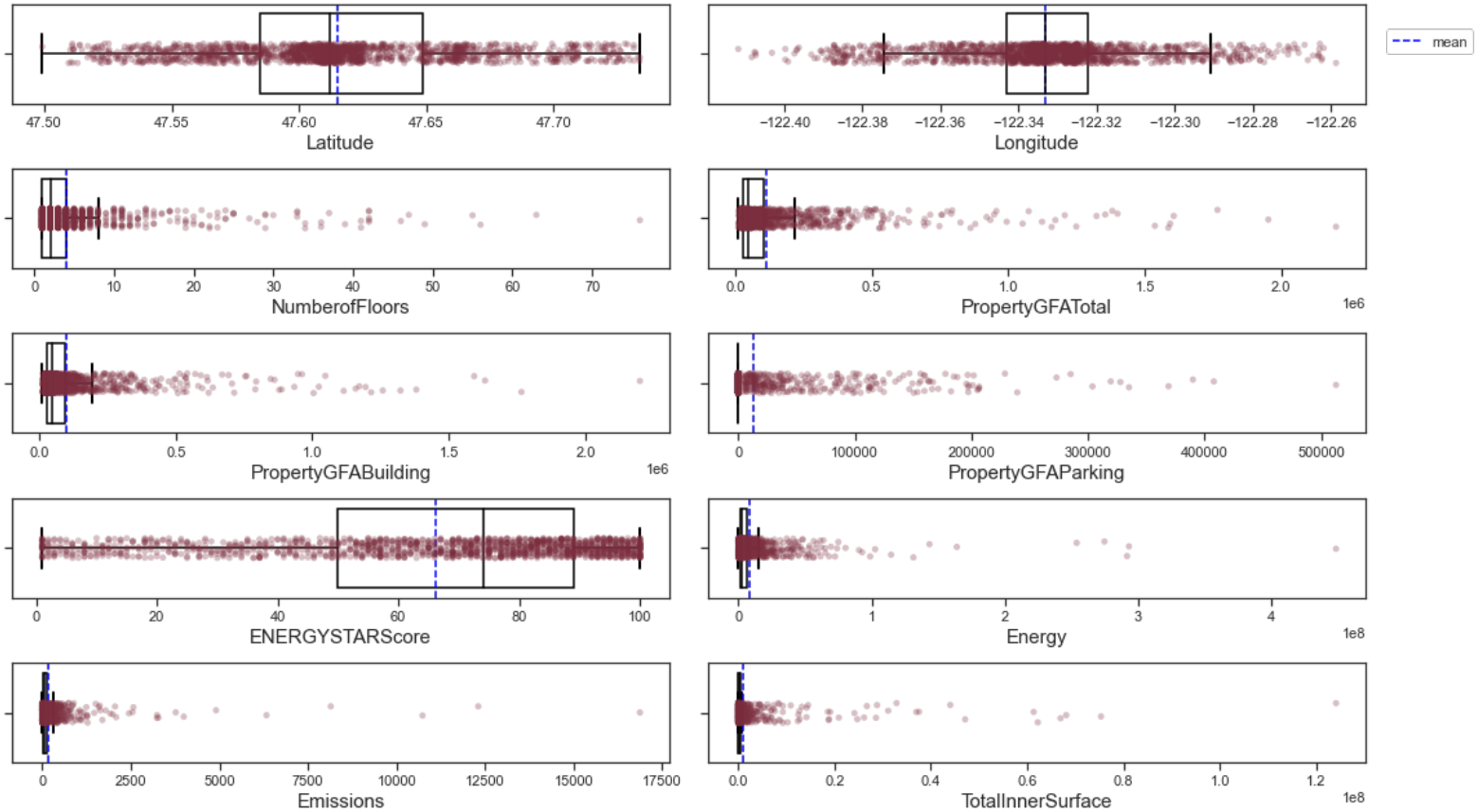
Univariate

Multivariate

- ▸▸▸ Correlation matrix
- ▸▸▸ PCA
- ▸▸▸ ANOVA

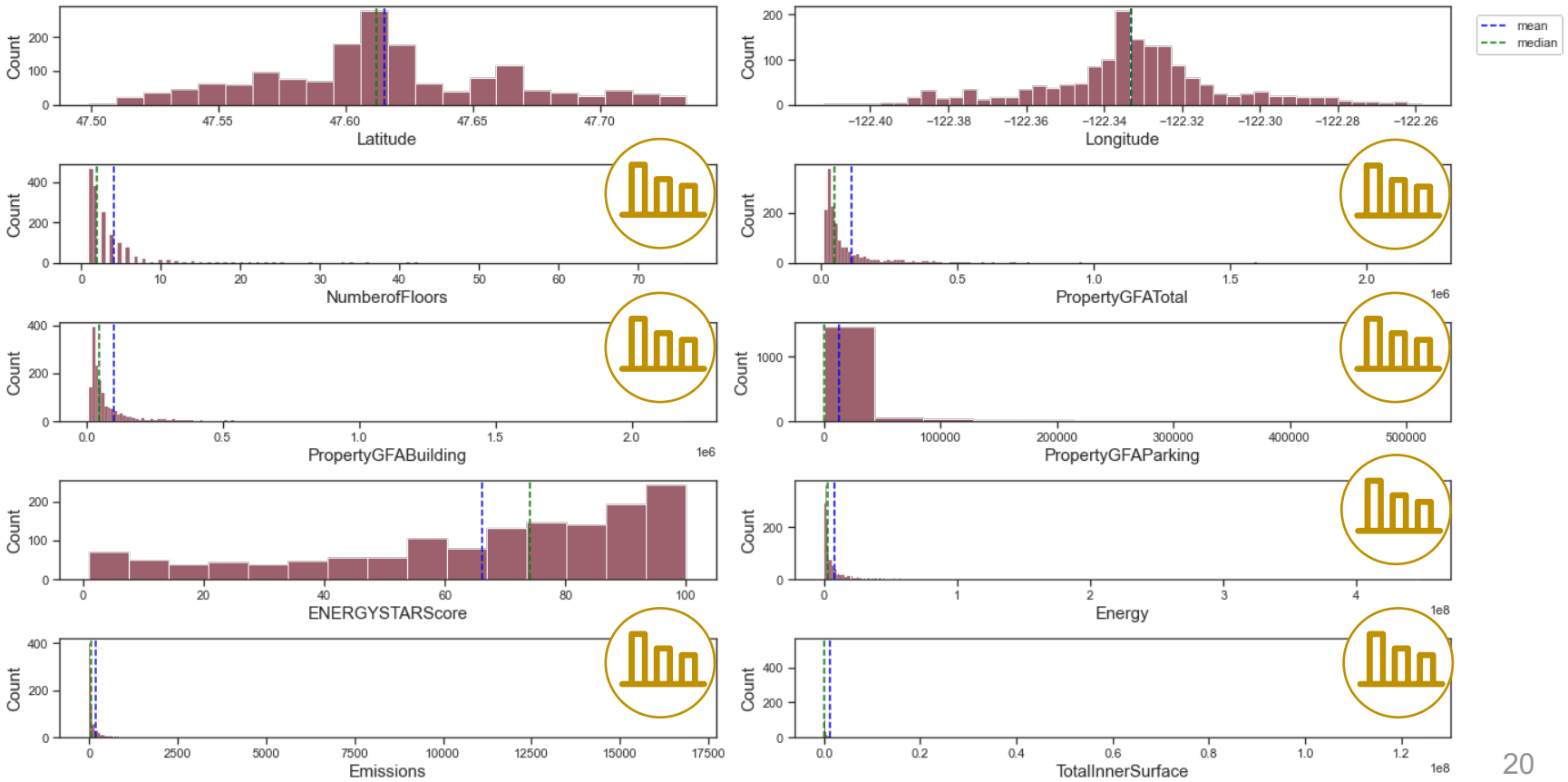
EXPLORATORY ANALYSIS

Numerical features



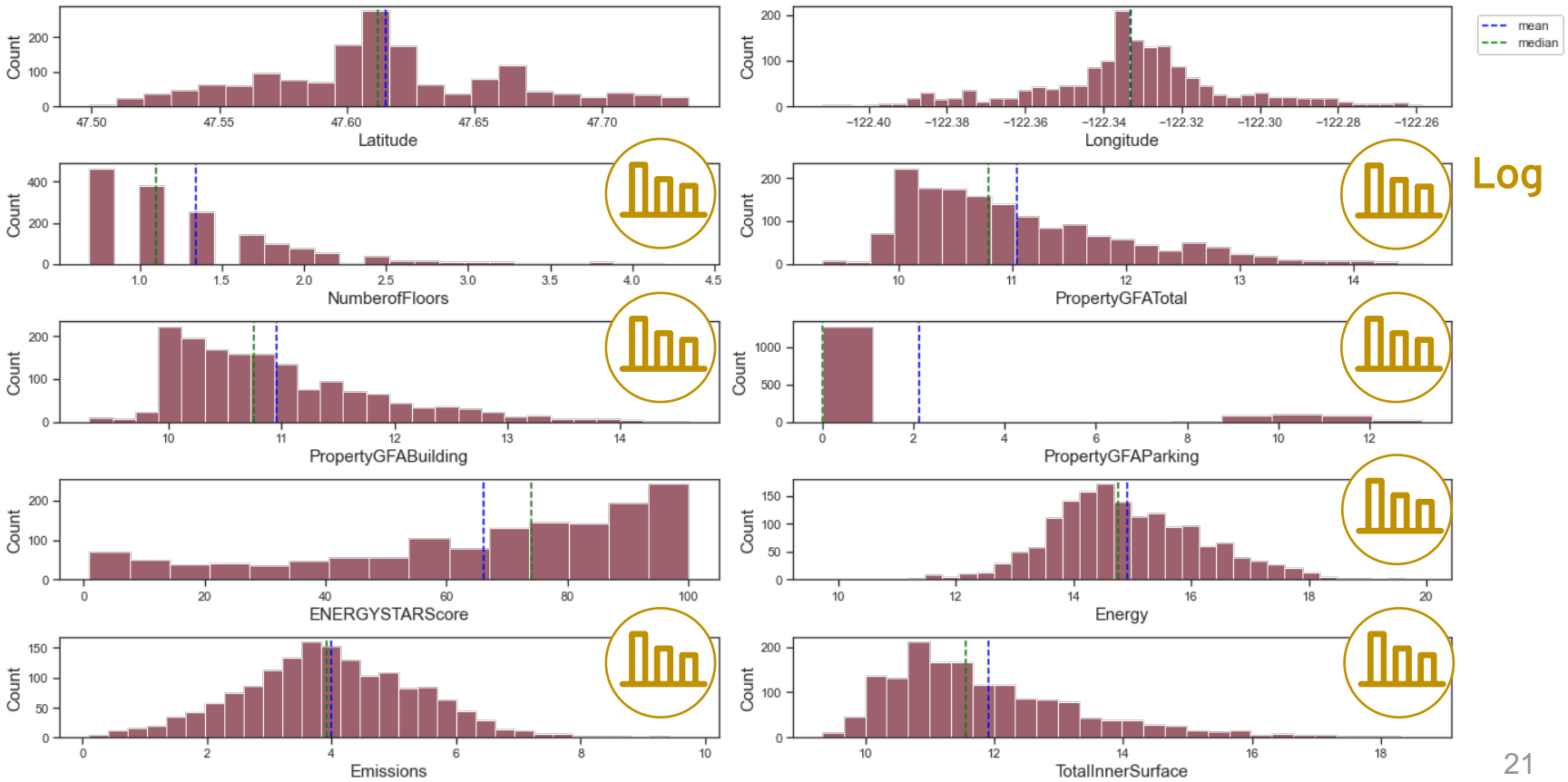
EXPLORATORY ANALYSIS

Numerical variables



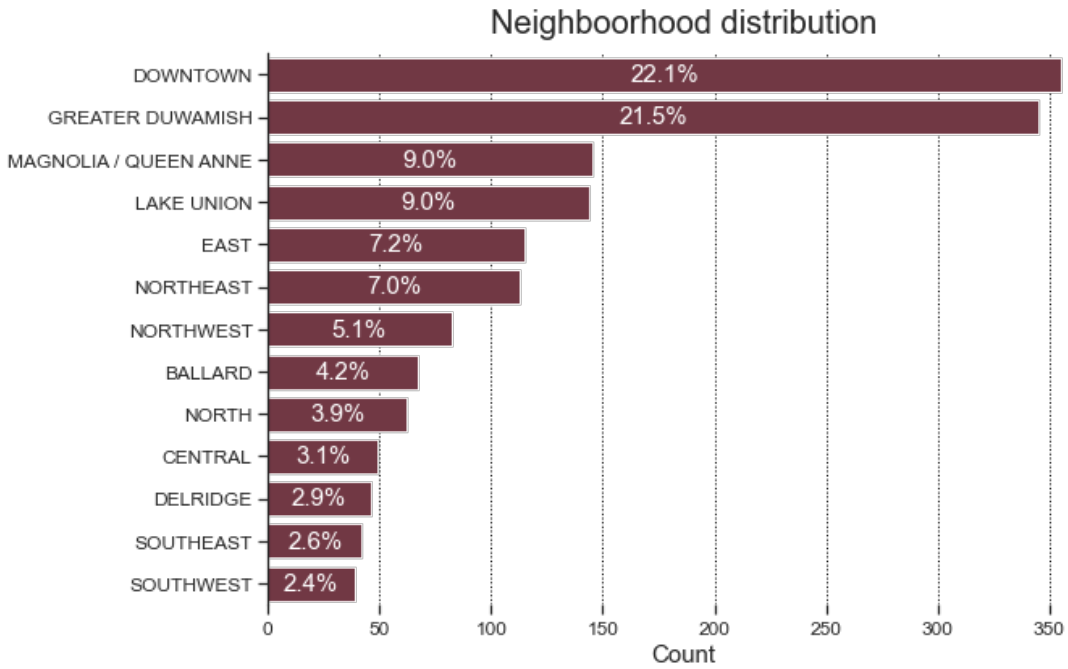
EXPLORATORY ANALYSIS

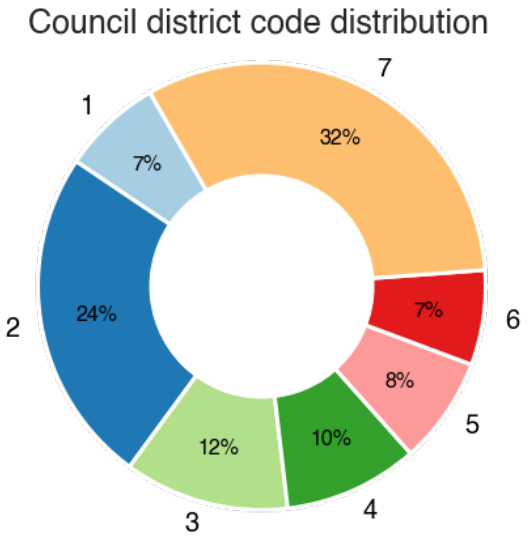
