



Seattle

Anticipez les besoins en consommation électrique de bâtiments





Agenda

- 1 Contexte
- 2 Approche
- 3 Présentation des jeux de données
- 4 Préparation
- 5 Machine Learning
- 6 Prochaines étapes
- 7 Annexe : l'environnement technique



1) Contexte

Objectif

La ville de **Seattle** a pour objectif de ville neutre en émissions carbonées en 2050. A ce stade, une attention particulière est donnée sur les émissions des bâtiments non destinés à l'habitation.

Données mises à disposition

La ville de Seattle a effectué des relevés minutieux en 2015 et 2016 sur les bâtiments.

Mission

Prédiction des émissions de CO2 et la consommation totale d'énergie de bâtiments pour lesquels elles n'ont pas encore été mesurées.



2) Approche



1) Présentation

Présentation des jeux de données (Taille du jeu de données, Type de données, nombre de lignes, nombre de colonnes, données manquantes...)

2) Préparation



Uniformisation du nom des colonnes, uniformisation des données au sein des categories, Sélection des colonnes pertinentes, enrichissement des données manquantes, Encoding (OneHotEncoding, ...), Normalisation / Standardisation des données,

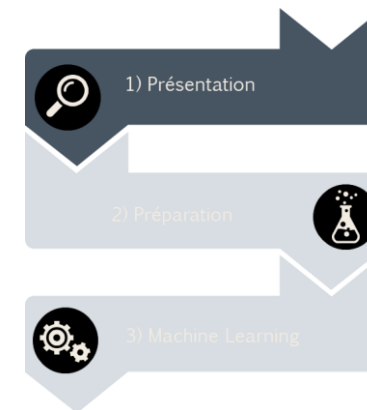


3) Machine Learning

Application de différents algorithmes de machine learning (linéaire, ...)



3) Présentation des jeux de données (1/2)



Jeu de données 2015

OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber	Location	CouncilDistrictCode	N
0	1	2015	NonResidential	Hotel	MAYFLOWER PARK HOTEL	659000030	{'latitude': '47.61219025', 'longitude': '-122...	7
1	2	2015	NonResidential	Hotel	PARAMOUNT HOTEL	659000220	{'latitude': '47.61310583', 'longitude': '-122...	7
2	3	2015	NonResidential	Hotel	WESTIN HOTEL	659000475	{'latitude': '47.61334897', 'longitude': '-122...	7
3	5	2015	NonResidential	Hotel	HOTEL MAX	659000640	{'latitude': '47.61421585', 'longitude': '-122...	7
4	8	2015	NonResidential	Hotel	WARWICK SEATTLE HOTEL	659000970	{'latitude': '47.6137544', 'longitude': '-122...	7
...

Extrait du jeu de données 2015

3340 lignes

47 colonnes

Indicateurs clés du jeu de données

Jeu de données 2016

OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode	TaxParcelIdentificationNumber	
0	1	2016	NonResidential	Hotel	Mayflower park hotel	405 Olive way	Seattle	WA	98101.0	0659000030
1	2	2016	NonResidential	Hotel	Paramount Hotel	724 Pine street	Seattle	WA	98101.0	0659000220
2	3	2016	NonResidential	Hotel	5673-The Westin Seattle	1900 5th Avenue	Seattle	WA	98101.0	0659000475
3	5	2016	NonResidential	Hotel	HOTEL MAX	620 STEWART ST	Seattle	WA	98101.0	0659000640
4	8	2016	NonResidential	Hotel	WARWICK SEATTLE HOTEL (ID8)	401 LENORA ST	Seattle	WA	98121.0	0659000970
...

Extrait du jeu de données 2016

3376 lignes

46 colonnes

Indicateurs clés du jeu de données



3) Présentation des jeux de données (2/2)



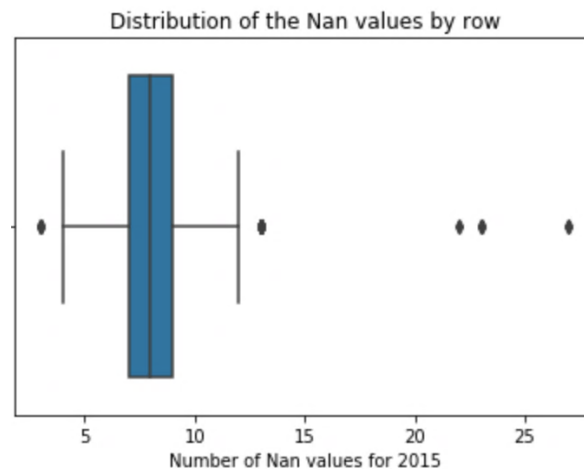
1) Présentation

2) Préparation



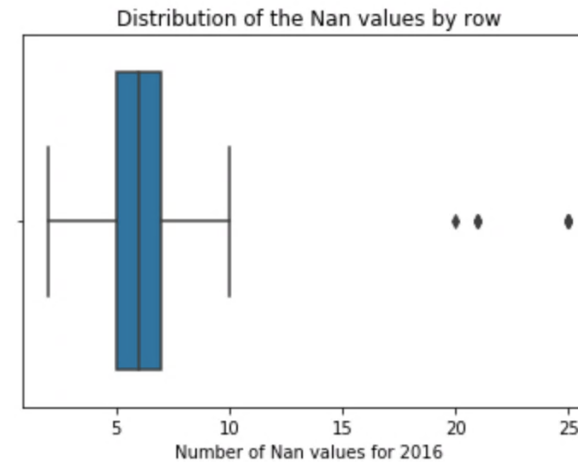
3) Machine Learning

Analyse des Nan pour 2015



Caractéristiques	Valeur
Nombre de lignes	3330
Moyenne du nombre de valeurs	7.937725
Ecart type	1.961370
Min de valeurs	3
25%	7
50%	8
75%	9
Max	27

Analyse des Nan pour 2016



Caractéristiques	Valeur
Nombre de lignes	3376
Moyenne du nombre de valeurs	5.909953
Ecart type	1.763215
Min de valeurs	2
25%	5
50%	6
75%	7
Max	25

Le jeu de données pour 2016 est plus complété que celui de 2015. Afin d'avoir un algorithme de Machine Learning efficace, un enrichissement des données est à effectuer.



