Predicting the energy consumption of buildings

FEBRUARY 2022

NAJWA MOULINE

Presentation Outline

1. Objectives
2. Dataset Preparation
3. Modeling process
4. Final model overview

Objectives

Context

- Seattle to become carbon neutral by 2050
- Consumption data available for buildings in the city of Seattle for the years 2015 and 2016

Business Problem

- Significant cost of obtaining statements / tedious to collect

Mission

- Predict CO_2 emissions and total energy consumption without annual readings and from existing data
- Assess the value of the ENERGY STAR Score
- Set up a reusable prediction model

Objectives

Approach

Interpretation

- 2 final models to identify – regression models:
  - Model to predict CO2 emissions
  - Model to predict total energy consumption
- Features:
  - intrinsic to buildings (easy to obtain)
  - EnergyStarScore feature: Testing without and with

Methodology

1. Carry out an AED to identify and observe targets and features

Cleaning	Feature Engineering	Exploratory Analysis

2. Evaluate different prediction models

Pre-processing	Modeling	Optimisation

Dataset Preparation

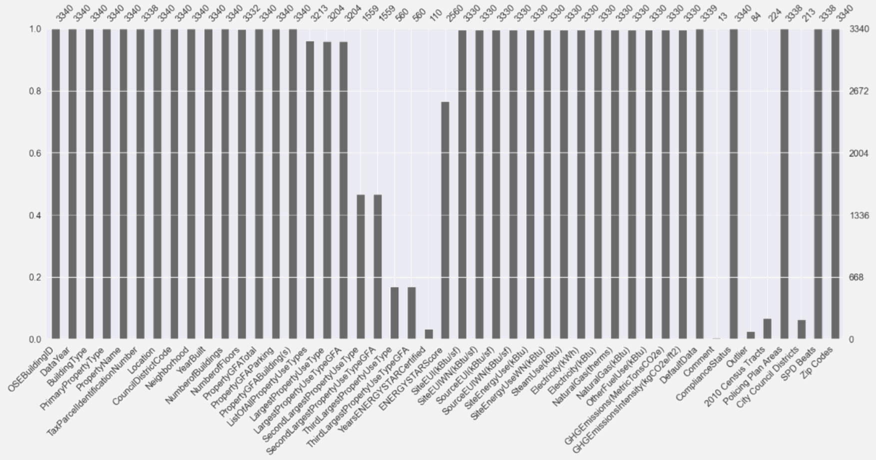
Dataset

Dataset

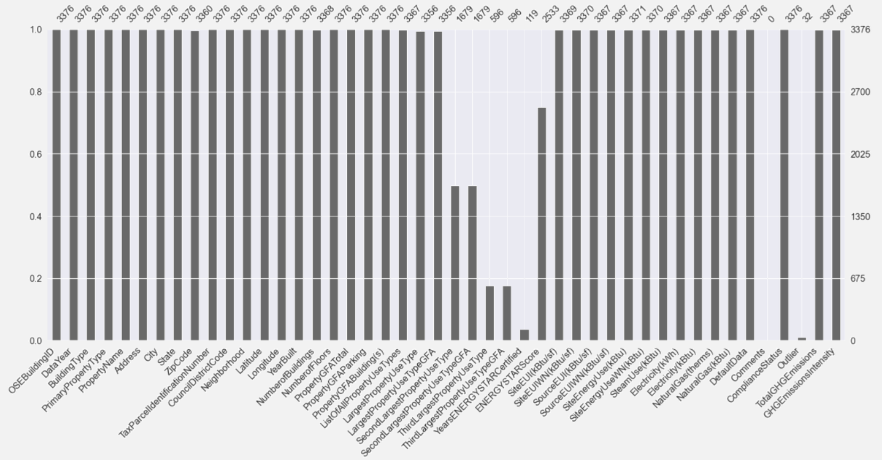
Observations

- A dataset: 2 files representing the year 2015 and the year 2016
- 3000 rows (buildings) and more than 46 columns (variables).
- Buildings: identified by a Unique ID
- Variables: 55% qualitative and 45% quantitative, some different variables
- A well-filled dataset – 14% NaN

Variables fill rate - 2015



Variables fill rate - 2016



Dataset Preparation

Methodology

Process followed in 3 steps

Vars : 47/46
Obs. : 3340/3376
NaN : 17%/13%

Cleaning

- Aggregating of the 2 files
- Selecting observations
- Selecting variables
- Cleaning

Vars : 22
Obs : 1684
NaN : 1.4%

Feature Engineering

- Creating variables
- Combining modalities
- Transforming variables

Vars : 37
Obs : 1684
NaN : 1%

Exploratory Analysis

- Distribution of variables
- Correlation between variables
- Selecting variables

Vars : 28
Obs : 1684
NaN : 1%

Dataset Preparation

Cleaning

Process followed in 4 steps

Aggregation of the 2 files
- Homogenisation of variable names
- Homogenisation of modalities

Selection of observations
- Non-duplicate buildings
- Buildings in Seattle
- Non-Residential Buildings