Numerical variables



EXPLORATORY ANALYSIS

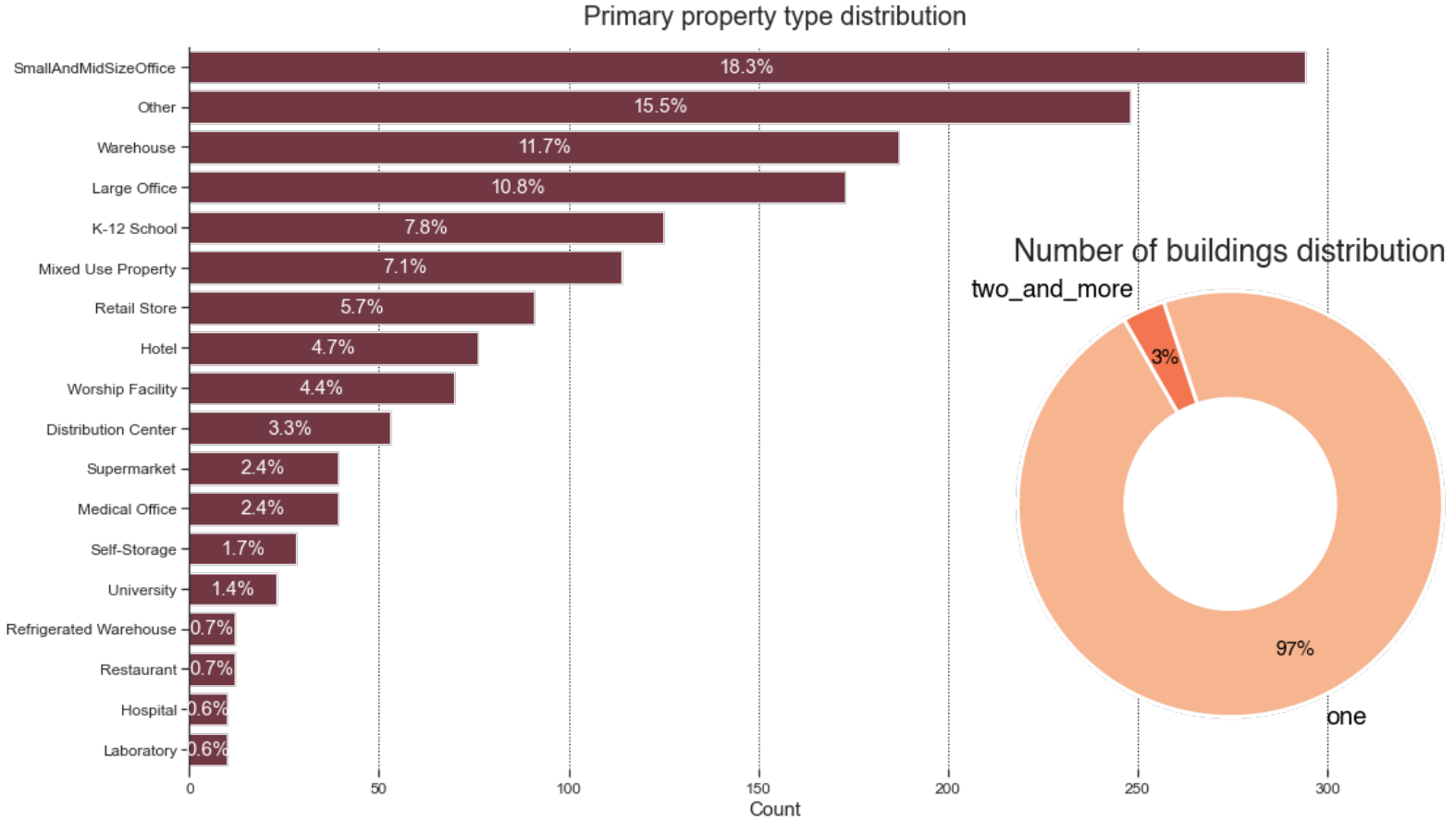
Categorical features





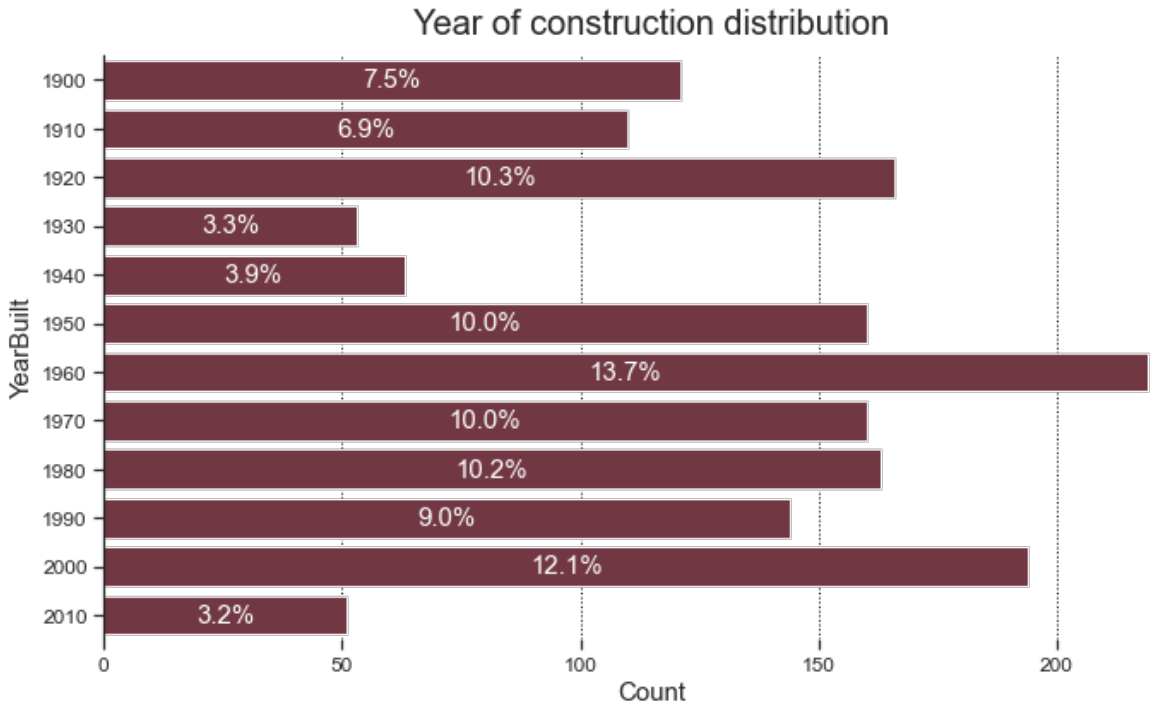
EXPLORATORY ANALYSIS

Categorical features



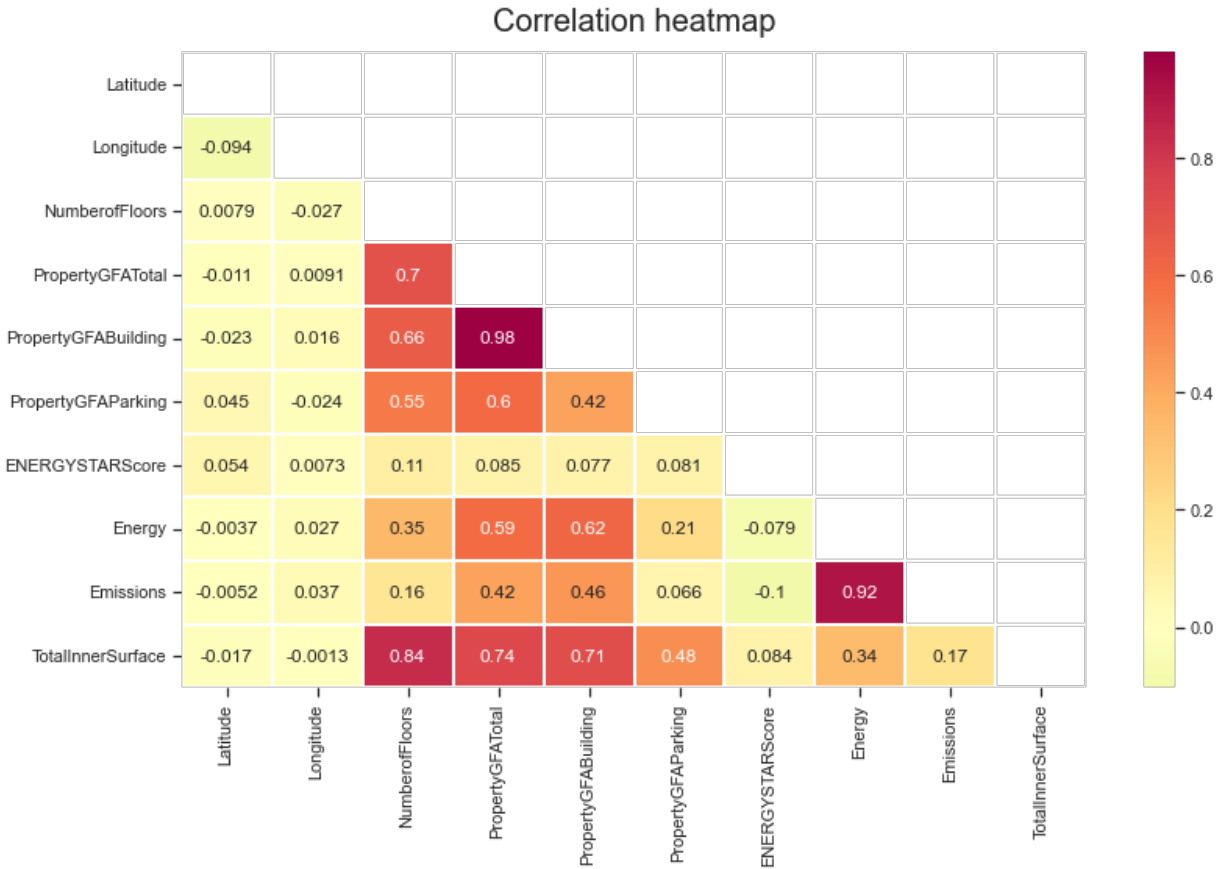
EXPLORATORY ANALYSIS

Categorical features



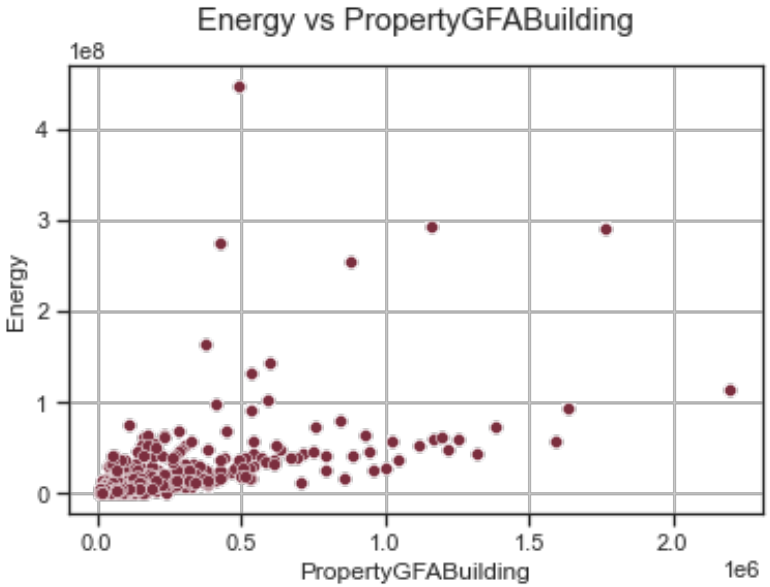
EXPLORATORY ANALYSIS

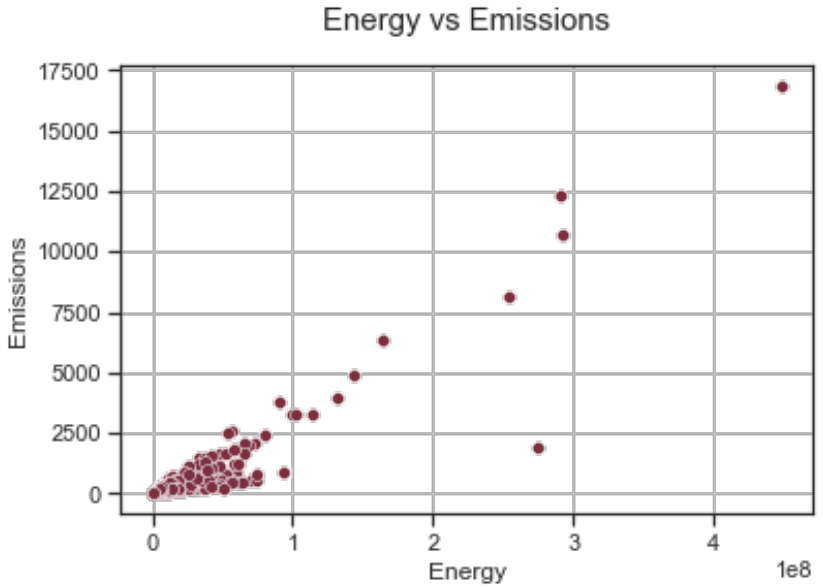
Multivariate analysis



