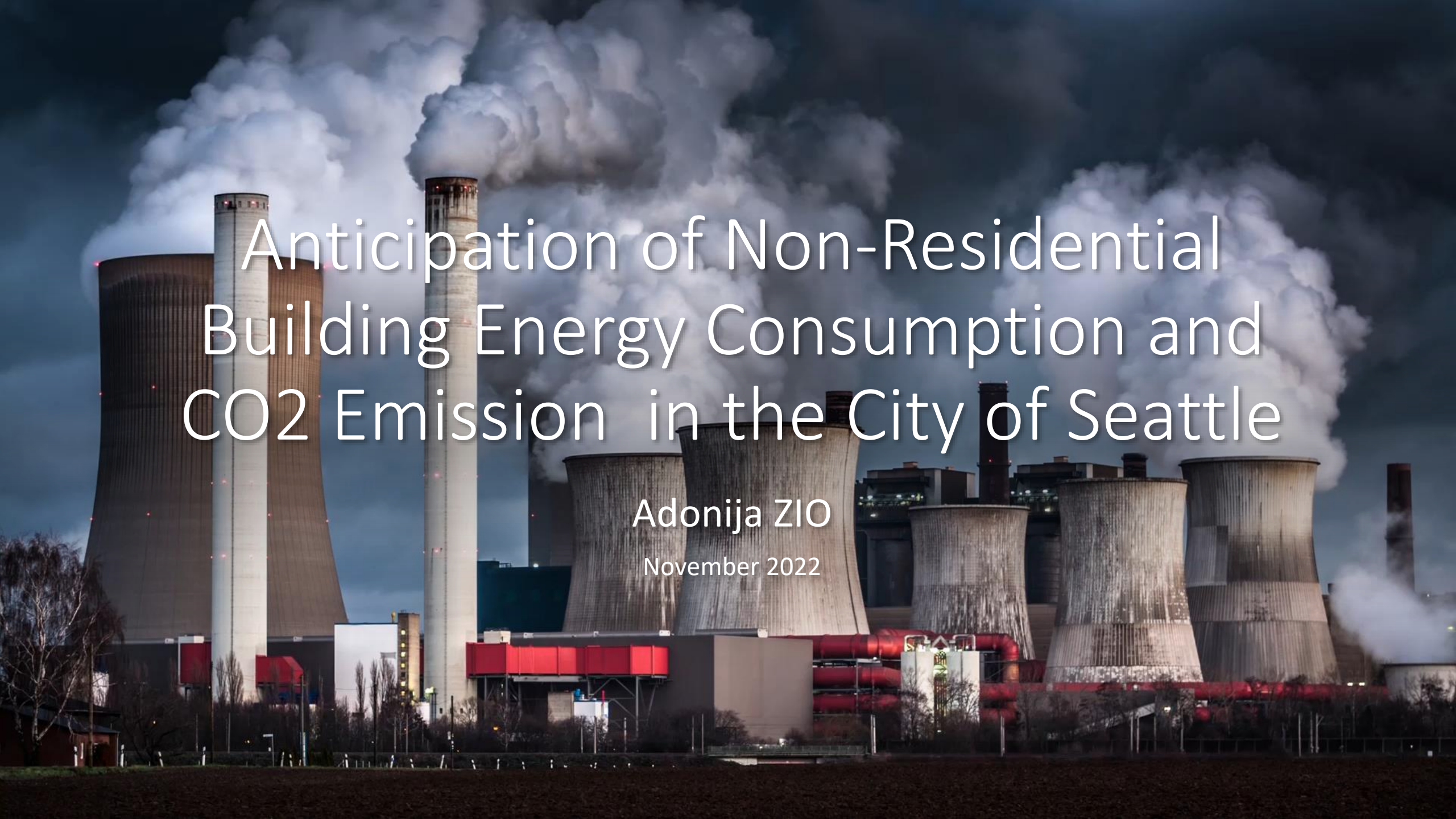


Anticipation of Non-Residential Building Energy Consumption and CO2 Emission in the City of Seattle

Adonija ZIO

November 2022

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

Executive Summary

Summary of Methodologies

Data collection and Data wrangling

Features engineering

Data preparation

Modelling

Summary of results

Ensemble model as the best model of energy consumption prediction

EnergyStarScore is one of the most important feature

Data improvement and news features engineering

Introduction

Project background and context

The growth of environmental concerns in the modern cities

High cost of environment sustaibility policies

Objective of Seattle City

Reduction of Seattle City CO2 Emission of nonresidential building

Reduce the cost of this policy

Question of interest

What are the determinants of energy consumption and CO2 emissions in non-residential buildings in the city of Seattle?

- Constraint: Energy consumption and CO2 anticipation with low cost

Methodology

FEATURES PREVIEW

FEATURES ENGINEERING

EXPLORATORY ANALYSIS

PREDICTIVE ANALYSIS (REGRESSION)