Selecting variables
- Variables filled in
- Non-redundant variables
- Relevant variables

Cleaning
- Handling missing values
- Removing outliers

Dataset Preparation

Feature Engineering

Process followed in 3 steps

Creating variables

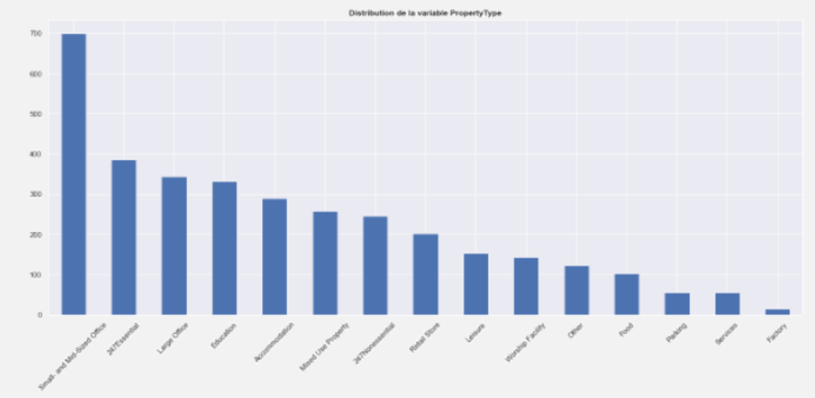
- Age = DataYear - YearBuilt
- FloorGFA = GFAB / (NbFloors + 1)
- ParkingRate = GFAParking / GFAT