4) Préparation

Uniformisation des noms des colonnes et des données

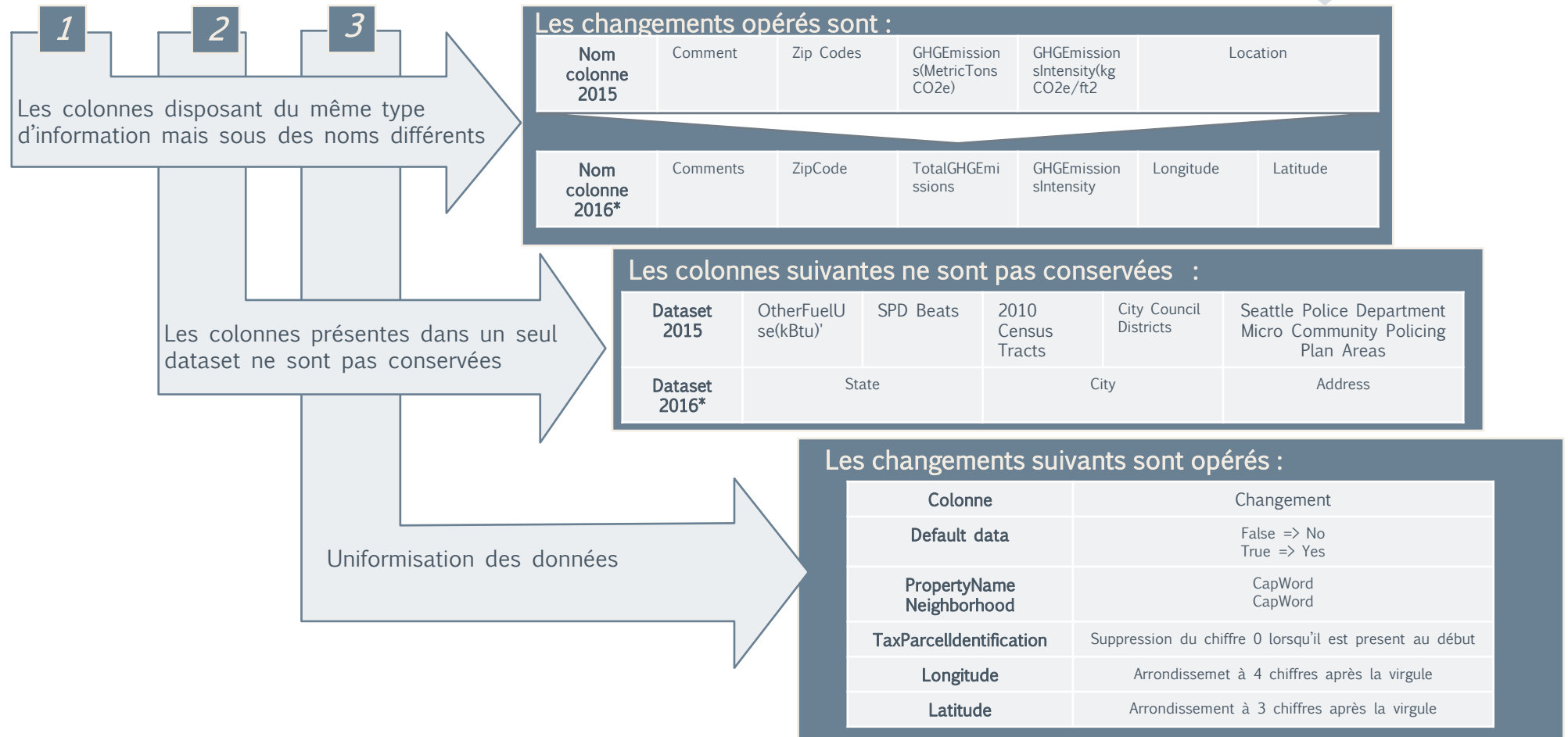


1) Présentation

2) Préparation



3) Machine Learning

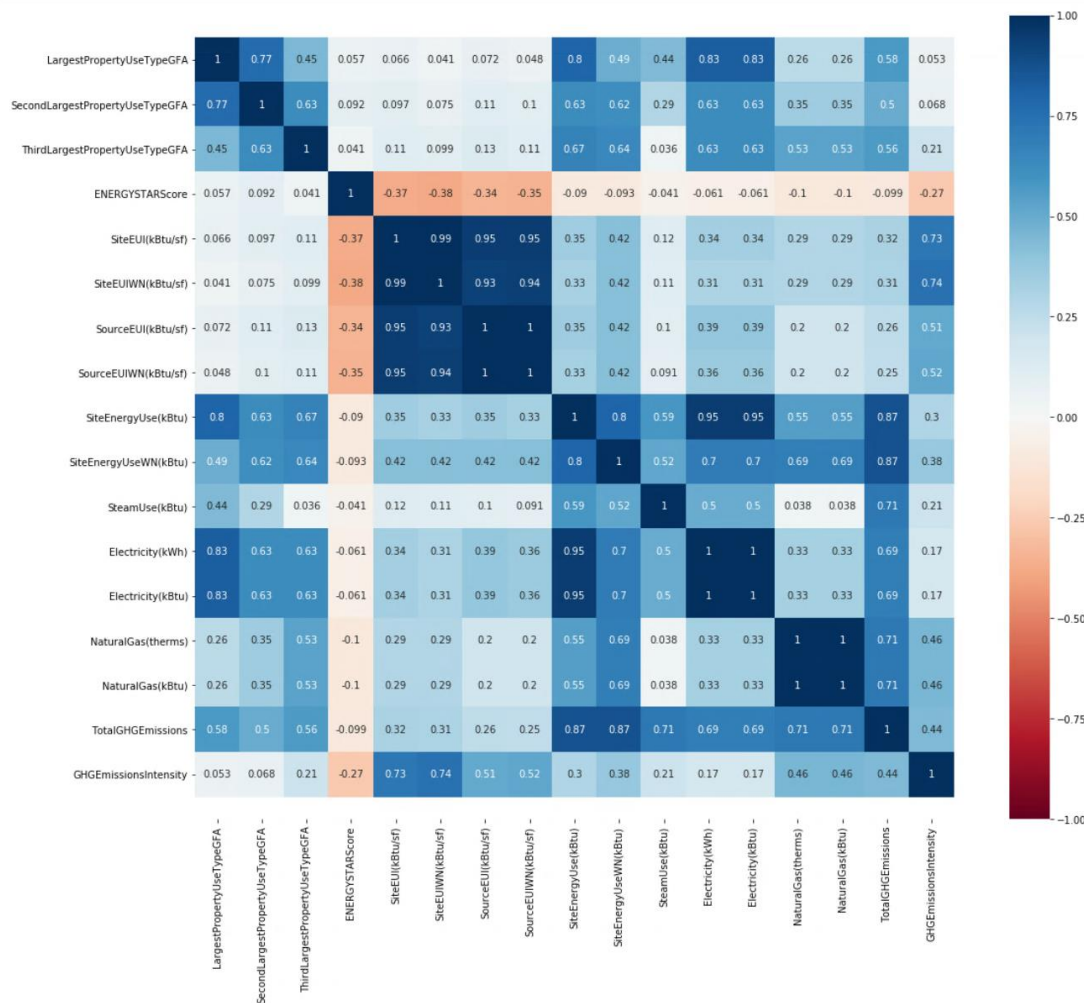


* Le dataset final est basé sur celui de 2016



4) Préparation

Analyse des corrélations



- Corrélation importante entre les éléments suivants :

Valeur 1*	Valeur 2
'SiteEUI(kBtu/sf)'	'SiteEUIWN(kBtu/sf)'
'SourceEUI(kBtu/sf)'	'SourceEUIWN(kBtu/sf)'
'SiteEnergyUse(kBtu)'	'SiteEnergyUseWN(kBtu)'
'Electricity(kBtu)'	'Electricity(kWh)'
'NaturalGas(kBtu)'	'NaturalGas(therms)'

*afin d'éviter le sur-apprentissage de l'algorithme, uniquement les valeurs en gras sont conservées.