FEATURES IMPORTANCES ANALYSIS

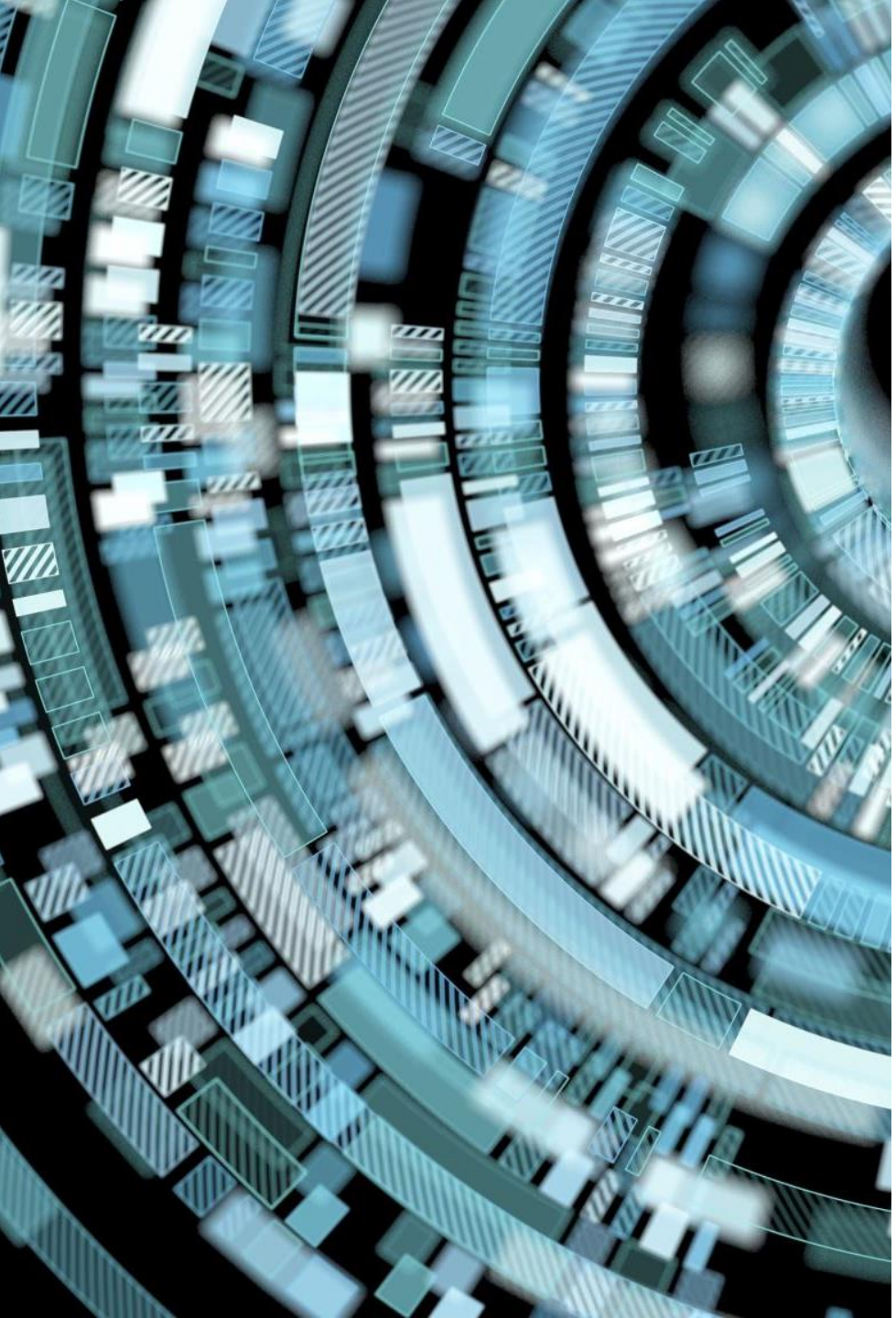
Data understanding and data wrangling

Categorical features preview

- Selection of the candidates of interest
- Outliers drop with the feature outlier
- Filling Categorical Missing value

Numerical features preview

- Droping numerical outlier
- Droping absud values
- Droping data leakage features



Features Engineering

- Catagorical features engineering
  - Grouping categorical features categories by their statistical characteristics
  - Creation of Boolean features (`parking, SecondUsage, ThirdUsage, SteamUse, GasUse`)
  - Creation of Categorical features (MainEnergyUse)
- Numerical features engineering
  - Numerical features (`LargestGFArate, SecondGFArate, ThirdGFArate, BuildingGFARate, …`)
  - Numerical Encoding (`PrimaryGFAProportion, LargestGFAProportion, SecondGFAProportion, ThirdGFAProportion,…`)

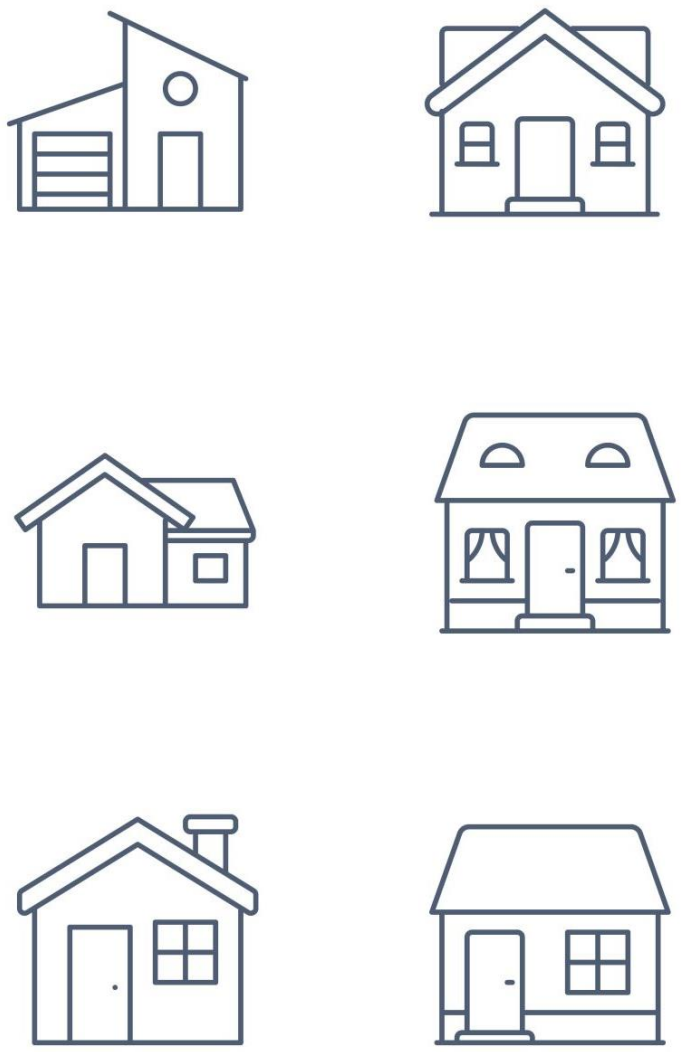
Exploratory analysis

Relationship between categorical features and Target

- Categorical distribution with the targets variables
- Geographical location effect on targets variables

Relationship between features and targets variables

- Correlation analysis
- Numerical features and targets variables



Predictive analysis (Regression)

- Building Model
  - Baseline definition
  - Model selection
- Model Evaluation
  - Metrics definition
  - Metrics calculation
- Best fit model
  - Hyperparameters tuning
  - Model comparison