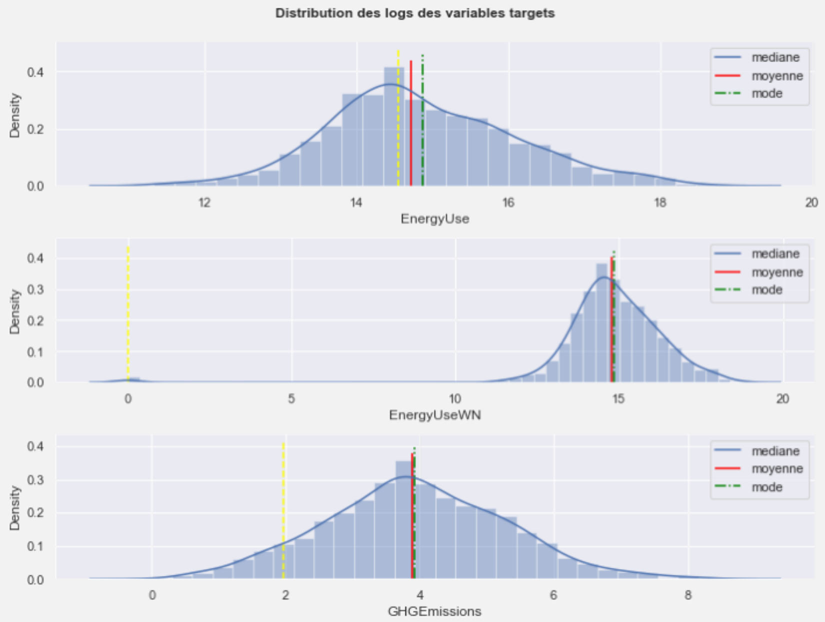
Combining modalities

- PropertyType - 15 modalities

Transforming variables

- PropertyType – onehotencoder
- Variables distributed in log/sqrt

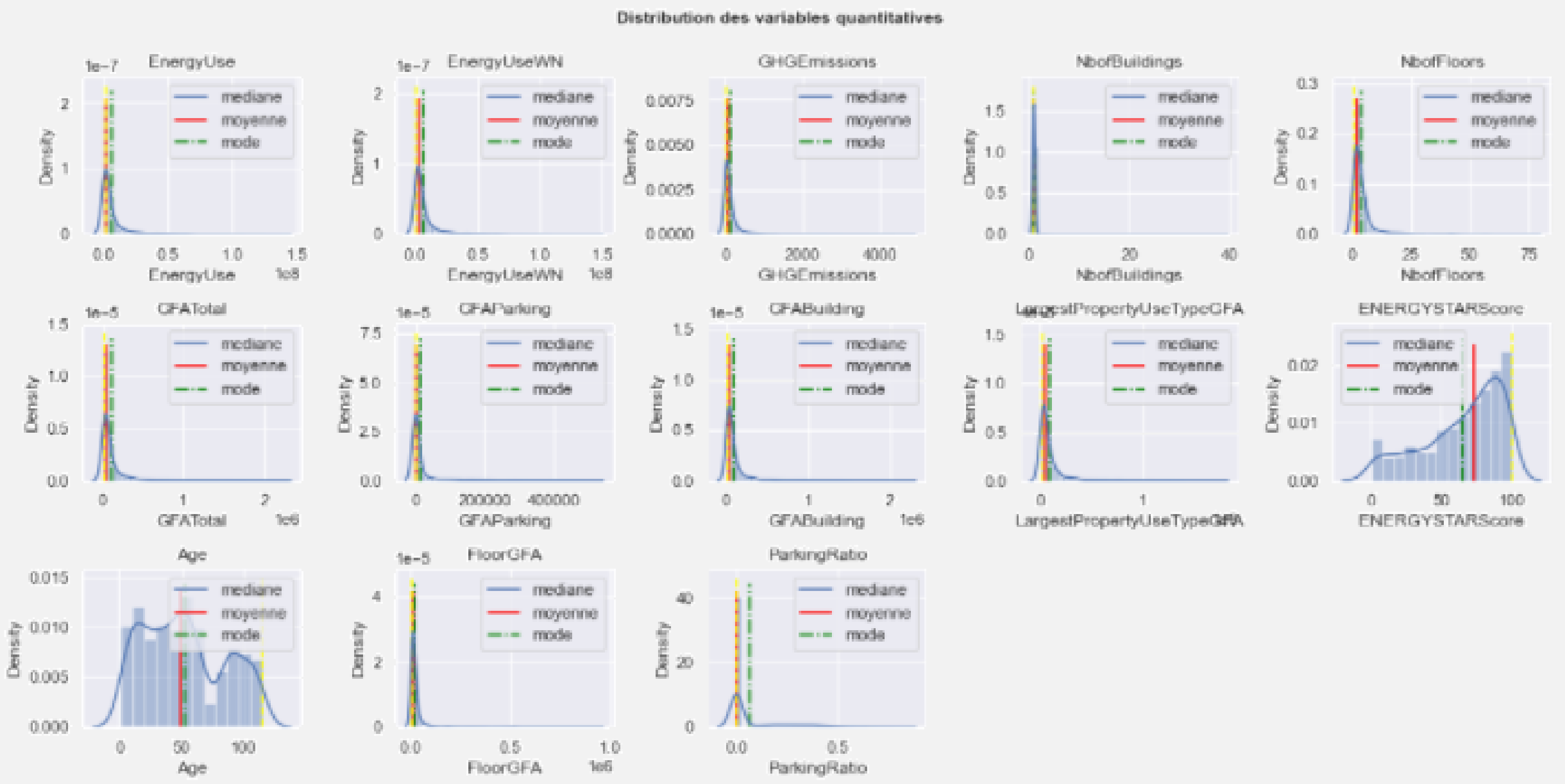




Dataset Preparation

Exploratory Analysis

Distribution of variables



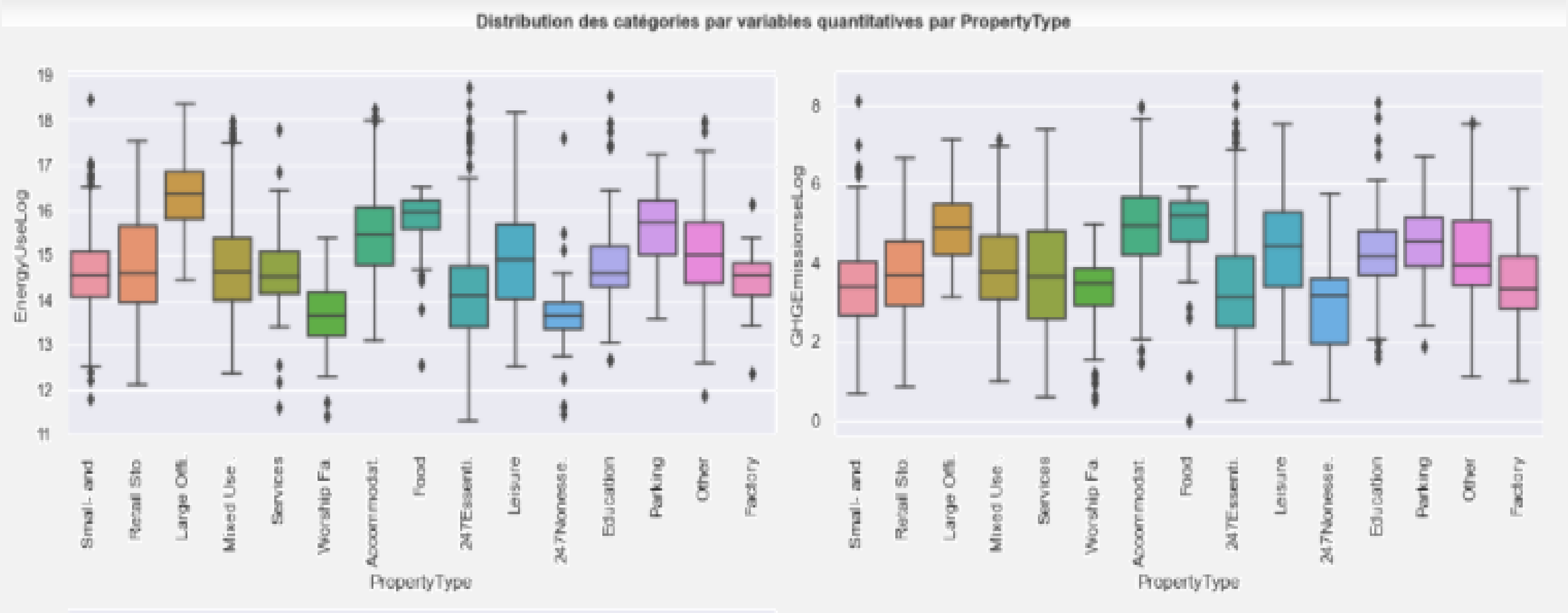
Observations

- Many distributions with a strong skewness on the right, hence the consideration of the transition to the log
- Many outliers