EXPLORATORY ANALYSIS

Multivariate analysis

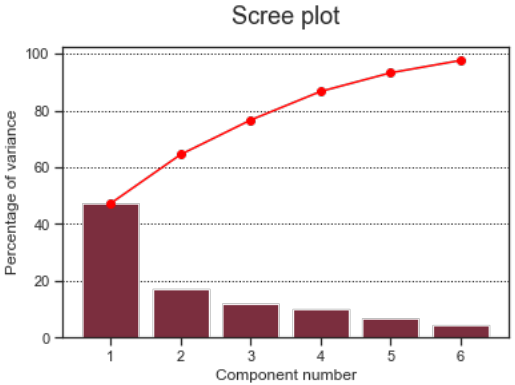


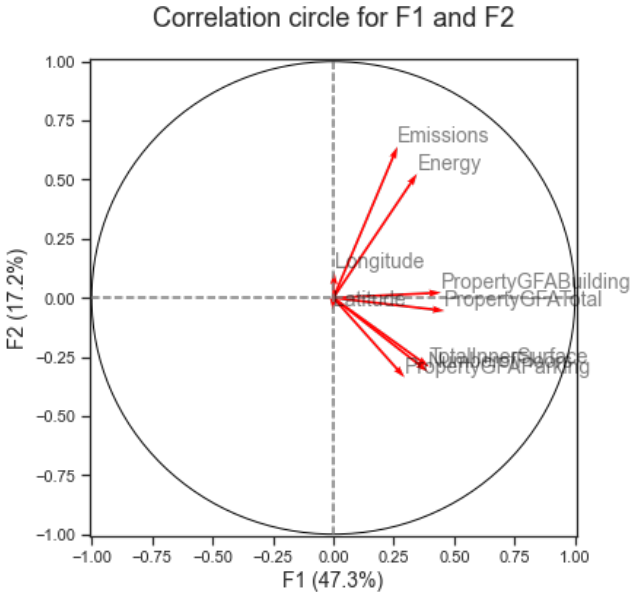


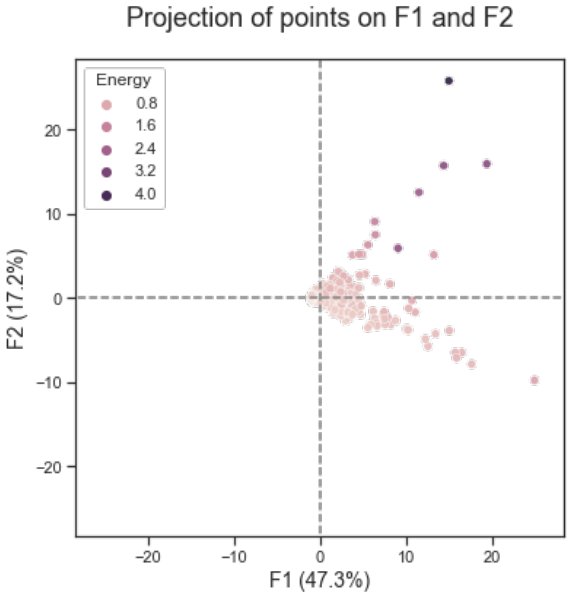
EXPLORATORY ANALYSIS

Multivariate analysis

Principal Component Analysis

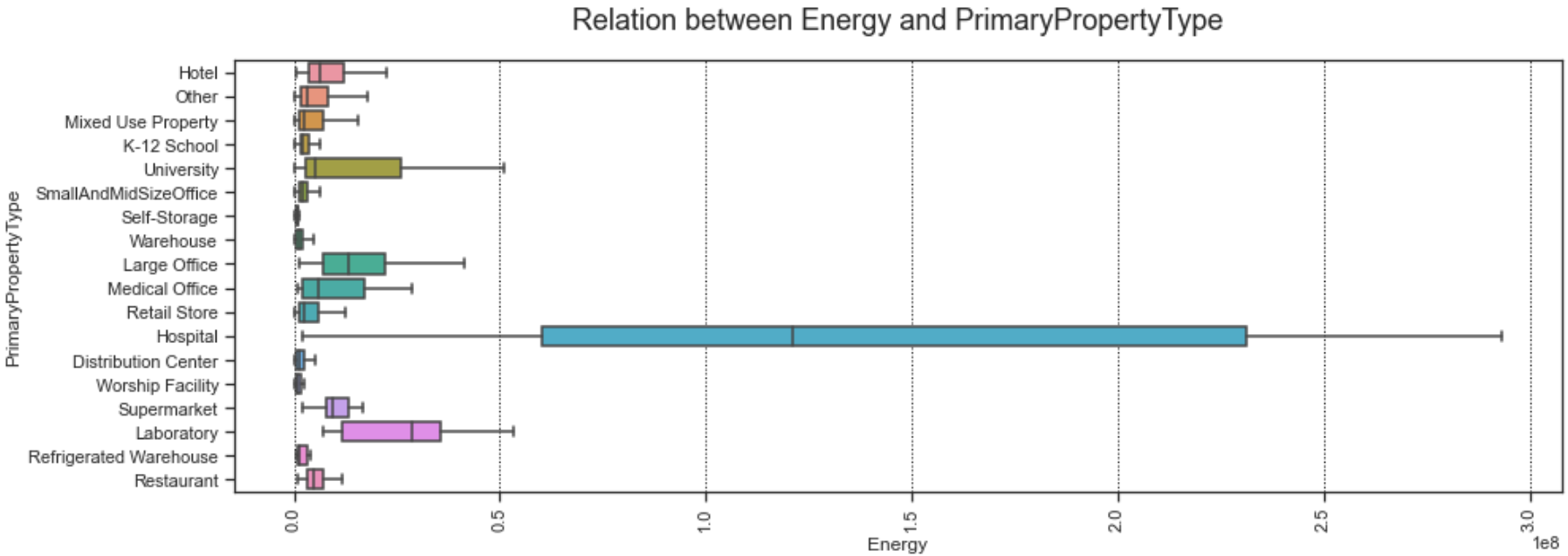






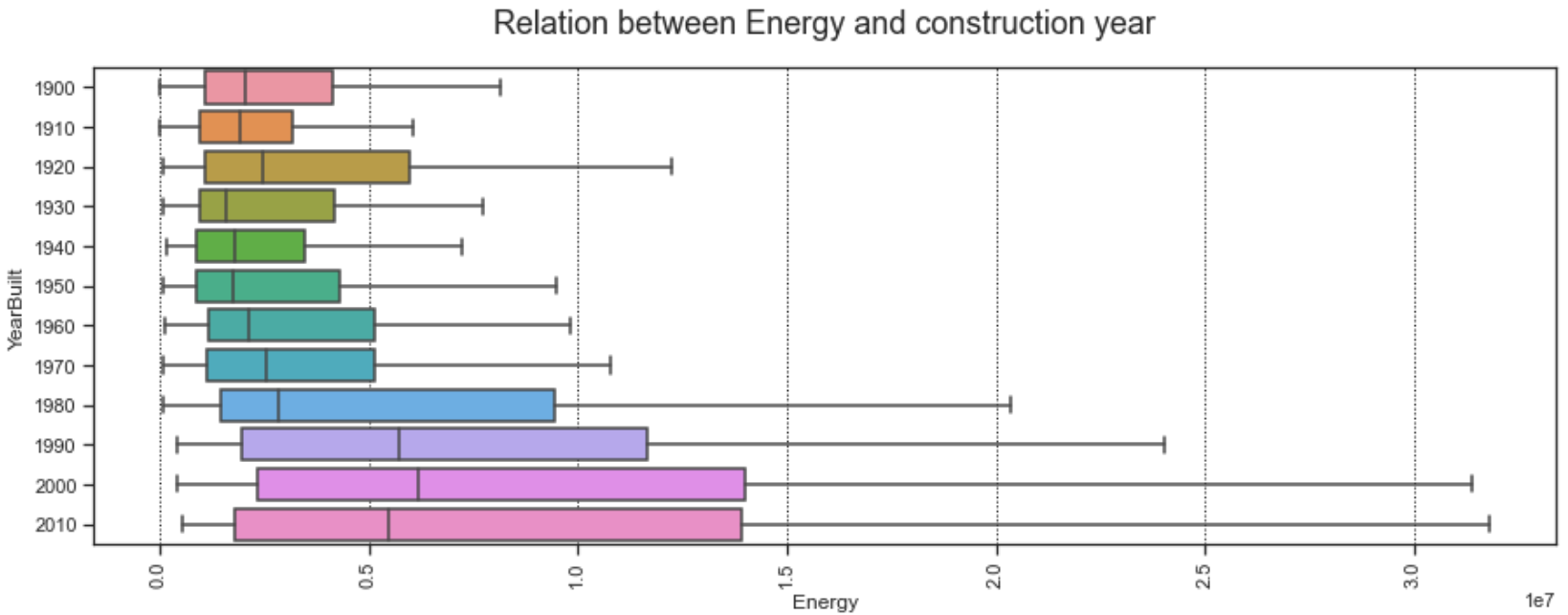
EXPLORATORY ANALYSIS

Multivariate analysis