Results

Exploratory Data Analysis

- Categorical features visualization
- Numerical features visualization

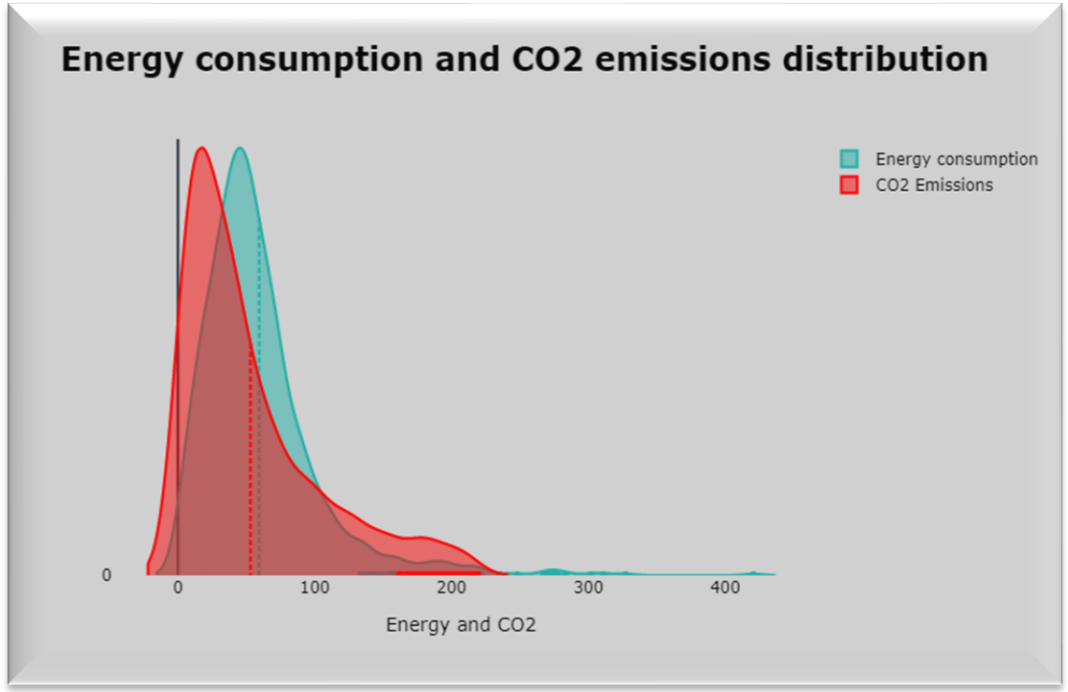
Predictive Analysis

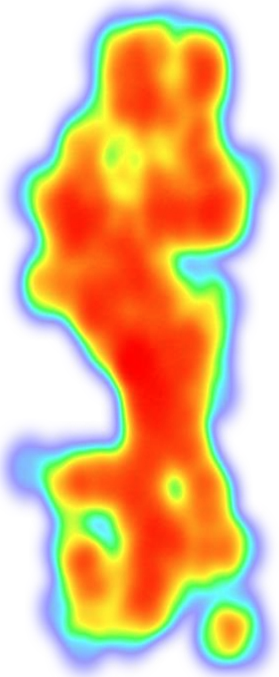
- Model Building
- Model Selection

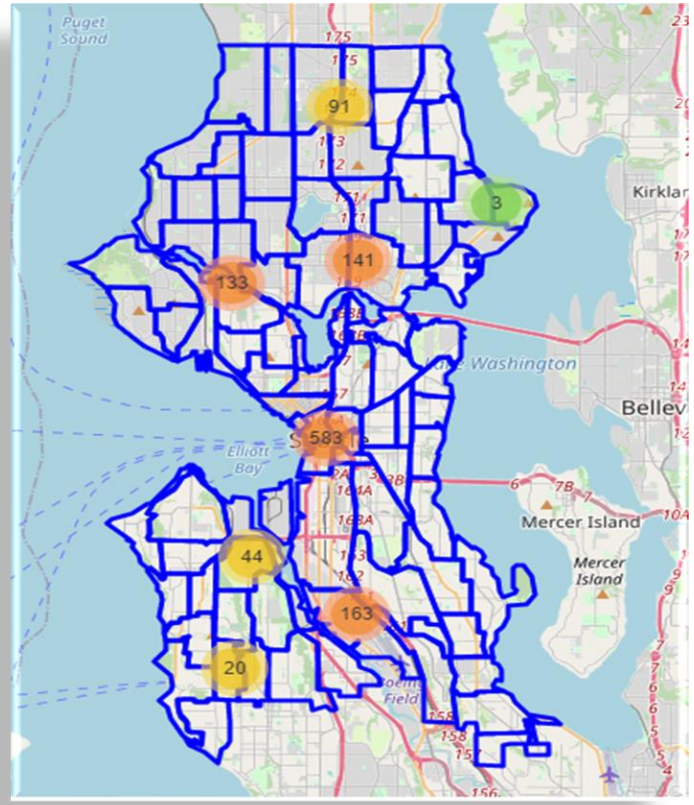