Dataset Preparation

Exploratory Analysis

ANOVA – Distribution of targets based on PropertyTypeUse



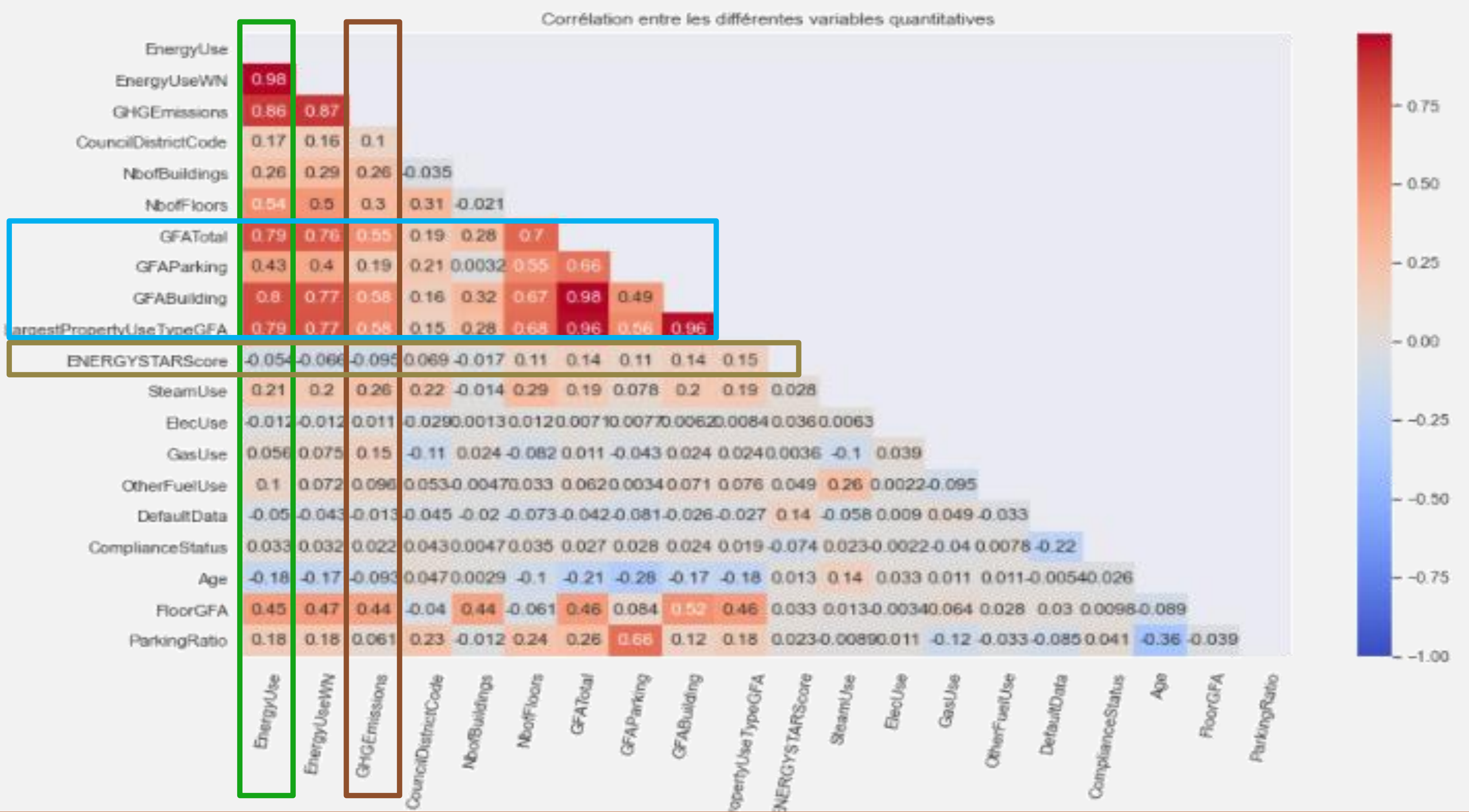
Observations

- Variables to predict linked to the PropertyType.