- L'energyScore est très faiblement corrélé avec la consommation d'énergie, cet indicateur ne semble pas pertinent, il est conservé cependant afin de vérifier son influence sur le modèle.



1) Présentation

2) Préparation



3) Machine Learning



4) Préparation

Les principales colonnes sélectionnées (1/4)














1) Présentation

2) Préparation



3) Machine Learning

#	Code	Description	Raison	Action
1	OSEBuildingID	Identifiant d'un bâtiment	 Donnée clé	
2	DataYear	Année des mesures	 Donnée clé	
3	BuildingType	Type de bâtiment	 Donnée input algo	 A encoder (texte)
4	PrimaryPropertyType	Type Principale de la propriété	 Donnée input algo	 A encoder (texte)
5	Neighborhood		 Donnée input algo	 A encoder (texte)
6	YearBuilt	Année de construction	 Donnée clé	
7	NumberOfBuilding	Nombre de bâtiments	 Donnée input algo	

Légende :



Données à encoder



Données à binariser



Aucune action



Données en entrée



Données en sortie



Données clés



4) Préparation

Les principales colonnes sélectionnées (2/4)



1) Présentation

2) Préparation



3) Machine Learning

#	Code	Description	Raison	Action
8	PropertyGFATotal	Surface au sol totale	 Donnée input algo	
9	PropertyGFAParking	Surface au sol des parkings	 Donnée input algo	
10	PropertyGFABuilding(s)	Surface au sol des bâtiments	 Donnée input algo	
11	LargestPropertyUseType	Usage principale de la propriété	 Donnée input algo	 A encoder (texte)
12	SecondLargestPropertyUseType	Usage secondaire de la propriété	 Donnée input algo	 A encoder (texte)
13	ThirdLargestPropertyUseType	Usage tertiaire de la propriété	 Donnée input algo	 A encoder (texte)
14	SiteEUI(kBtu/sf)	Quantité d'énergie nécessaire annuelle consommée	 Donnée input algo	 A binariser (fuite)

Légende :



Données à encoder



Données à binariser



Aucune action



Données en entrée



Données en sortie



Données clés



4) Préparation

Les principales colonnes sélectionnées (3/4)

















1) Présentation

2) Préparation



3) Machine Learning

#	Code	Description	Raison	Action
15	SourceEUI(kBtu/sf)	Quantité d'énergie nécessaire annuelle pour fonctionner	 Donnée input algo	 A binariser (fuite)
16	SteamUse(kBtu)	Quantité de vapeur consommée	 Donnée input algo	 A binariser (fuite)
17	Electricity(kBtu)	Quantité d'électricité consommée	 Donnée input algo	 A binariser (fuite)
18	NaturalGas(kBtu)	Quantité de gaz consommé	 Donnée input algo	 A binariser (fuite)
19	GHGEmissionsIntensity		 Donnée input algo	 A binariser (fuite)
20	EnergySTARSCORE	Score Energy	 Donnée input algo	 A encoder (texte)
21	SiteEnergyUse(kBtu)	Consommation énergétique du site	 Donnée output algo	

Légende :



Données à encoder



Données à binariser



Aucune action



Données en entrée



Données en sortie





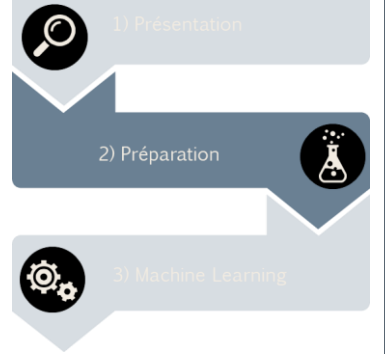
Données clés



4) Préparation

Les principales colonnes sélectionnées (4/4)

#	Code	Description	Raison	Action
22	TotalGHGEmissions	Quantité des gaz à effet de serre émis	 Donnée output algo	



Légende :

















4) Préparation

L'enrichissement des données (1/2)

Principe : pour un bâtiment donné, si une donnée signalétique est présente une année et absente l'autre, cette dernière est recopiée.

#	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType
1	1	2015		
2	1	2016		
3	2	2015		
4	2	2016		
5			
6	3000	2015		
7	3000	2016		



1) Présentation

2) Préparation



3) Machine Learning



4) Préparation

L'enrichissement des données avec homogénéisation des données (2/2)