Features importances

- Features Importances without ESS
- Features importances with ESS

Exploratory Data Analysis

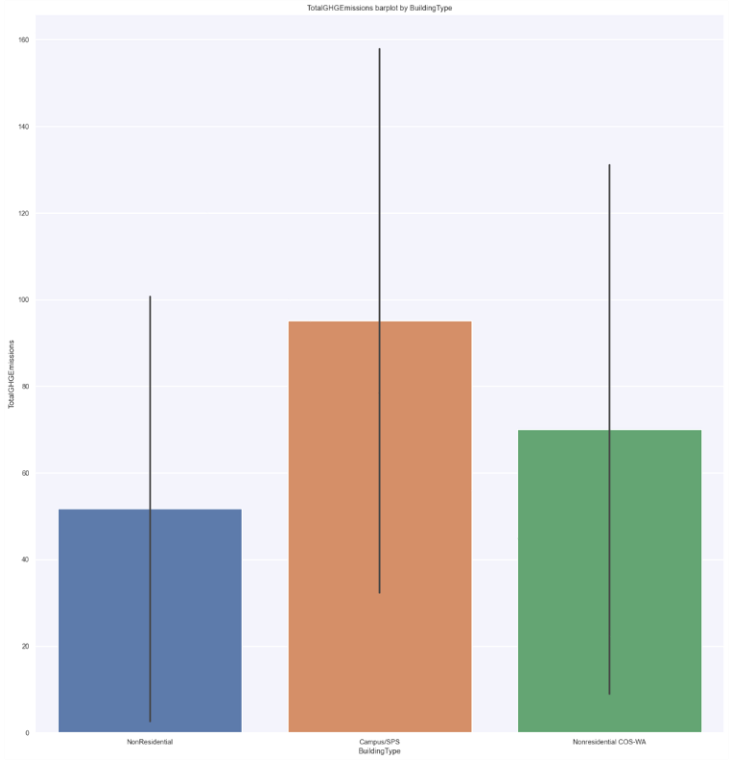


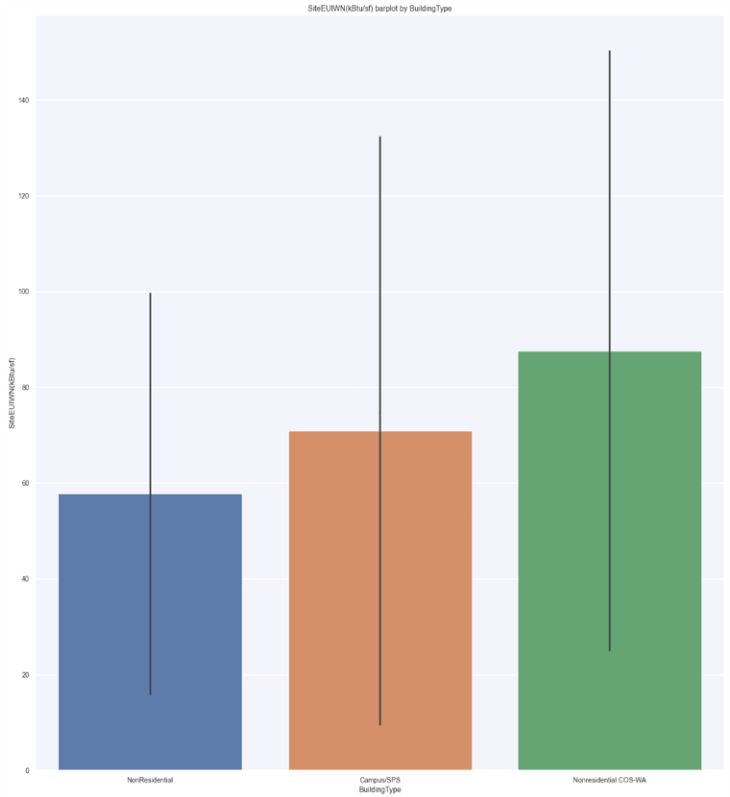


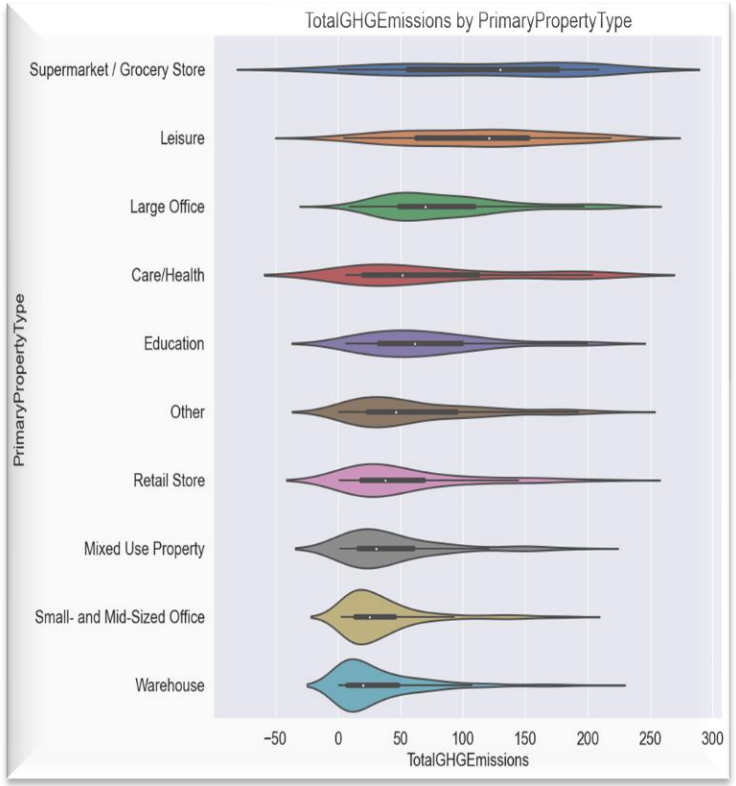


Exploratory Data Analysis