Dataset Preparation

Exploratory Analysis

Correlation between quantitative variables



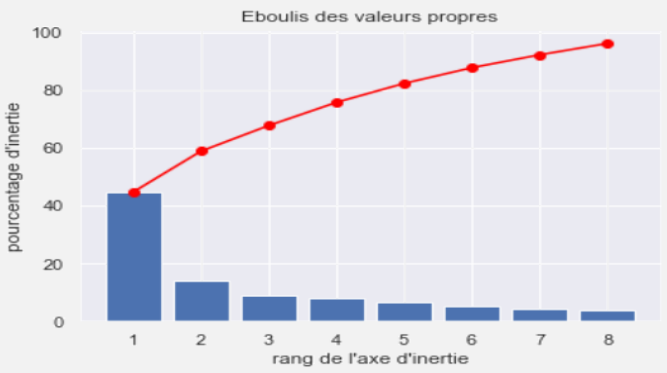
Observations

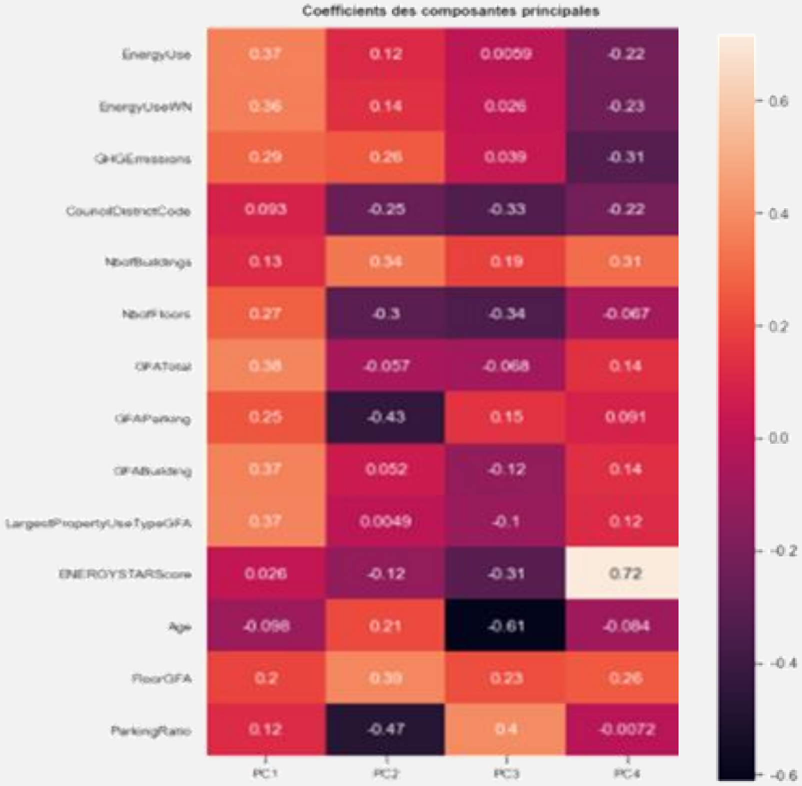
- Strong correlation between variables *EnergyUse* and *GHGEmissions*
- Strong correlation between the variables to be predicted and the *surfaces*
- Strong correlation between *surfaces*
- No correlation between the variables to be predicted and *EnergyStarCode*

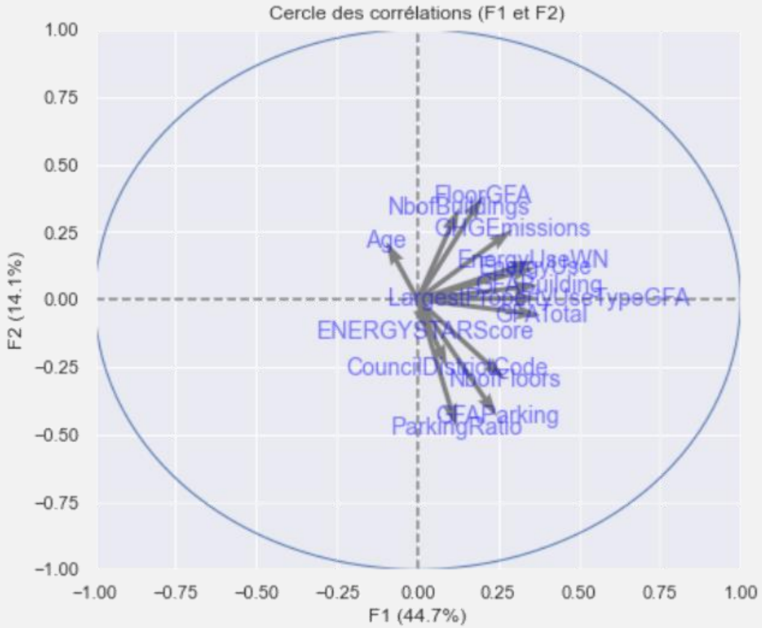
Dataset Preparation

Exploratory Analysis

Principal Component Analysis







Observations

- The first 4 components contain 77% of the variance
- Axis PC1 representing 45% is related to targets and surfaces to be heated mainly