1) Présentation

2) Préparation



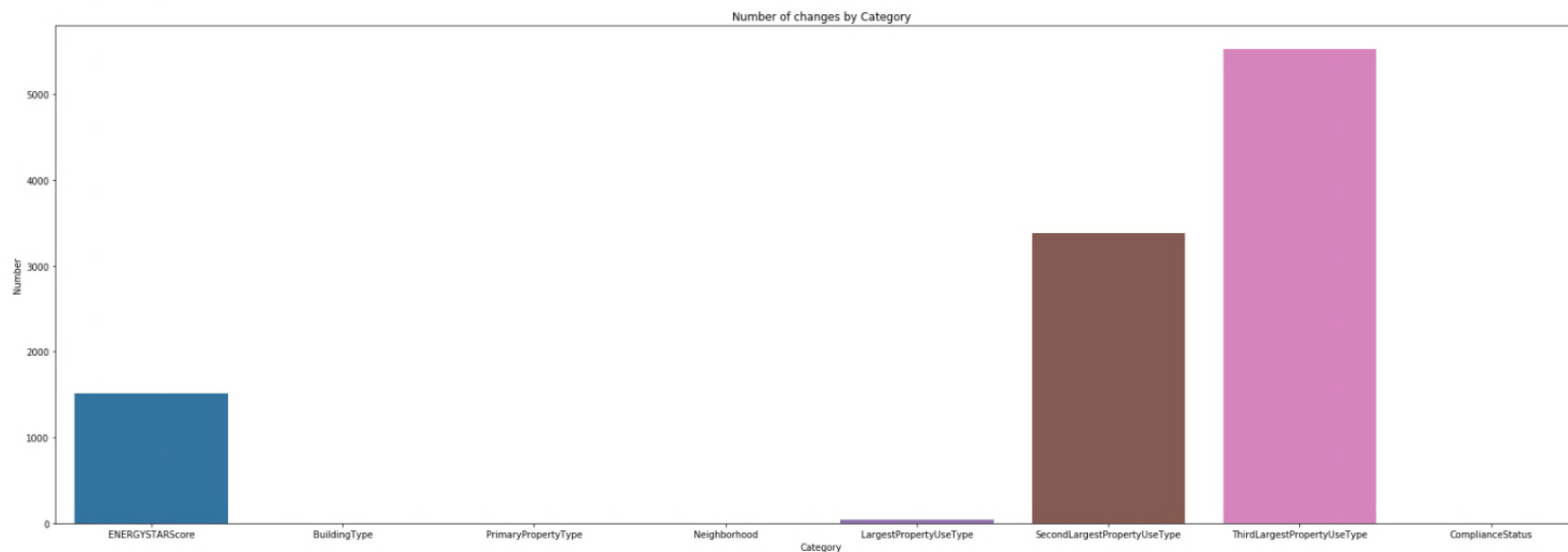
3) Machine Learning

1) Homogénéisation

Les changements opérés sont :

Intitulé source	University	Warehouse	Residence Hall	SPS-District K12	Supermarket / Grocery Store	Other-Mall	Supermarket/Grocery
Intitulé cible	College University	Non-Refrigerated Warehouse	Residence Hall / Dormitory	K12 School	Retail Store	Retail Store	Retail store

2) Résultats de l'homogénéisation et l'enrichissement des données





4) Préparation

Encoding

Label Encoding

#	DataToEncode	#	DataToEncode	EncodedData
1	Label1	1	Label1	1
2	Label2	2	Label2	2
3	Label3	3	Label3	3
4	4	
5	Label3	5	Label3	3
6	Label2	6	Label2	2
7	Labe	7	Labe	1

Critère	Commentaires
Nombre de colonnes	18 dans le jeu de données
Temps d'exécution	1.29s
Inconvénients	Les valeurs encodées influencent le modèle

One Hot Encoding

#	DataToEncode	#	DataTo-Encode	Label1	Label2	Label3
1	Label1	1	Label1	1		
2	Label2	2	Label2		1	
3	Label3	3	Label3			1
4	4			
5	Label3	5	Label3			1
6	Label2	6	Label2		1	
7	Labe	7	Labe	1		

Critères	Commentaires
Nombre de colonnes	207 dans le jeu de données
Temps d'exécution	5.59s
Inconvénient	Augmentation des temps d'exécution des algorithmes Augmentation du nombre de colonnes

Target Encoding

#	DataTo Encode	ValueTo Assess	#	DataToEncode	EncodedData
1	Label1	1	1	Label1	Mean(ValueToAssess)ForLabel1
2	Label2	2	2	Label2	Mean(ValueToAssess)ForLabel2
3	Label3	3	3	Label3	Mean(ValueToAssess)ForLabel3
4		4	
5	Label3	4	5	Label3	Mean(ValueToAssess)ForLabel3
6	Label2	4	6	Label2	Mean(ValueToAssess)ForLabel2
7	Label	6	7	Labe1	Mean(ValueToAssess)ForLabel1

Critère	Commentaires
Nombre de colonnes	18 dans le jeu de données
Temps d'exécution	1.29s
Commentaires	Meilleur compromis pour cet analyse. Les résultats décrits par la suite seront basés sur cet encoding



1) Présentation

2) Préparation



3) Machine Learning



5) Machine Learning

SmartData

Encoding

- LabelEncoding
- OneHotEncoding
- TargetEncoding

DataConversion

- Normalisation
- Standardisation

Output

- Energy
- GHG

- Dummy Regression
- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- SGD Regression
- Decision Tree Regression
- Random Forest Regression
- ExtraTrees Regression
- Gradient Boosting Regression

- Extra Tree Regression

1ère étape

- CrossValidation à 5
- Passage à l'échelle log
- Métriques : temps, R2, Variance

2ème étape

- Suppression des outliers
- Pertinence EnergySTARScore
- Hyperparamètres via GridSearchCV