- Relationship between categorical features and target variables

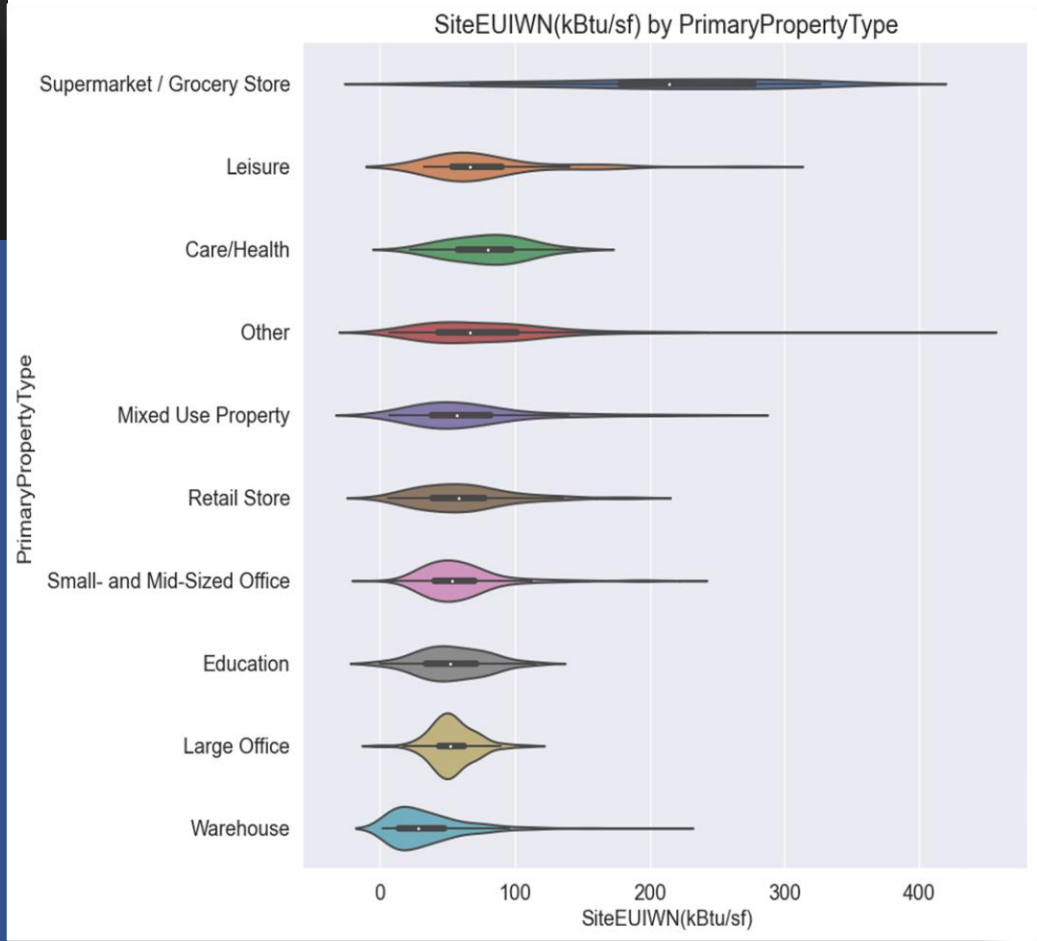


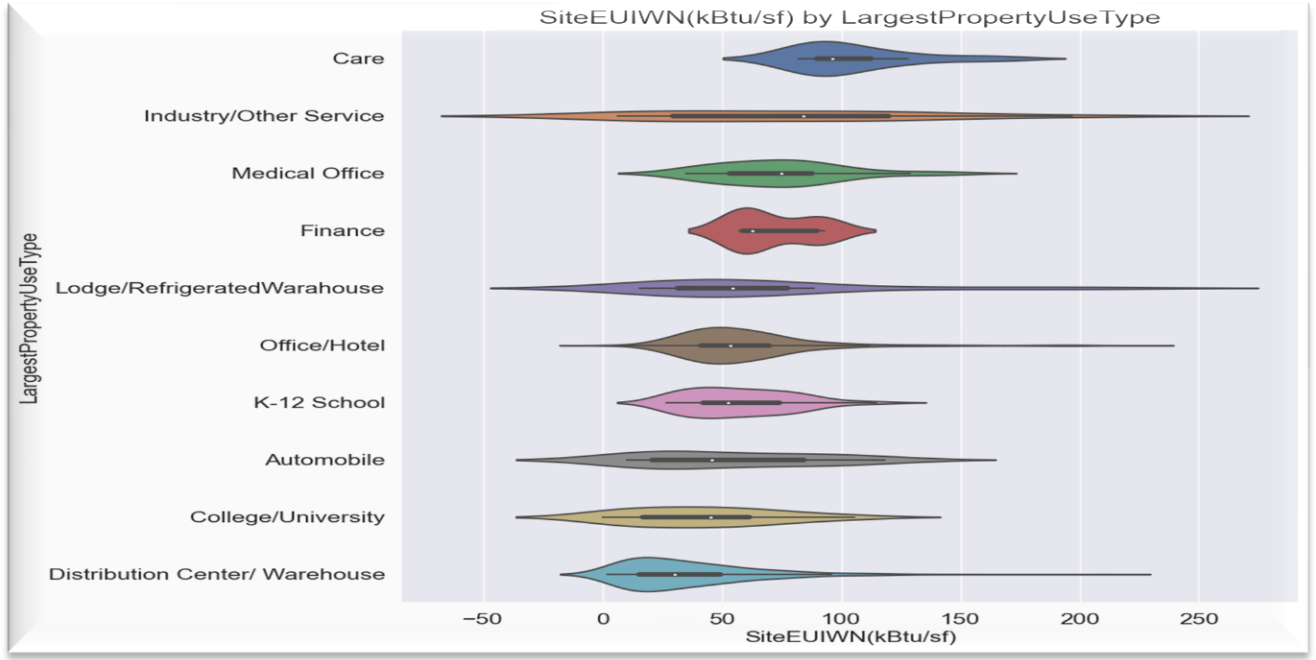


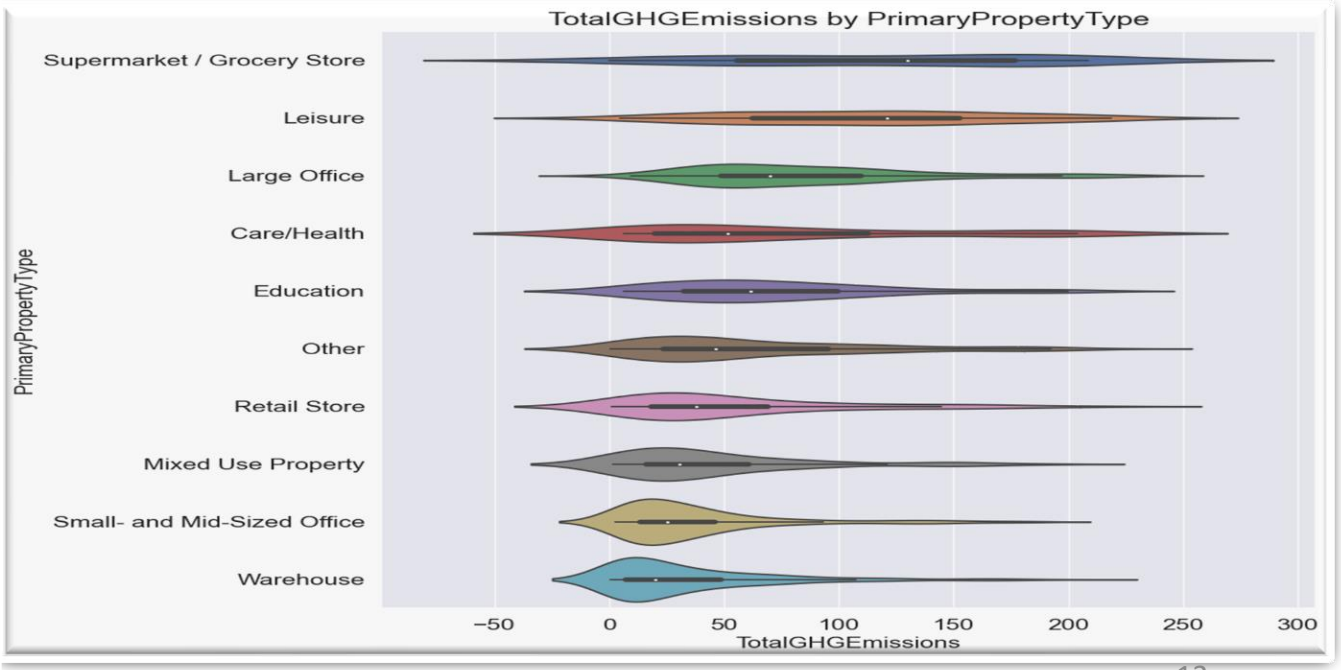


Exploratory Data Analysis

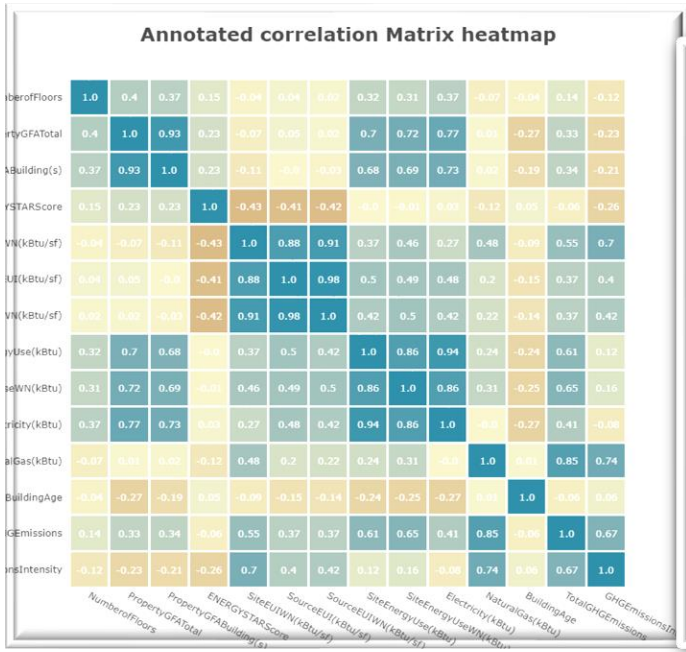
Relationship between categorical features and target variables