Dataset Preparation

Exploratory Analysis

Selecting variables

Target variables

- EnergyUse
- EnergyUseWN
- GHGEmissions

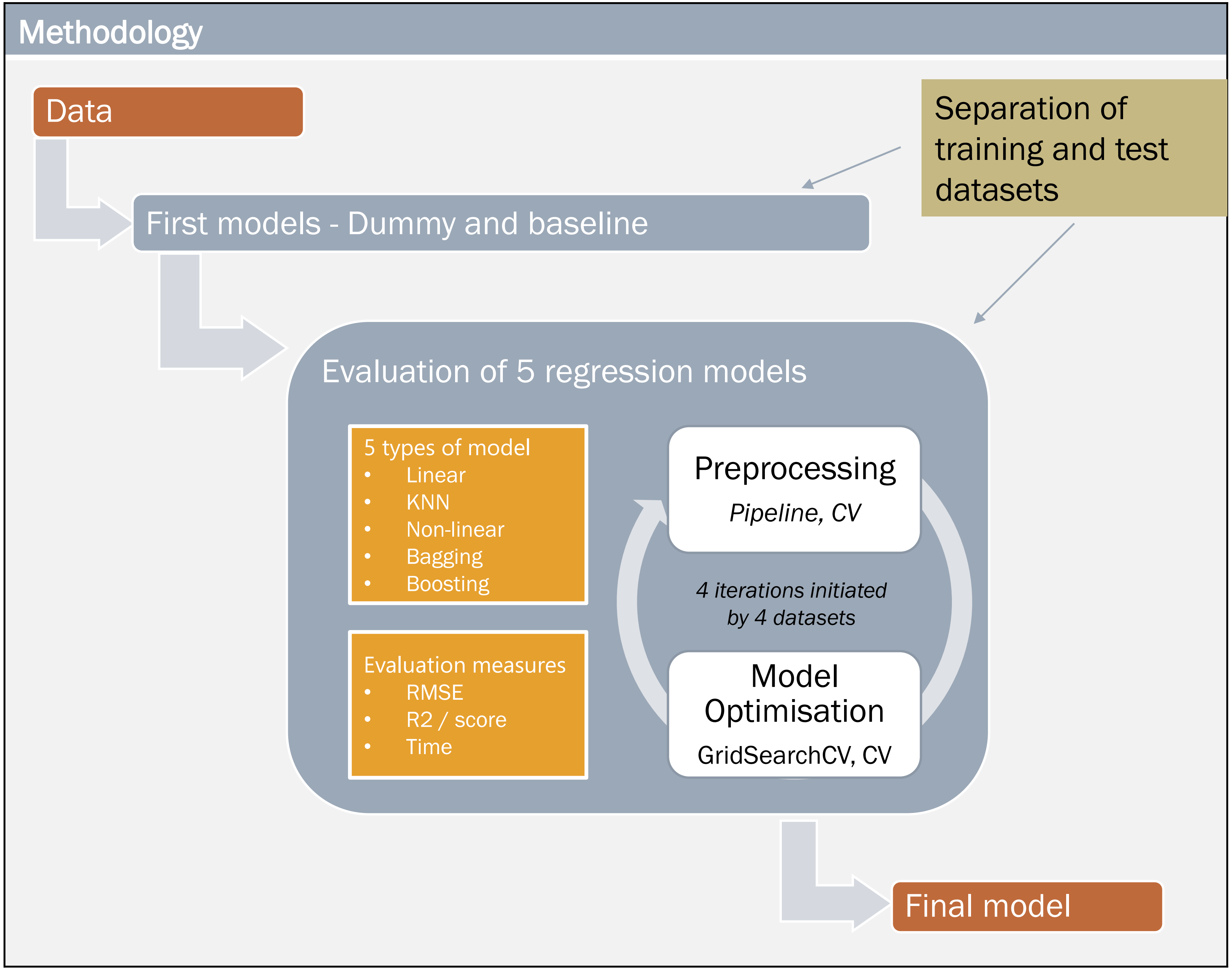
Variables features - quantitatives

Selection of variables a minimum correlated to the variables to be predicted, not correlated to other features.

CouncilDistrictCode	247Essential
NbofBuildings	247Nonessential Accommodation
NbofFloors	Education
GFAParking	Food
GFABuilding	Large Office
ENERGYSTARScore	Leisure
SteamUse	Mixed Use Property
GasUse	Other
Age	Parking
FloorGFA	Retail Store
ParkingRatio	Services
	Small- and Mid-Sized Office Worship Facility'