1) Présentation

2) Préparation



3) Machine Learning



Algorithme optimisé



Données générées pour les valeurs manquantes



5) Machine Learning

1ère étape : Energy



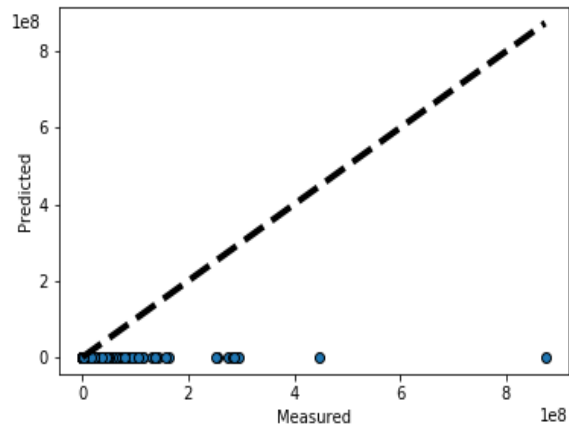
1) Présentation

2) Préparation

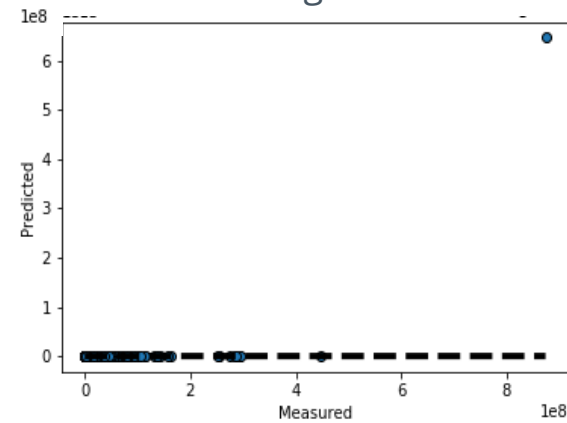


3) Machine Learning

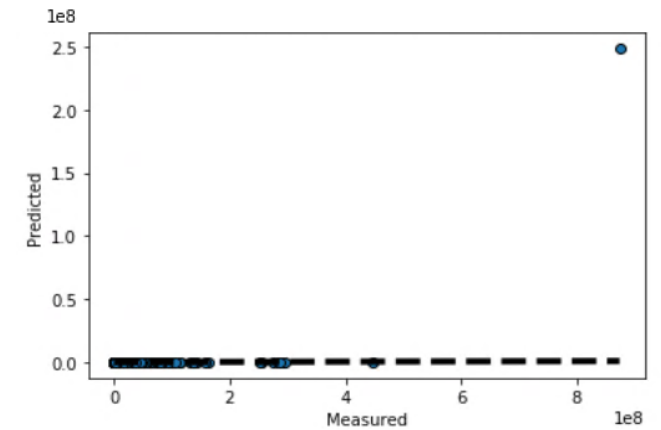
Dummy Regressor



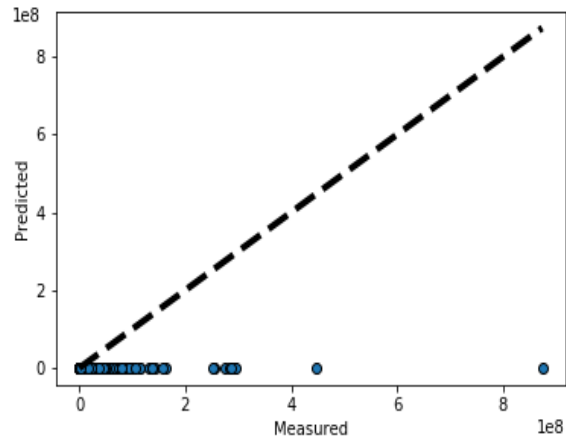
Linear Regression



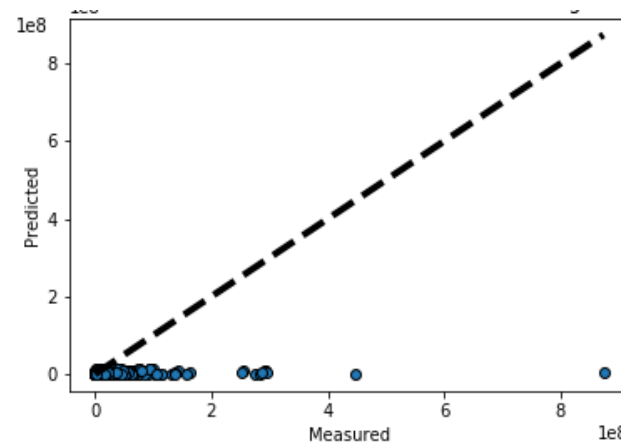
Ridge Regression



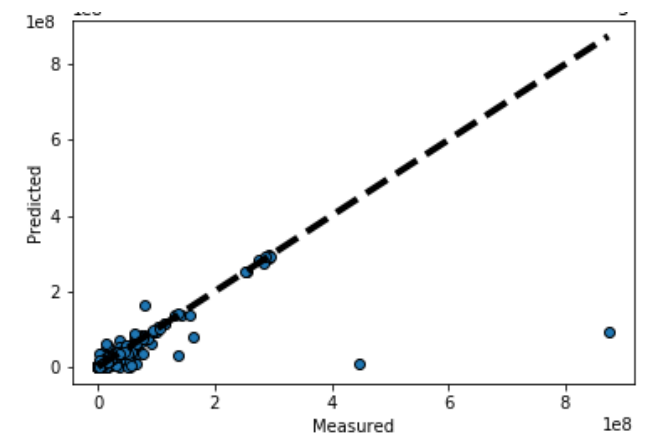
ElasticNet



SGD Regression



Decision Tree Regression





5) Machine Learning

1ère étape : Energy



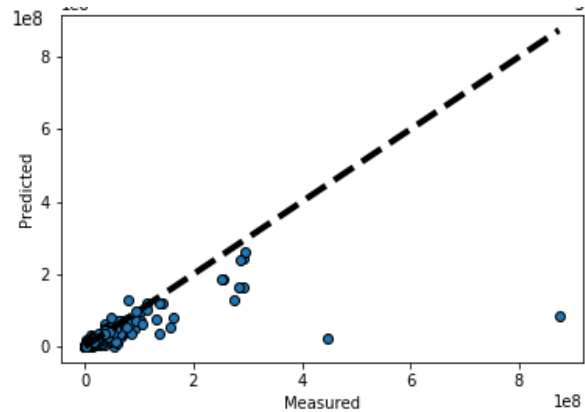