EXPLORATORY ANALYSIS

Multivariate analysis



EXPLORATORY ANALYSIS

Multivariate analysis

ANOVA $\eta^2 = \frac{ESS}{TSS}$

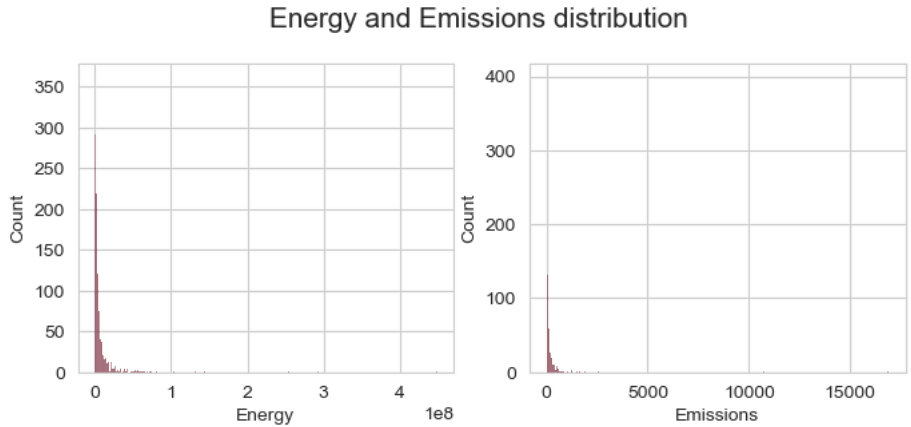
Numerical_features	PrimaryPropertyType	CouncilDistrictCode	Neighborhood	YearBuilt	NumberofBuildings
Latitude	0.113138	0.883750	0.909015	0.020111	0.000429
Longitude	0.029328	0.501332	0.789693	0.004865	0.002595
NumberofFloors	0.335104	0.142691	0.233428	0.063139	0.001189
PropertyGFATotal	0.279441	0.055096	0.074704	0.067121	0.012621
PropertyGFABuilding	0.254042	0.045427	0.065854	0.045427	0.018079
PropertyGFAParking	0.194210	0.049316	0.065302	0.108966	0.000942
Energy	0.299168	0.022276	0.025728	0.030201	0.044216
Emissions	0.323389	0.016503	0.018439	0.017813	0.049964
TotalInnerSurface	0.131981	0.034819	0.068777	0.035087	0.000015