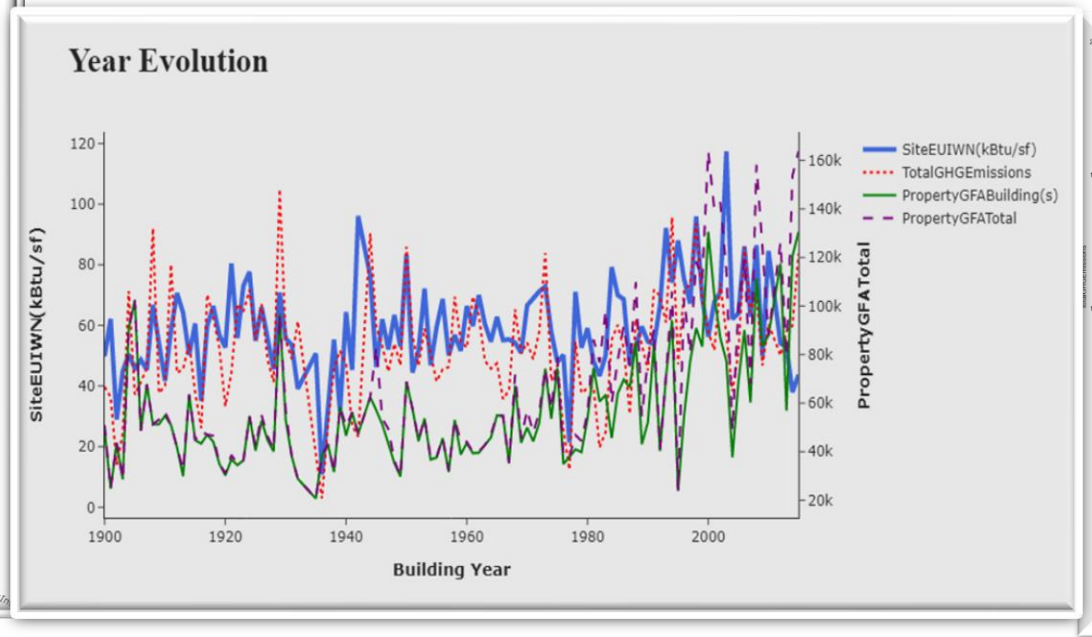


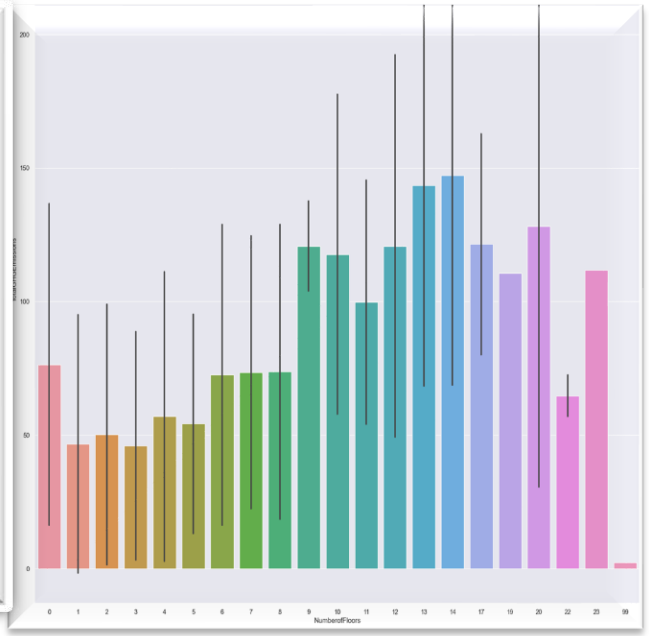




Exploratory Data Analysis: Numerical features relationship







Predictive analysis: Model Building

Energy Consumption

	Dummy	Elastic Net	SVR	Random Forest	Extra Tree	Gradient Boosting	KNeighbors	Linear Regression	Decision Tree	XGB	Lasso	Ridge	SGD	AdaBoost
R^2	-0.000	5.391e-1	-1.334e-1	6.921e-1	7.106e-1	7.354e-1	6.179e-1	5.786e-1	4.656e-1	7.118e-1	5.786e-1	5.791e-1	-2.28574e+4	4.283e-1
MAE	2.146e+6	1.435e+6	1.9868e+6	1.1322e+6	1.0965e+6	1.0699e+6	1.269e+6	1.3117e+6	1.4840e+6	1.1168e+6	1.3117e+6	1.3050e+6	2.6456e+8	1.8953e+6
RMSE	2.825e+6	1.918e+6	3.0078e+6	1.5676e+6	1.5197e+6	1.4533e+6	1.7463e+6	1.834e+6	2.0653e+6	1.5166e+6	1.834e+6	1.8329e+6	4.2715e+8	2.1362e+6
MAPE	1.567	9.648e-1	9.949e-1	6.406e-1	6.095e-1	6.034e-1	7.308e-1	7.216e-1	7.018e-1	6.302e-1	7.216e-1	7.179e-1	2.23092e+2	1.7299

TOTAL GHG Emissions

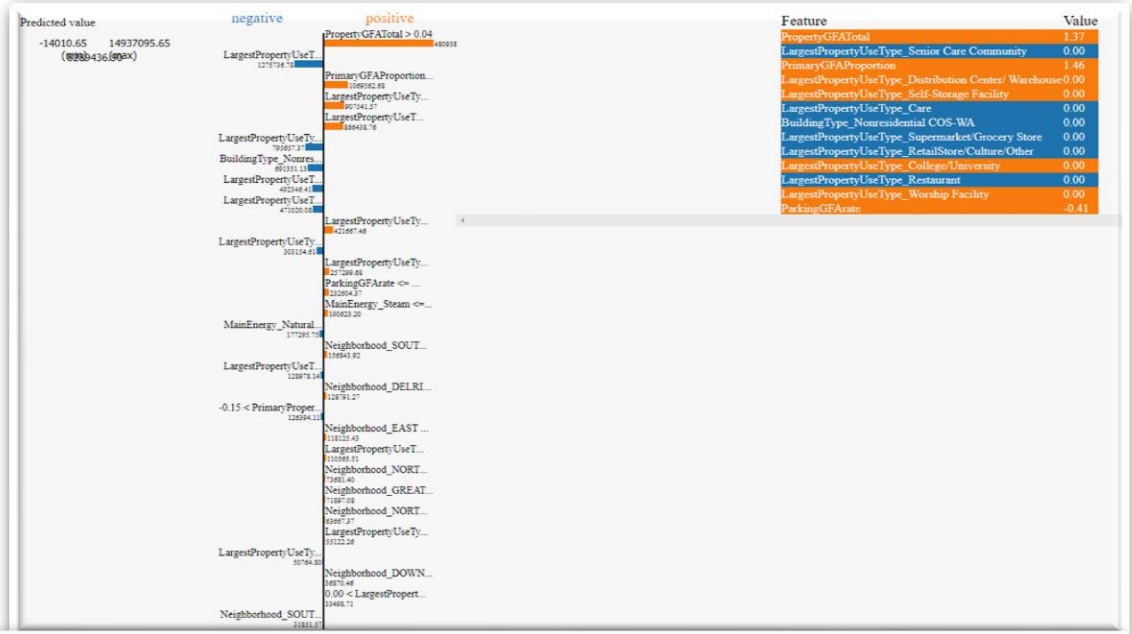
	Dummy	Elastic Net	SVR	Random Forest	Extra Tree	Gradient Boosting	KNeighbors	Linear Regression	Decision Tree	XGB	Lasso	LassoCV	LassoLars CV	Ridge	SGD	AdaBoost
R^2	-0.00	0.2517	-0.06	0.3321	0.188	0.42	0.31	0.4	-0.001	0.17	0.37	0.4122	0.41	0.40	-4.88e+11	0.2022
MAE	38.42	32.26	34.94	28.75	31.48	27.57	29.51	28.83	34.7264	32.31	29.84	28.64	28.68	28.78	1.58e+7	38.2765
RMSE	50.06	43.3	51.65	40.91	45.11	38.19	41.53	38.71	50.0879	45.66	39.69	38.37	38.45	38.58	3.49e+7	44.7128
MAPE	2.64	2.07	1.66	1.32	1.3626	1.1939	1.37	1.36	1.2660	1.61	1.58	1.36	1.37	1.36	6.75e+5	2.68