1) Présentation

2) Préparation

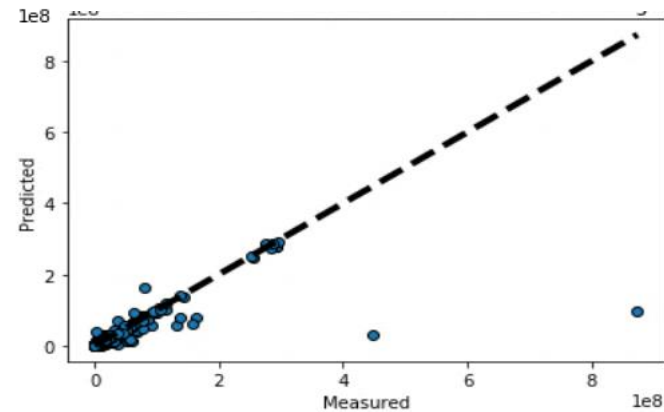


3) Machine Learning

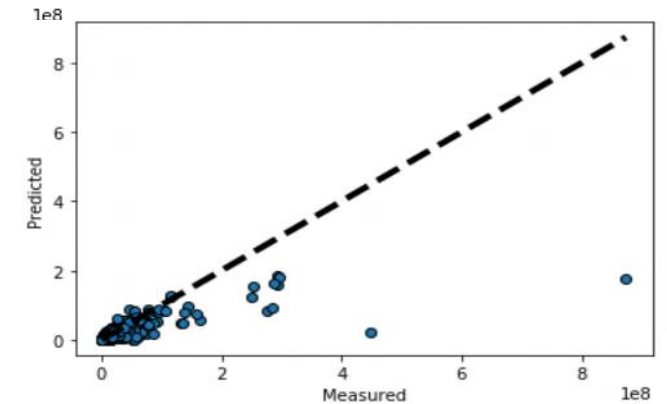
Random Forest Regression



Extra Tree Regression



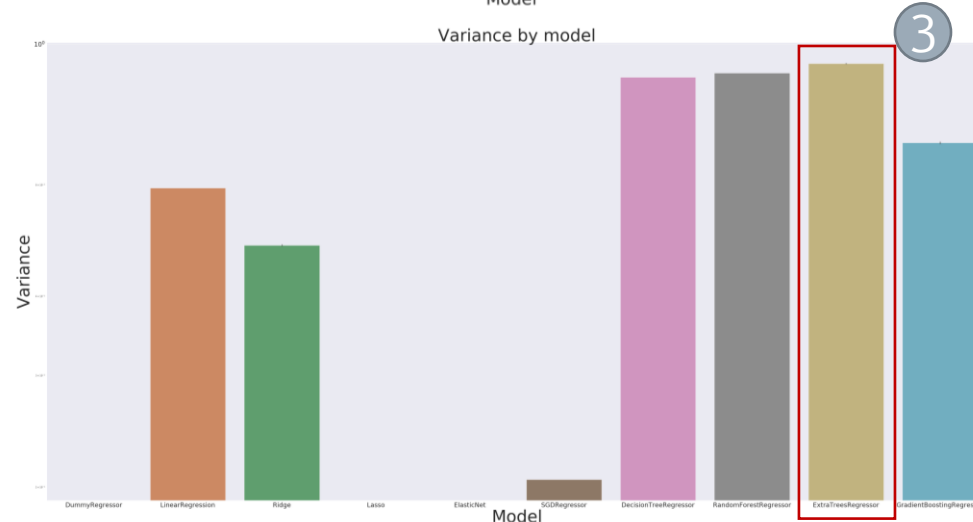
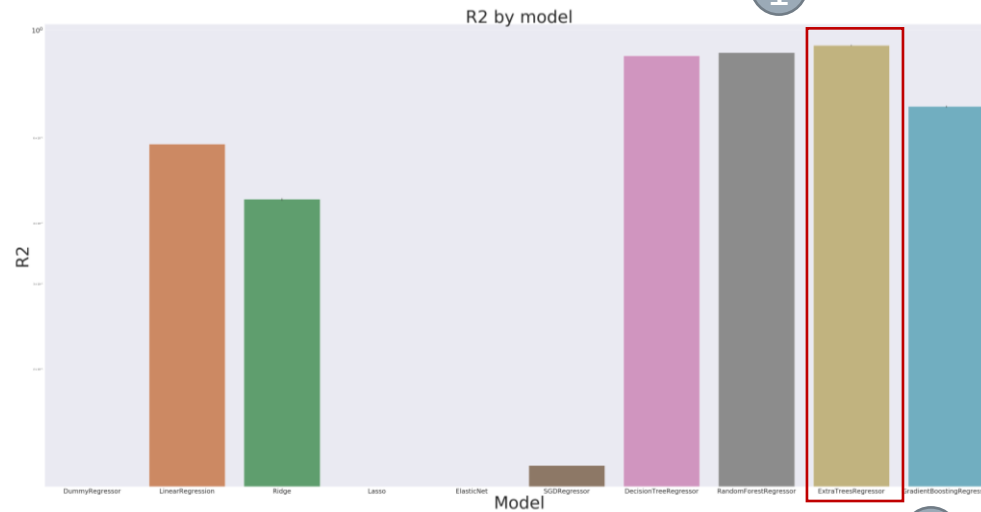
Gradient Boosting Regression





5) Machine Learning

1ère étape : Energy

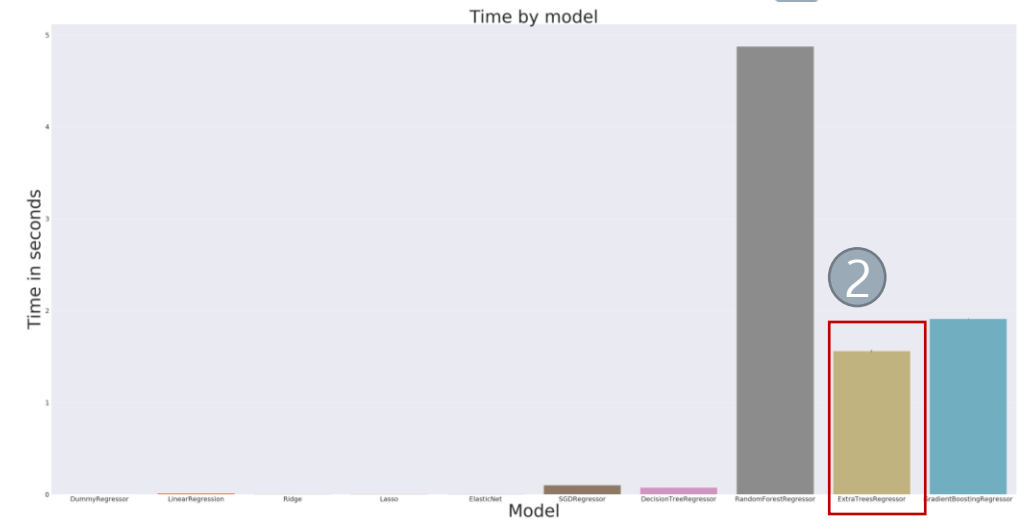


1) Présentation

2) Préparation



3) Machine Learning



ExtraTrees est sélectionné car :

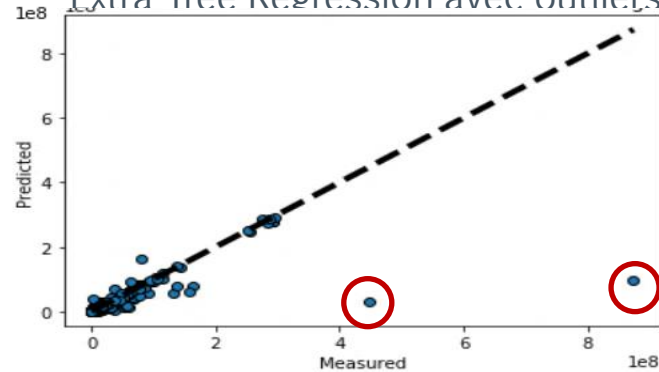
- 1) Meilleur R2 (avec données logarithmique)
- 2) Temps d'exécution bon au regard des résultats
- 3) Bonne couverture des données