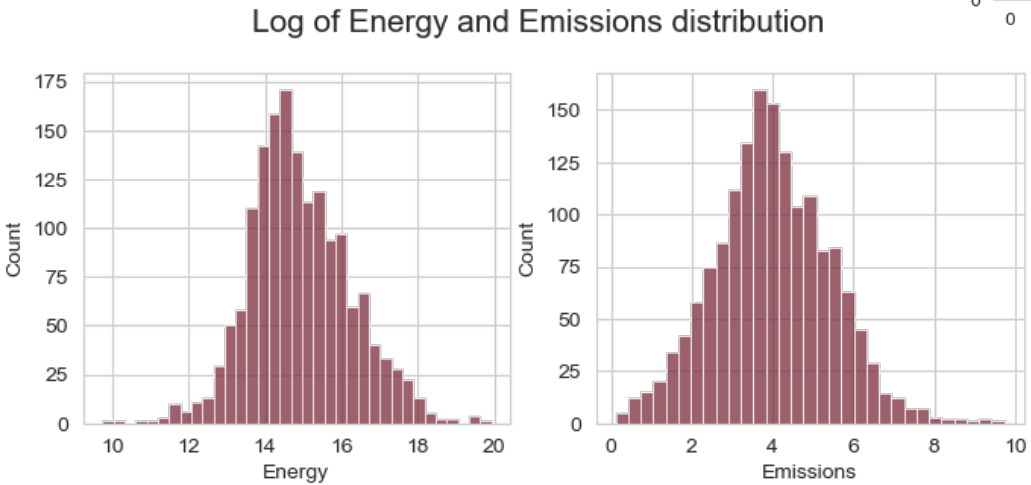
03 MACHINE LEARNING MODELS

MACHINE LEARNING MODELS

Feature engineering

Log transformation of targets

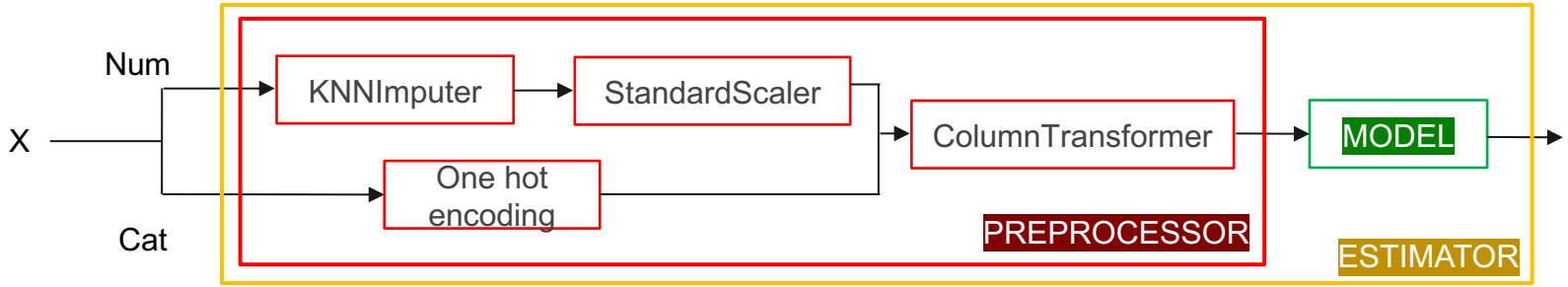


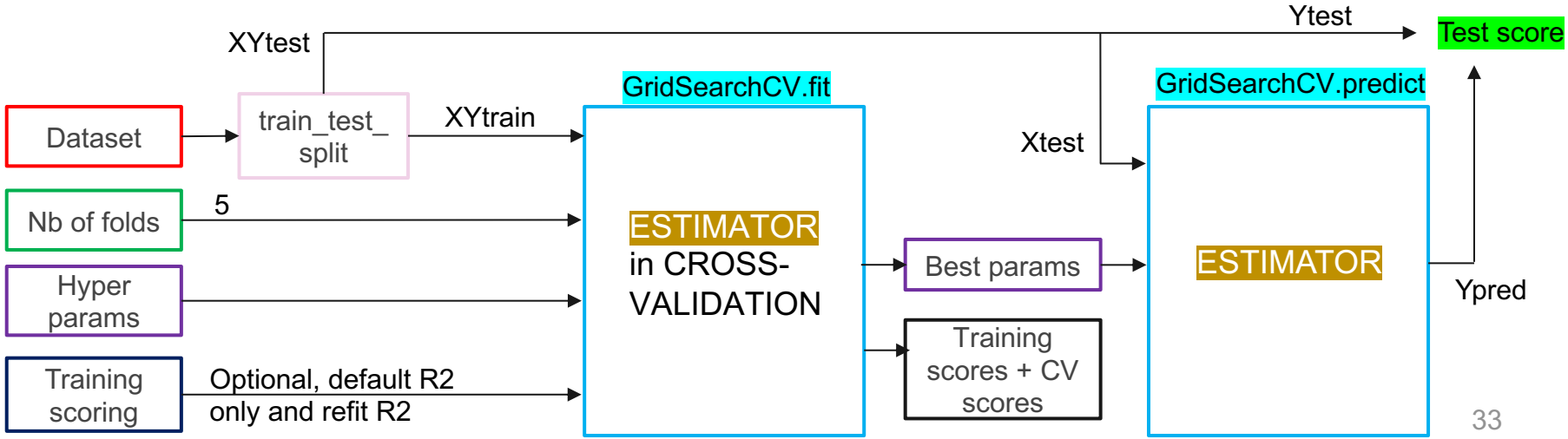


Log

MACHINE LEARNING MODELS

Regression pipeline





MACHINE LEARNING MODELS

Tested models

- Reference simple model
  - Dummy regressor
- Linear models
  - Linear regression
  - Linear regression with Ridge regularization
  - Linear regression with Lasso regularization
- Support Vector Machine (SVM)
  - Support vector regression
- Ensemble learning
  - Random forest
  - Gradient boosting

MACHINE LEARNING MODELS

Scoring metrics

- R^2 – coefficient of determination

$$R^2 = 1 - \frac{unexplained\ variation}{total\ variation} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

  - Between 0 and 1
  - The larger, the more variance explained by the regression model
  - Is the default sklearn criterion for regression
  - Also used for cross-validation in GridSearchCV by default

- MAE – Mean Absolute error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

  - Individual differences are weighted equally on the average (linear score)
  - Between 0 and ∞

- RMSE – Root Mean Square Error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

  - Gives a relatively high weight to large errors
  - Always >= MAE, greater difference between them = greater variance in the individual errors

MACHINE LEARNING MODELS - ENERGY

Results for Energy consumption

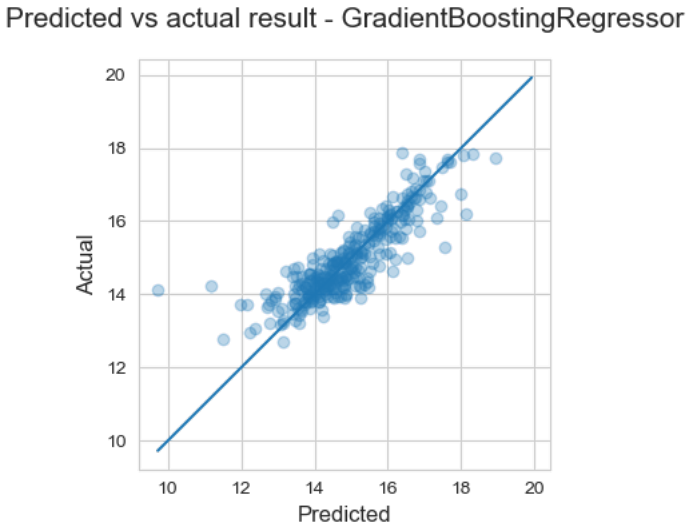
	Models	R2	MAE	RMSE
6	Gradient boosting	0.723084	0.451320	0.668488
5	Random forest	0.703442	0.463639	0.691790
4	SVM	0.694990	0.480456	0.701579
3	Lasso regression	0.591314	0.588084	0.812109
2	Ridge regression	0.588792	0.590862	0.814612
1	Linear regression	0.588451	0.591793	0.814949
0	Dummy regressor	-0.000030	0.996292	1.270358

MACHINE LEARNING MODELS - ENERGY

Gradient boosting hyperparameters optimization

- Parameters are adjusted to increase the score.
- Highest score obtained is R^2 = 0.742

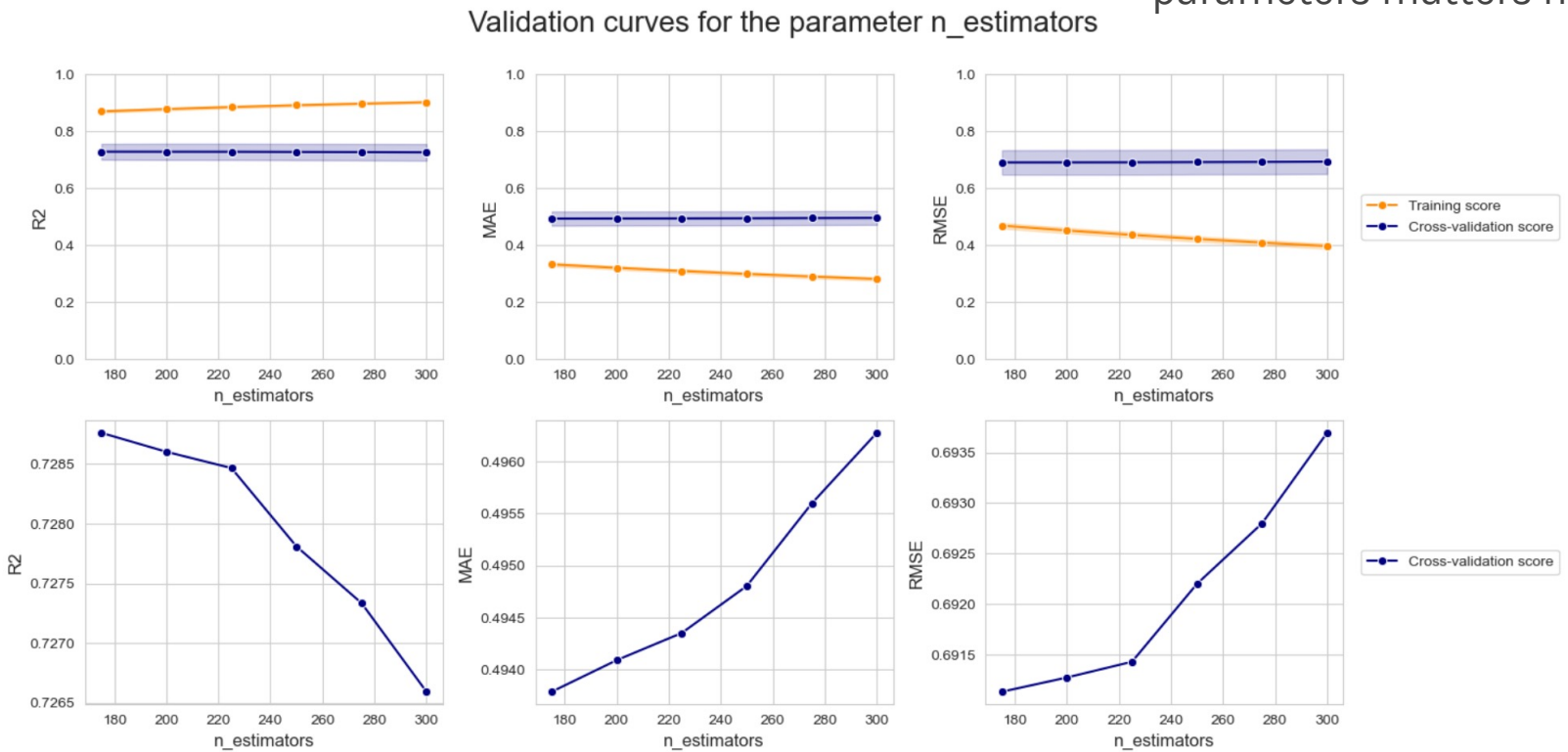
```
CROSS VALIDATION RESULTS
        - Best parameters:
                - model__max_depth: 3
                - model__max_features: log2
                - model__min_samples_leaf: 3
                - model__min_samples_split: 6
                - model__n_estimators: 300
        - Metrics:
                - R2: 0.744 (+/-0.028)
                - MAE: 0.481 (+/-0.023)
                - RMSE: 0.671 (+/-0.042)
TEST RESULTS
        - Metrics:
                - R2: 0.742
                - MAE: 0.440
                - RMSE: 0.645
```