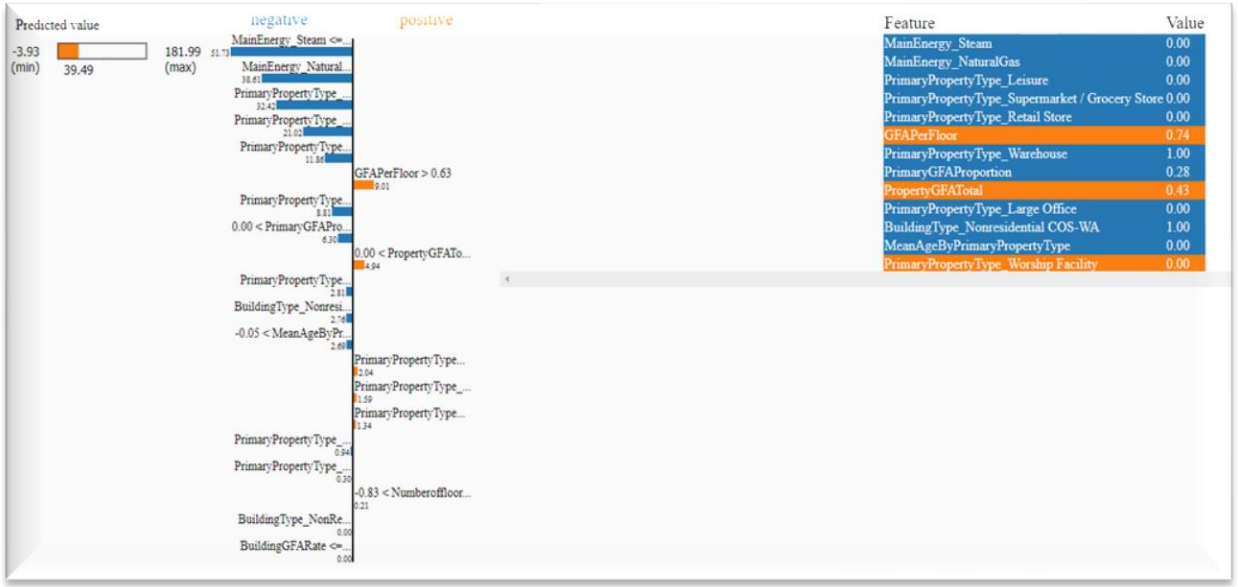
Hyperparameter tuning and Model selection

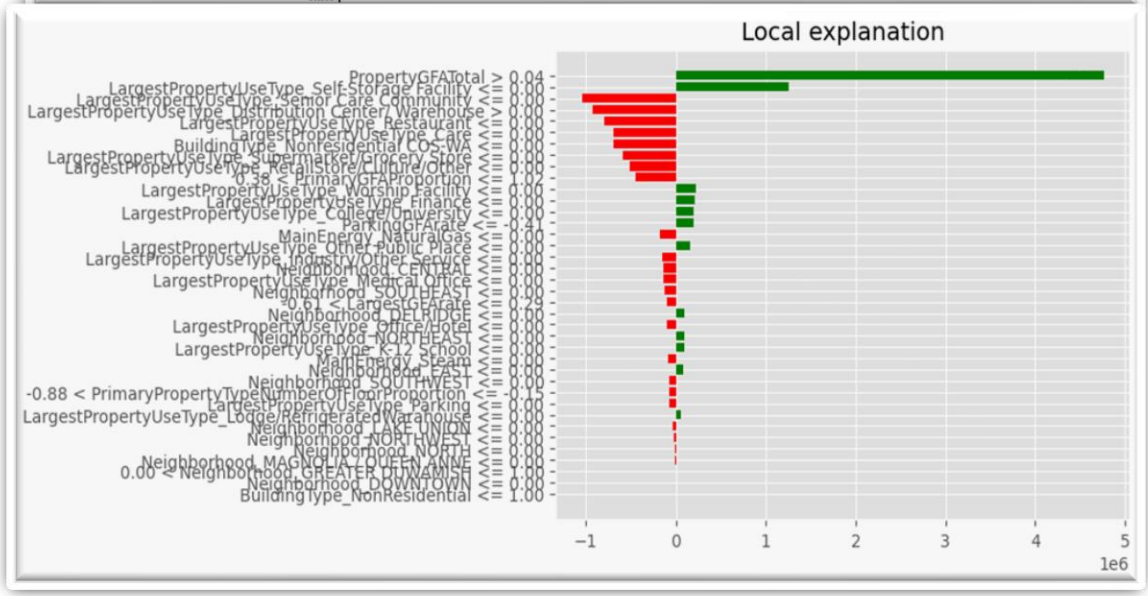
Energy Consumption			
	mean_test_score	test_score	train_score
Random Forest	0.628583	0.719983	0.917572
Gradient Boosting	0.655375	0.709855	0.772623
Extra Tree	0.655338	0.740479	0.813963
XGB	0.651438	0.717441	0.788149
Elastic Net	0.349125	0.409408	0.378343
KNeighbors	0.560731	0.628901	0.716398
Linear Regression	0.630524	0.578580	0.684501
Lasso	0.635667	0.579087	0.683825
Ridge	0.637002	0.577901	0.682150