5) Machine Learning

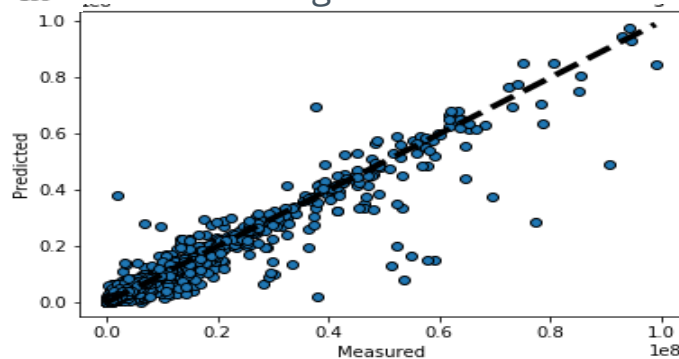
2ème étape : Energy

Extra Tree Regression avec outliers



R2 Score (no log)	SMAPE
0.6194145930865151	23.698713771324854

Extra Tree Regression sans outliers



R2 Score (no log)	SMAPE
0.9411664179150365	23.55269271584376

Identification des outliers

	OSEBuildingID	BuildingType	PrimaryPropertyType	LargestPropertyUseType	SecondLargestPropertyUseType	ThirdLargestPropertyUseType
3274	49967	Campus	College/University	College/University	NaN	NaN
35	43	Campus	Mixed Use Property	Office	Laboratory	Non-Refrigerated Warehouse
3546	276	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
170	276	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
618	828	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
3997	828	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
3936	753	NonResidential	Other	Data Center	Office	NaN
558	753	NonResidential	Other	Data Center	Office	NaN
124	198	NonResidential	Hospital	Hospital (General Medical & Surgical)	NaN	NaN
3499	198	NonResidential	Hospital	Hospital (General Medical & Surgical)	NaN	NaN
3264	49940	NonResidential	Hospital	Hospital (General Medical & Surgical)	NaN	NaN
6648	49859	Campus	Other	Other	NaN	NaN
167	268	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
3543	268	NonResidential	Hospital	Hospital (General Medical & Surgical)	Parking	NaN
3717	477	Campus	Other	Other	Parking	NaN
340	477	Campus	Other	Other	Parking	NaN
4884	22062	Campus	College/University	College/University	Parking	NaN
1494	22062	Campus	College/University	College/University	Parking	NaN
5085	23113	NonResidential	Medical Office	Medical Office	Parking	Other/Specialty Hospital
1690	23113	NonResidential	Medical Office	Medical Office	Parking	Other/Specialty Hospital



1) Présentation

2) Préparation

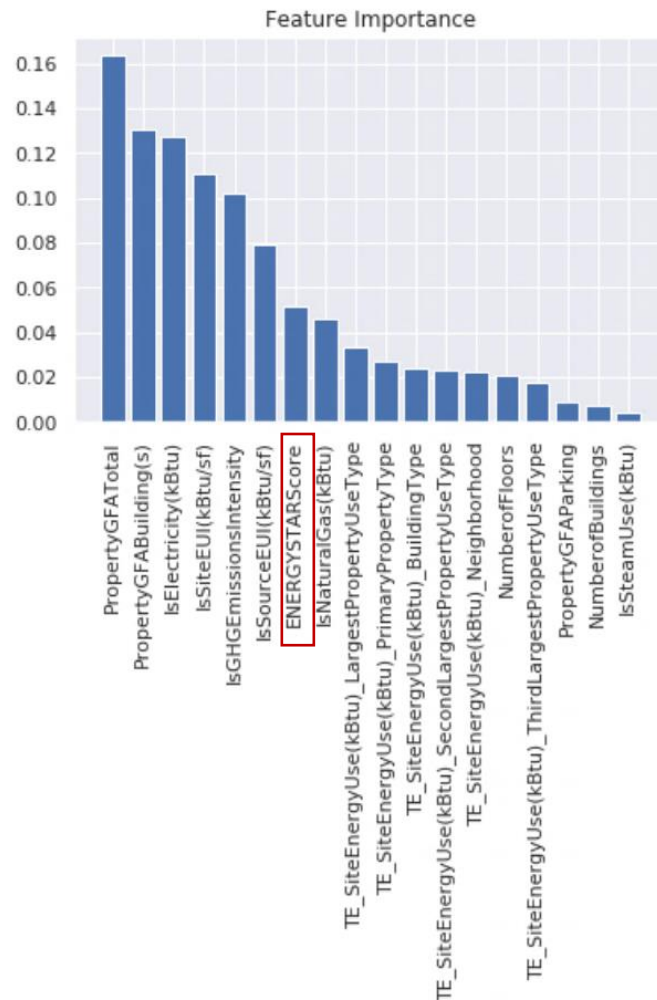


3) Machine Learning



5) Machine Learning

2ème étape : Energy



- D'un point de vue importance, l'energyScore arrive en 7ème position sur 18.. Comme présenté, sa pertinence est limitée et présente peu d'intérêt.