MACHINE LEARNING MODELS - ENERGY

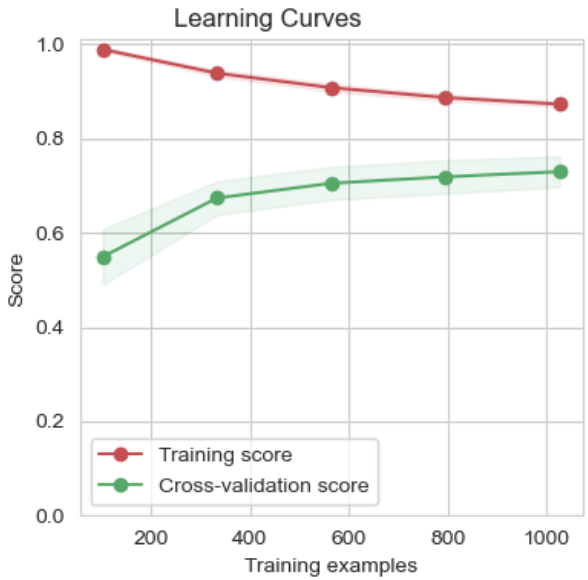
Gradient boosting cross-validation curves

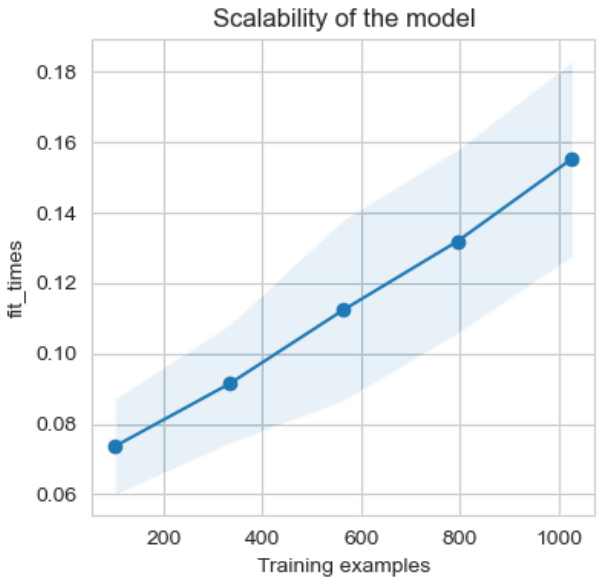
- Observation: combination of parameters matters more

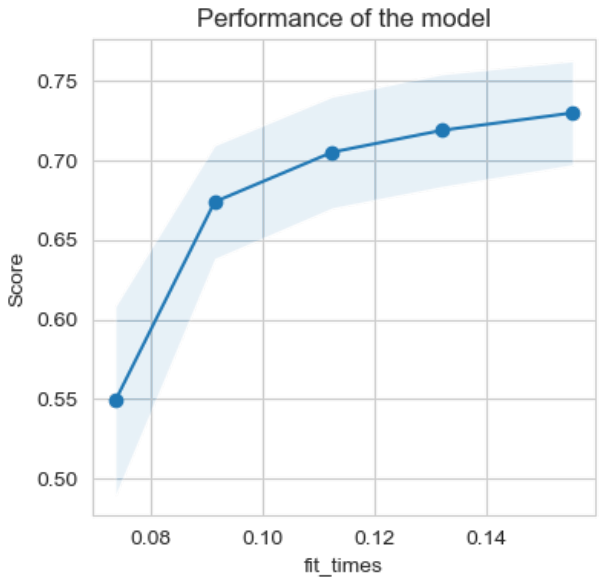


MACHINE LEARNING MODELS - ENERGY

Gradient boosting learning curves



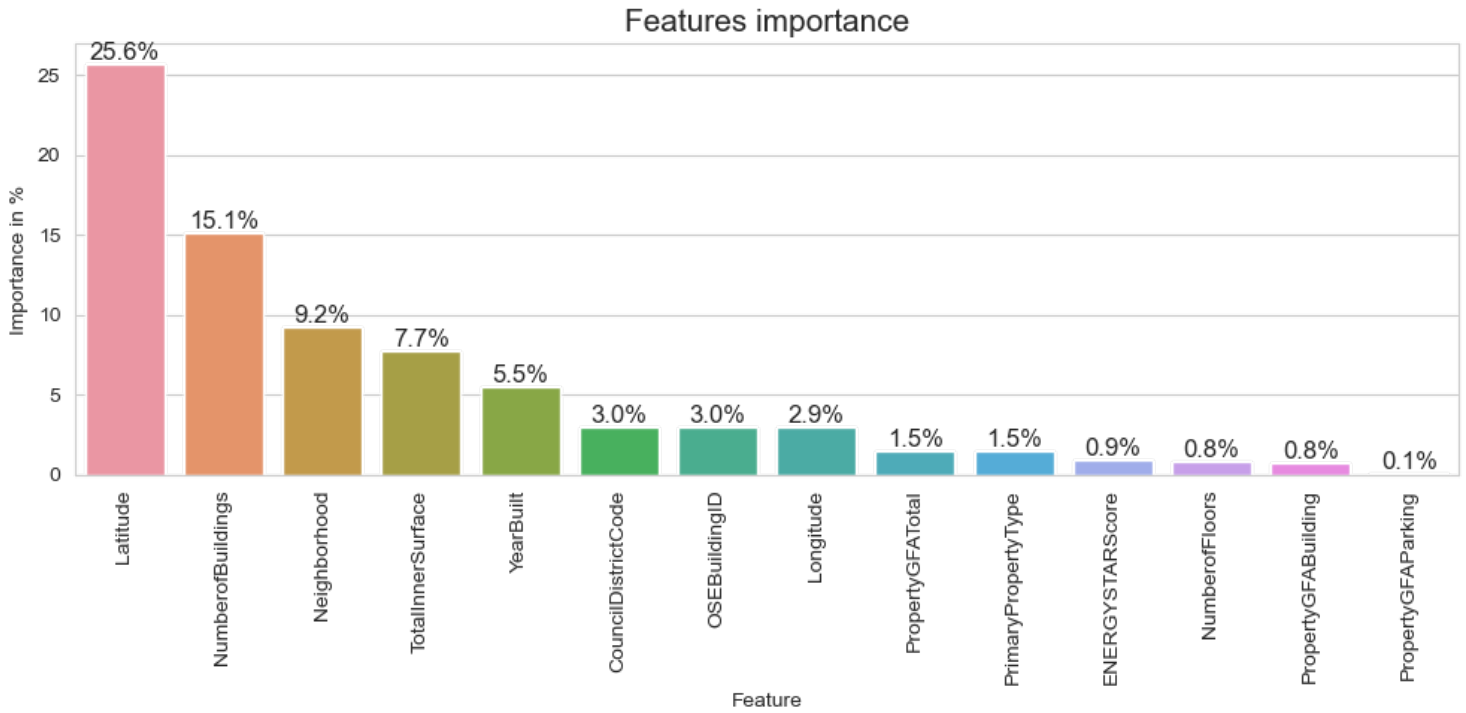




- Observation: there is a margin for improvement if more samples are added

MACHINE LEARNING MODELS - ENERGY

Energy star score influence



MACHINE LEARNING MODELS - EMISSIONS

Results for CO2 emissions

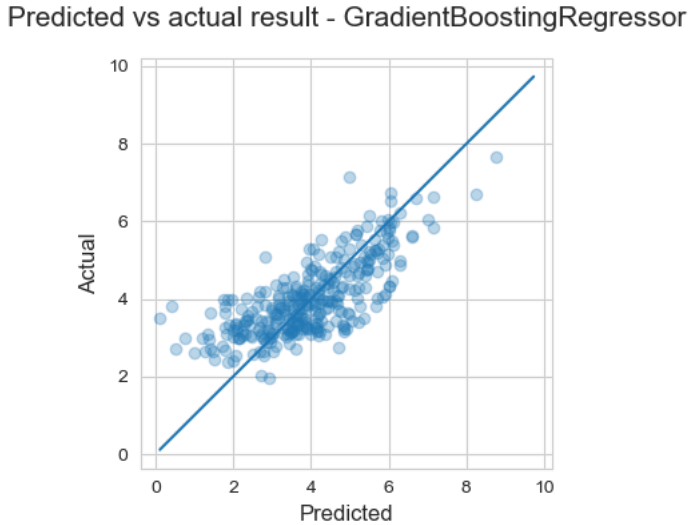
	Models	R2	MAE	RMSE
6	Gradient boosting	0.544780	0.723260	0.926146
5	Random forest	0.541935	0.722915	0.929036
4	SVM	0.515755	0.711980	0.955215
3	Lasso regression	0.435798	0.799182	1.031065
2	Ridge regression	0.433940	0.802601	1.032761
1	Linear regression	0.431366	0.801473	1.035107
0	Dummy regressor	-0.002113	1.089652	1.374128