Modeling process

Methodology



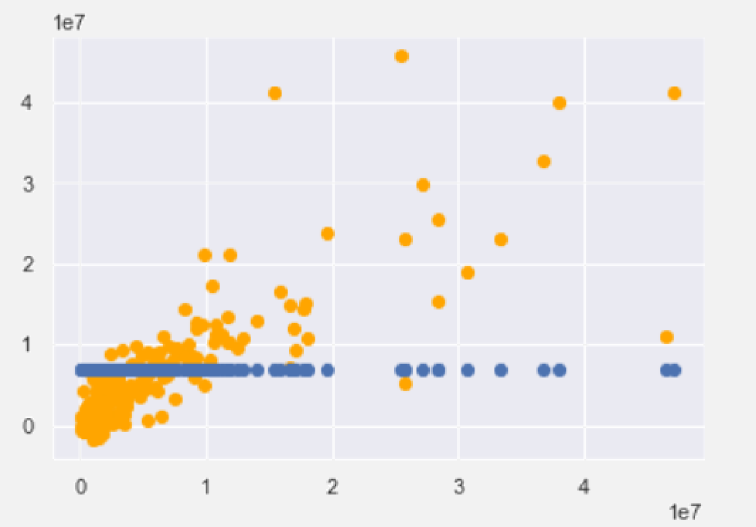
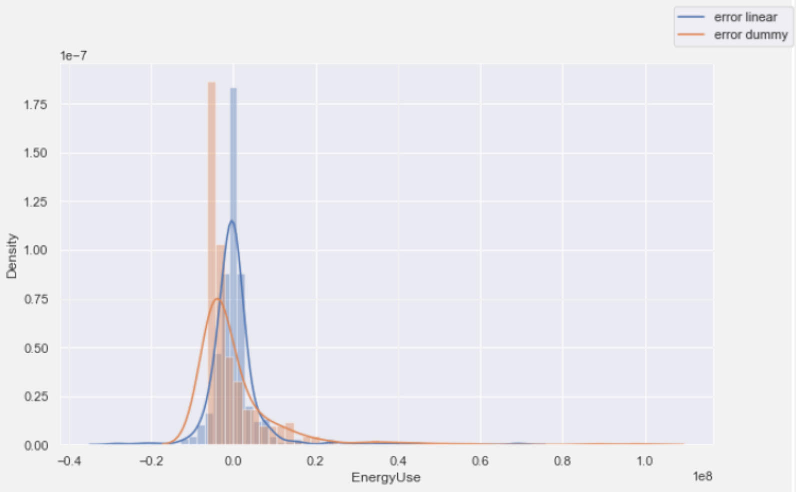
Modeling process

Energy Consumption

(without EnergyScore)

Dummy and Baseline

	Dummy	Linear (baseline)
RMSE	1e7	6.9e6
R2	0.00	0.64

Ypred vs. Ytest	Error model
	

Modeling process

Energy Consumption

(without EnergyScore)

Elements of pre-processing

Dataset selection

- Dataset 1A - EnergyUseLog / All features
- Dataset 1B - EnergyUseLog / All features including those relevant to the Log
- Dataset 2A - EnergyUseLog / 4 Features quant. + PtyType
- Dataset 2B - EnergyUseLog / 4 Features quant. to log+ PtyType

Features transformation selection

- No additional transformation
- PolynomialFeatures
- PCA

Scaling selection

- StandardScaler
- RobustScaler
- QuantileTransformer

Modeling process

Energy Consumption

(without EnergyScore)

First Iteration

Dataset selection
- Dataset 1A - EnergyUseLog / all features

Feature transformation selection
- Poly1

Scaling selection
- Quantile100

Top 3 models

Estimator	Best params	RMSE	R2	Time
RamdomForest	{'max_features': 'sqrt', 'min_samples_leaf': 1, 'n_estimators': 600}	0.42	0.72	1.18
XGBRegressor	{'n_estimators': 20}	0.47	0.70	0.13
LinearRegression		0.48	0.67	0.00

Modeling process

Energy Consumption

(without EnergyScore)

Second iteration

Dataset selection

- Dataset 1B - EnergyUseLog / all features log

Feature transformation selection

- Aucune

Scaling selection

- Robust

Top 3 models

Estimator	Best params	RMSE	R2	Time
SVR	{'C': 10, 'epsilon': 0.1, 'gamma': 0.01}	0.41	0.72	0.05
LinearRegression		0.41	0.72	0.00
ElasticNet	{'alpha': 0.01}	0.41	0.72	0.00

Modeling process

Energy Consumption

(without EnergyScore)

Third Iteration

Dataset selection

- Dataset 2A - EnergyUseLog / 4 Features quant. + PtyType

Feature transformation selection

- PCA4

Scaling selection

- Standard

Top 3 models

Estimator	Best params	RMSE	R2	Time
RandomForest	{'max_features': 'sqrt', 'min_samples_leaf': 5, 'n_estimators': 800}	0.55	0.62	1.05
KNN	{'n_neighbors': 15}	0.57	0.62	0.00
XGBRegressor	{'n_estimators': 20}	0.58	0.61	0.11

Modeling process

Energy Consumption

(without EnergyScore)

Fourth Iteration

Dataset selection

- Dataset 2B - EnergyUseLog / 4 Features quant. to log+ PtyType

Feature transformation selection

- Poly1

Scaling selection

- Robust

Top 3 models

Estimator	Best params	RMSE	R2	Time
RandomForest	{'max_features': 'sqrt', 'min_samples_leaf': 3, 'n_estimators': 500}	0.43	0.71	1.00
LinearRegression	{}	0.44	0.70	0.00
SVR	'C': 10, 'epsilon': 0.01, 'gamma': 0.01}	0.44	0.70	0.05