1) Présentation

2) Préparation



3) Machine Learning



5) Machine Learning

2ème étape : Energy : Hyperparamètres

1) Hyperparamètres

Best score : 0.9295826758300855

Best params : {'max_features': 6, 'n_estimators': 500}

2) Données estimées avec n_estimators à 1000



1) Présentation

2) Préparation

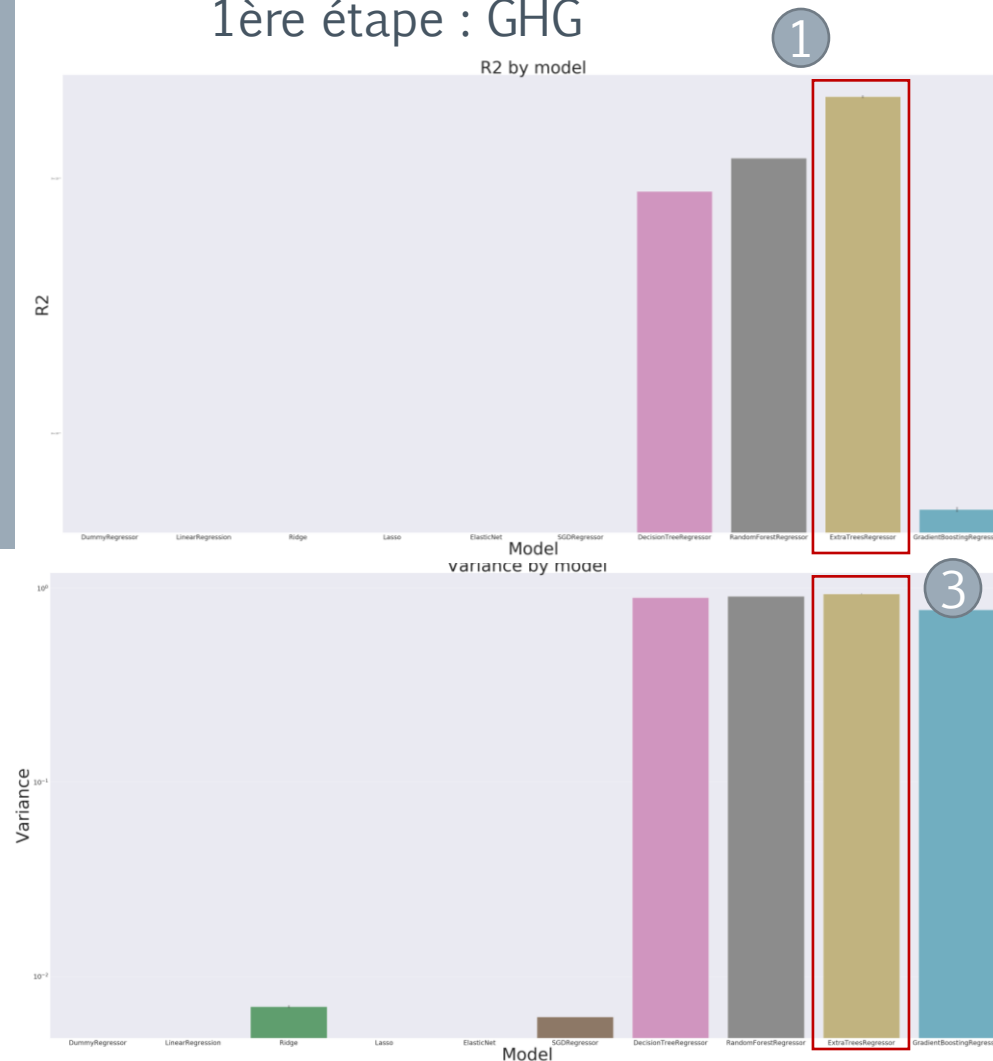


3) Machine Learning



5) Machine Learning

1ère étape : GHG

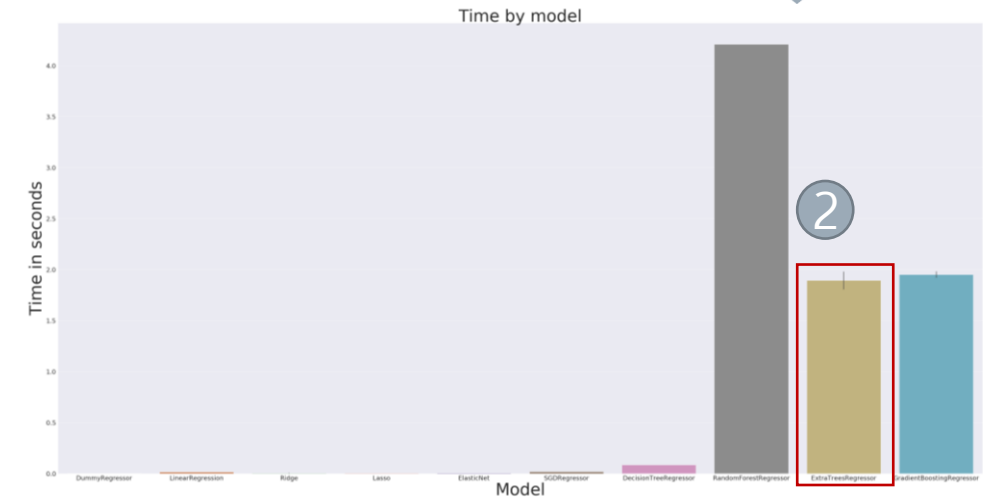


1) Présentation

2) Préparation



3) Machine Learning



ExtraTrees est sélectionné car :

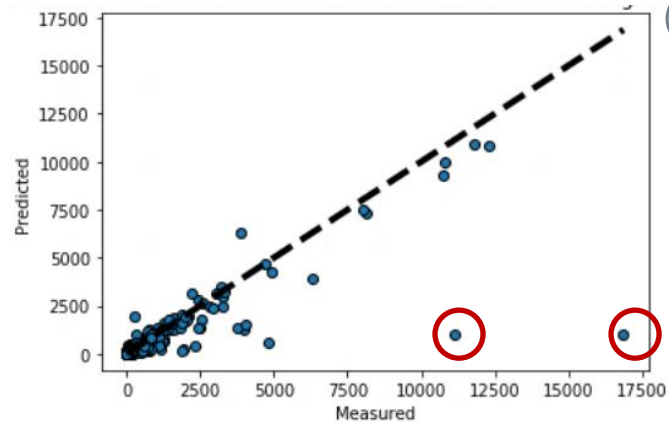
- 1) Meilleur R2 (avec données logarithmique)
- 2) Temps d'exécution bon au regard des résultats
- 3) Bonne couverture des données



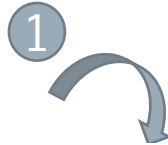
5) Machine Learning

2ème étape : GHG

Extra Tree Regression avec GHG



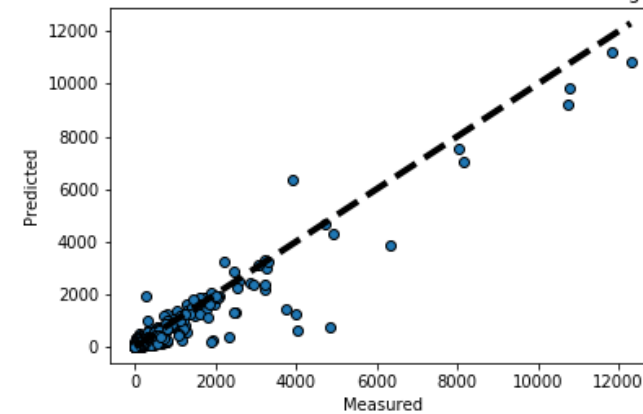
R2 Score (no log)	SMAPE
0.6224492981514331	23.70099636173587



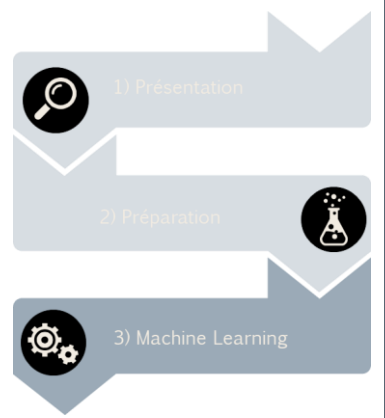
Identification des outliers

OSEBuildingID	BuildingType	PrimaryPropertyType	LargestPropertyUseType	SecondLargestPropertyUseType	ThirdLargestPropertyUseType
35	43	Campus	Mixed Use Property	Office	Laboratory
3274	49967	Campus	College/University	College/University	Non-Refrigerated Warehouse

Extra Tree Regression sans outliers



R2 Score (no log)	SMAPE
0.94213323731237	23.55720410962813



2



5) Machine Learning

2ème étape : GHG : Hyperparamètres

1 Hyperparamètres

Best score : 0.9293377567117144

Best params : {'max_features': 5, 'n_estimators': 500}

2 Données estimées avec n_estimators à 1000



1) Présentation

2) Préparation



3) Machine Learning



6) Prochaines étapes

Amélioration du modèle

- Récupérer les données des années 2017 à 2019 sur le site OpenData de la ville de Seattle.
- Intégrer une seule fois un bâtiment donné dans le modèle
- Récupérer des données complémentaires permettant de mieux gérer les outliers (Nombre de lits, nombre d'employés pour les hôpitaux, Nombre d'étudiants, nombre de professeurs pour les campus,...)



7) Environnement technique