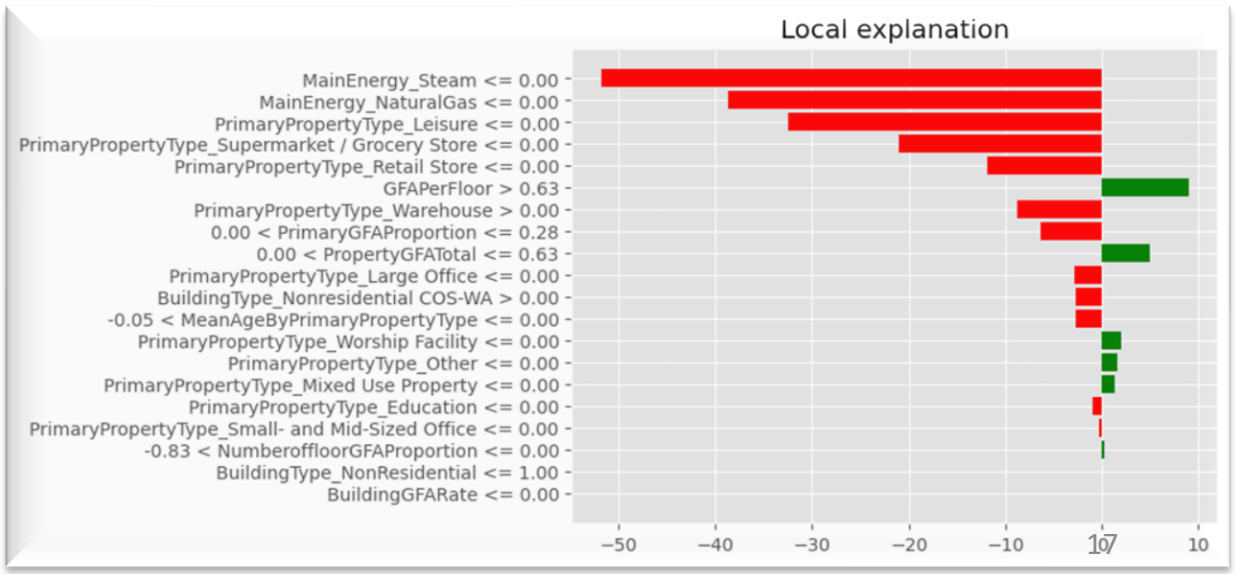
TOTAL GHG Emissions			
	mean_test_score	test_score	train_score
Random Forest	0.327674	0.393327	0.754677
Gradient Boosting	0.390514	0.447643	0.538767
Extra Tree	0.371981	0.410987	0.540685
XGB	0.383194	0.420030	0.555418
Elastic Net	0.117615	0.143895	0.128237
KNeighbors	0.260316	0.325260	0.406968
Linear Regression	0.373277	0.401896	0.410746
Lasso	0.376755	0.405900	0.410340
Ridge	0.377404	0.407571	0.409690

Features importances Without ESS









Model with ENERGYSTARSCORE

	mean_test_score	test_score	train_score
Random Forest	0.810030	0.841324	0.972957
Gradient Boosting	0.853881	0.866077	0.975143
Extra Tree	0.816429	0.820377	0.956928
XGB	0.827313	0.852216	0.942340
Elastic Net	0.371047	0.423996	0.383837
KNeighbors	0.696142	0.706665	1.000000
Linear Regression	0.758458	0.793009	0.808901
Lasso	0.760788	0.794415	0.808391
Ridge	0.762564	0.796845	0.804506

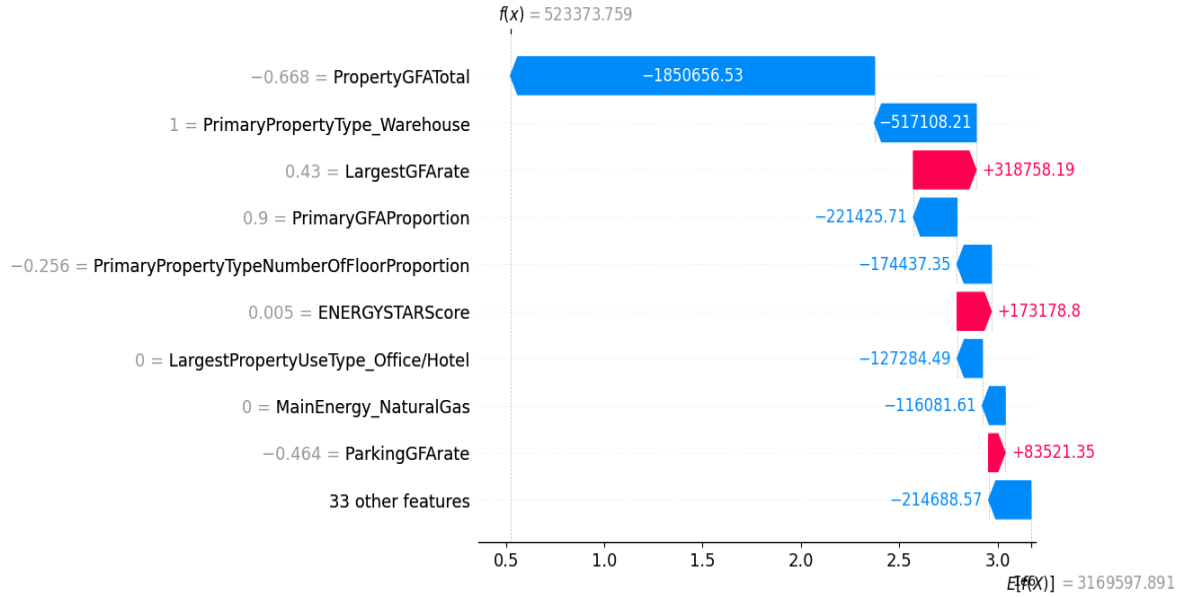
The overall best cross-validated score is : R^2 = 0.854
The best model is **Gradient Boosting Regressor** with parameters:
- learning_rate: 0.01
- - max_depth: 5
- - n_estimators: 1000
- - subsample: 0.3

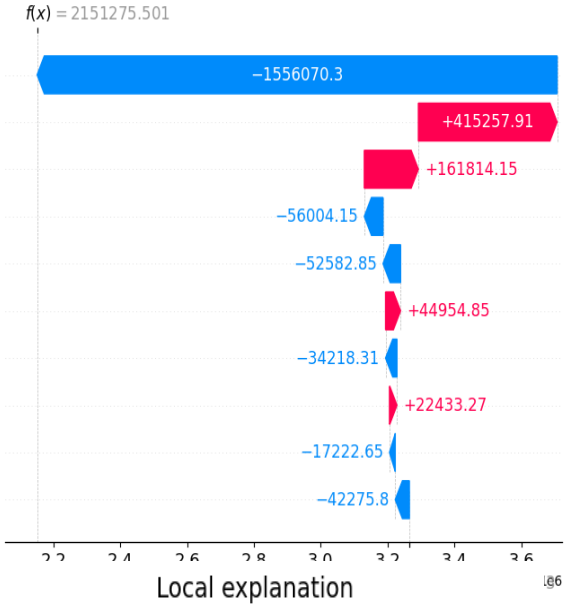
	mean_test_score	test_score	train_score
Random Forest	0.480830	0.416615	0.923479
Gradient Boosting	0.478084	0.504279	0.707597
Extra Tree	0.494845	0.471096	0.838995
XGB Regressor	0.478417	0.481708	0.777819
Elastic Net	0.143629	0.184598	0.159149
KNeighbors	0.328874	0.289181	1.000000
Linear Regression	0.450028	0.500401	0.503901
Lasso	0.450395	0.508548	0.502631
Ridge	0.450823	0.506174	0.503401

The overall best cross-validated with GridSearchCV score is : R^2 = 0.854
The best model is **Gradient Boosting Regressor** with parameters:
- learning_rate: 0.01
- - max_depth: 5
- - n_estimators: 1000
- - subsample: 0.3

The overall best cross-validated with BayesSearch CVscore is : R^2 = 0.505
The best model is **Extra Tree Regressor** with parameters:
- max_features: 10
- - min_samples_leaf: 2
- - min_samples_split: 2
- - n_estimators: 100

Features importances with ESS





Conclusion

- Data size extending
- Improving score by a new features engineering
- New features collection
- Collect ENERGYSTARSCORE for each building