MACHINE LEARNING MODELS - EMISSIONS

Gradient boosting hyperparameters optimization

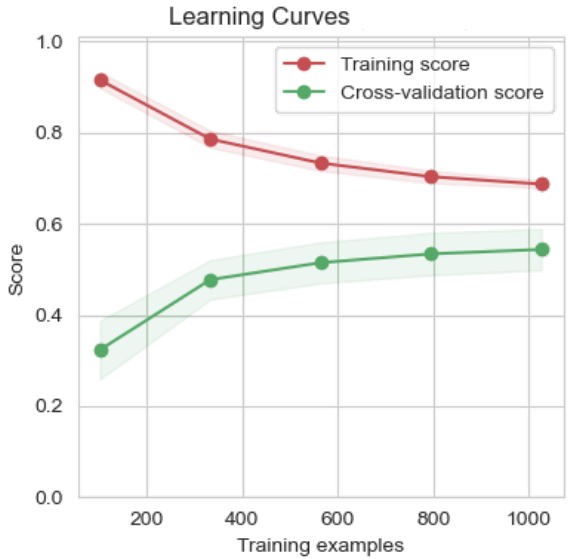
- Parameters are adjusted to increase the score.
- Highest score obtained is R^2 = 0.552

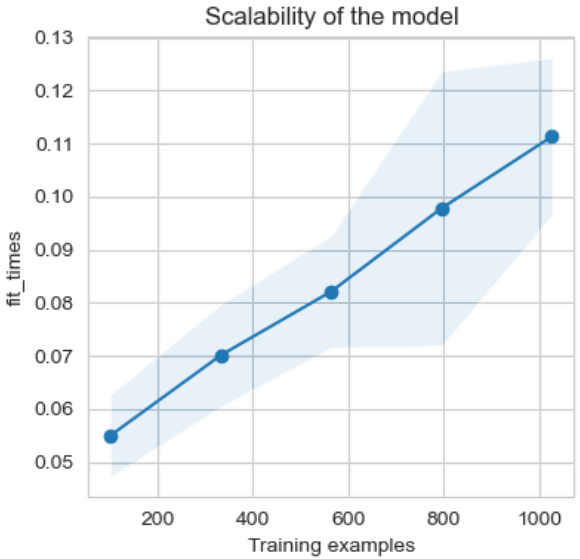
```
CROSS VALIDATION RESULTS
        - Best parameters:
                - model__max_depth: 2
                - model__max_features: sqrt
                - model__min_samples_leaf: 4
                - model__min_samples_split: 6
                - model__n_estimators: 250
        - Metrics:
                - R2: 0.569 (+/-0.032)
                - MAE: 0.755 (+/-0.030)
                - RMSE: 0.953 (+/-0.036)
TEST RESULTS
        - Metrics:
                - R2: 0.552
                - MAE: 0.716
                - RMSE: 0.919
```

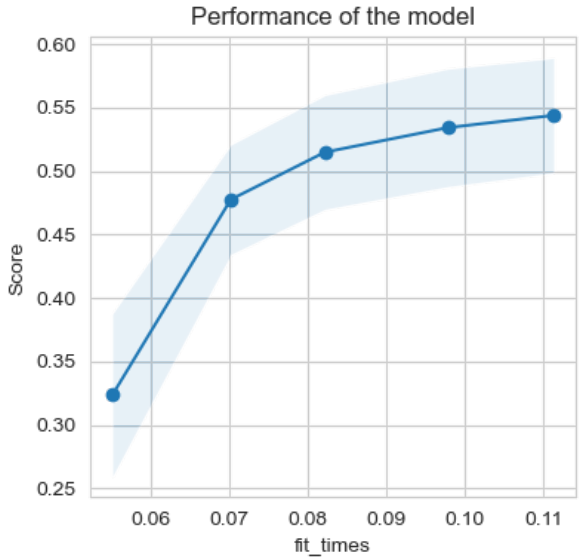


MACHINE LEARNING MODELS - EMISSIONS

XGBoost learning curves



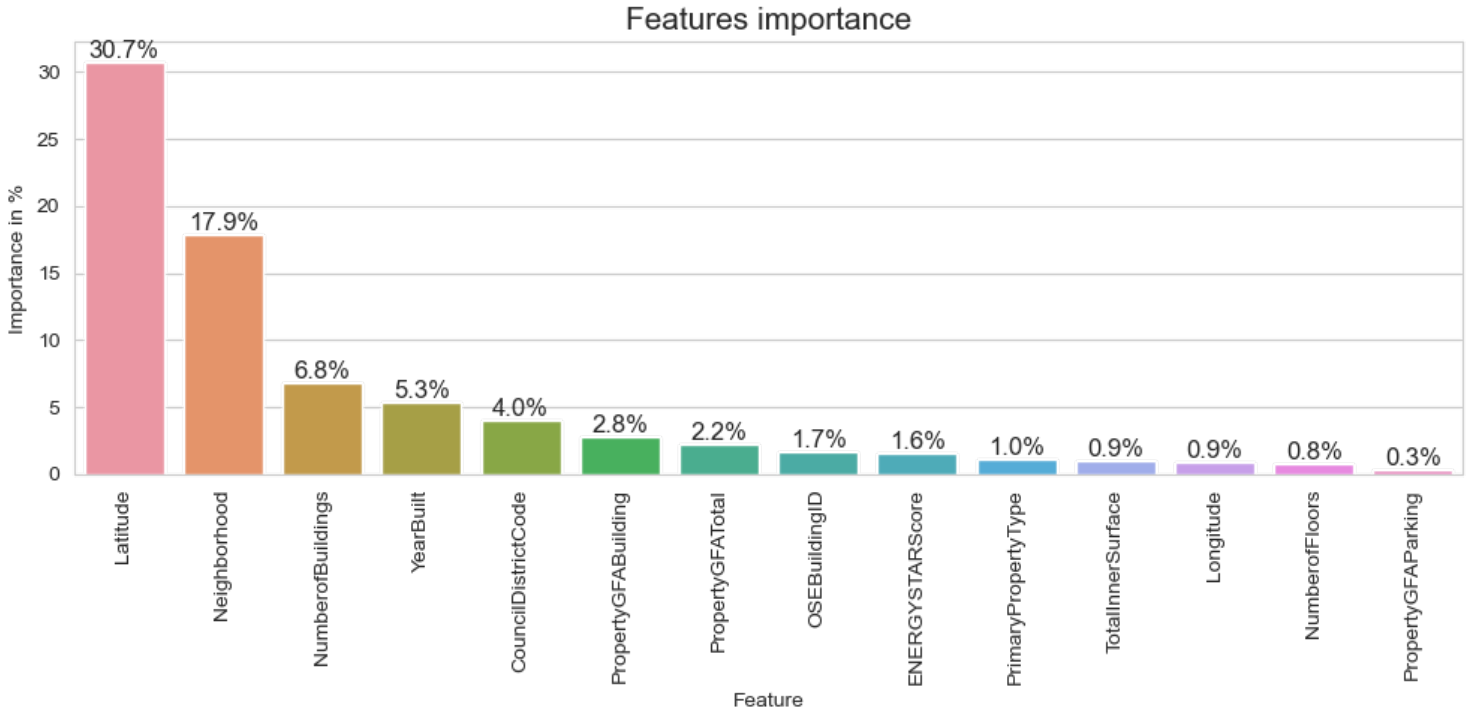




- Observation: there is a margin for improvement if more samples are added

MACHINE LEARNING MODELS - EMISSIONS

Energy star score influence



04

CONCLUSION

CONCLUSION

- Initial data: for 2 years: 2015 and 2016, (3340, 47) and (3376, 46) resp.
- Target features
  - Energy consumption -> ‘SiteEnergyUse(kBtu)’
  - CO2 emissions -> ‘TotalGHGEmissions’
- Correlations found between targets and property type, location
- Best regression model: Gradient boosting, for both targets
  - R^2 = 0.742 for Energy consumption prediction
  - R^2 = 0.552 for CO2 emissions prediction
- Model improvement:
  - Energy star score has insignificant influence on results, is not worth measuring
  - A larger sample would allow error reduction in selected regression models

THANK YOU

CentraleSupélec

Victor Benard