Presentation of the final model

Energy Consumption

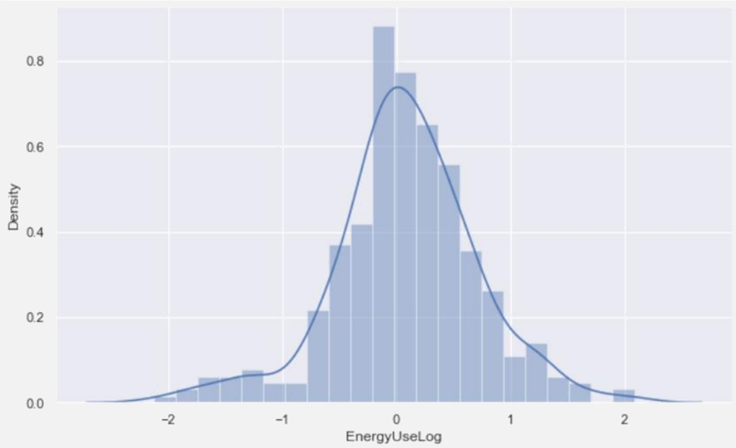
(without EnergyScore)

Final model

Dataset	Dataset 1B - EnergyUseLog / all features log
Scaler	RobustScaler
Features transfo	None
Estimator	SVR
Params	{'C': 10, 'epsilon': 0.1, 'gamma': 0.01}
RMSE	0.41
R2	0.73
Time	0.05

Ypred vs Ytest

Error model



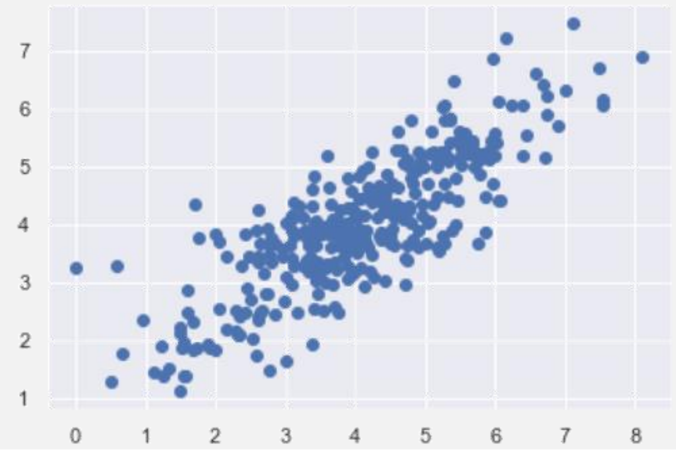
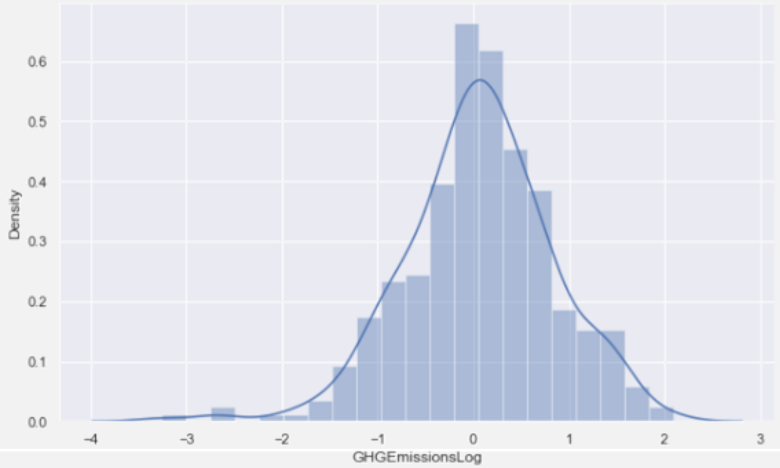
Presentation of the final model

CO2 Emission

(sans EnergyScore)

Final model

Dataset	Dataset 3B - GHGEmissionsLog / all features log
Scaler	StandardScaler
Features transfo	Non
Estimator	XGBRegressor
Params	n_estimators = 20
RMSE	0.61
R2	0.68
Time	0.13

Ypred vs Ytest	Error model
	

Presentation of the final model

Impact of the EnergyStarScore

EnergyStarScore's impact on the Best Prediction Model of each target

Impact on EnergyUse's Best Prediction Model (dataset reduced to buildings with energystarScore)

	Without EnergyStarScore	With EnergyStarScore
RMSE	0.26	0.15
R2	0.82	0.89
Time	0.02	0.02

Impact on Best GHGEmissions prediction model (dataset reduced to buildings with energystarScore)

	Without EnergyStarScore	With EnergyStarScore
RMSE	0.44	0.34
R2	0.75	0.81
Time	0.02	0.02

Presentation of the final model

Conclusions

Relevance and areas of improvement

Relevance of models

- Two models were identified for the two target variables.
- These models are significantly improved with the consideration of EnergyStarScore.

Areas for model improvement

- Identification of optimal features for better modeling
- Further identification of outliers
- Further identification of the most optimal hyperparameters for each model
- Consideration of the EnergyUseWN target variable?

Thank you for your attention!