Projet 3

Anticipez les besoins en consommation d’énergie de bâtiments

Seattle

Loridan Adrien

student machine learning engineer

medium.com/@adrien.loridan

linkedin.com/in/adrien-loridan

github.com/loridan

Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

Présentation

La ville de Seattle a pour objectif de réduire ses émissions de carbone. Afin d'anticiper les besoins en consommation d’énergie de bâtiments, des agents ont effectués des relevés minutieux en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir !

Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

Problème

Pouvons nous réduire les côuts d'acquisition des relevés ?

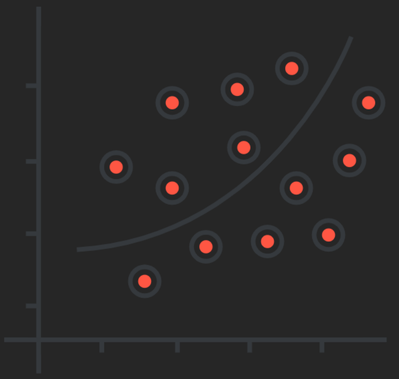
Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

Une solution envisagée par la ville...

Mettre à disposition un modèle de régression qui fournit une estimation, à partir des données déclaratives du permis d'exploitation commerciale, de la consommation d'énergie totale des bâtiments et des émissions de C02



Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

... ainsi qu'une demande spécifique

Evaluer l’intérêt de l’ENERGY STAR SCORE pour la prédiction d’émissions

Information métier sur l'ENERGYSTAR® SCORE

L'ENERGYSTAR® est un label du gouvernement des USA qui valorise l’efficacité énergétique.
L'ENERGYSTAR® SCORE d'un batiment est un score entre 0 et 100 qui est calculé à partir des divers caractéristiques du batiment et de ceux du même secteur d'activité au niveau national.

Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

Hypothèse principale

Pouvons nous estimer les émissions de CO2 et la consommation totale d’énergie des batiments non mesurées à l'aide de modèles de régression ?

Seattle

A problem without a solution is a misstated problem.

Une difficulté qu'il faut résoudre

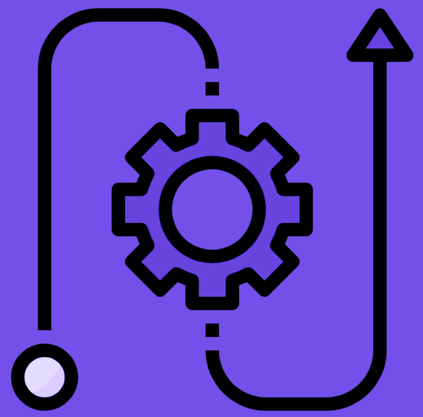
Hypothèse secondaire

L’'ENERGY STAR SCORE améliore t-il notre modèle de prédiction d’émissions ?

Seattle

La science des données

Expérimentation



1. Collecter des données
2. Préparer et explorer les données
3. Modéliser

Seattle

1

kaggle

Lien de téléchargement
kaggle.com/city-of-seattle/sea-building-energy-benchmarking

4 fichiers
2015-building-energy-benchmarking.csv (1.51 MB)
2016-building-energy-benchmarking.csv (1.18 MB)
socrata_metadata_2015-building-energy-benchmarking.json (53.34 KB)
socrata_metadata_2016-building-energy-benchmarking.json (44.73 KB)

Quelle est la réglementation sur la protection de ces données ?
Aucune, le dataset est publique d'apres Kaggle.
Faible, d'après la source officielle data.seattle.gov

Ou sont stockées les données?
Elles sont sur les serveurs distants publiques et sur notre ordinateur en sdd local

Quel est le type des données
Structurée

Seattle

2

P03_01_notebook_nettoyage.ipynb

Nettoyage et assemblage

Importer et observer les données

Supprimer les outliers identifiés par la source

Renommer les colonnes identiques

Vérifier la cohérence entre 2015 et 2016

Assembler les deux ensembles de données

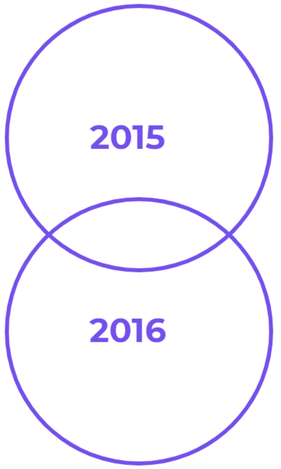
Identifier et traiter les doublons, les valeurs aberrantes

Convertir les valeurs qualitatives identiques

Supprimer les colonnes qui ne sont pas utiles ou accessibles pour le permis d'exploitation (sauf targets)

Imputation de certaines valeurs manquantes

Analyser les données



Avant (3340, 47)
(3376, 46)

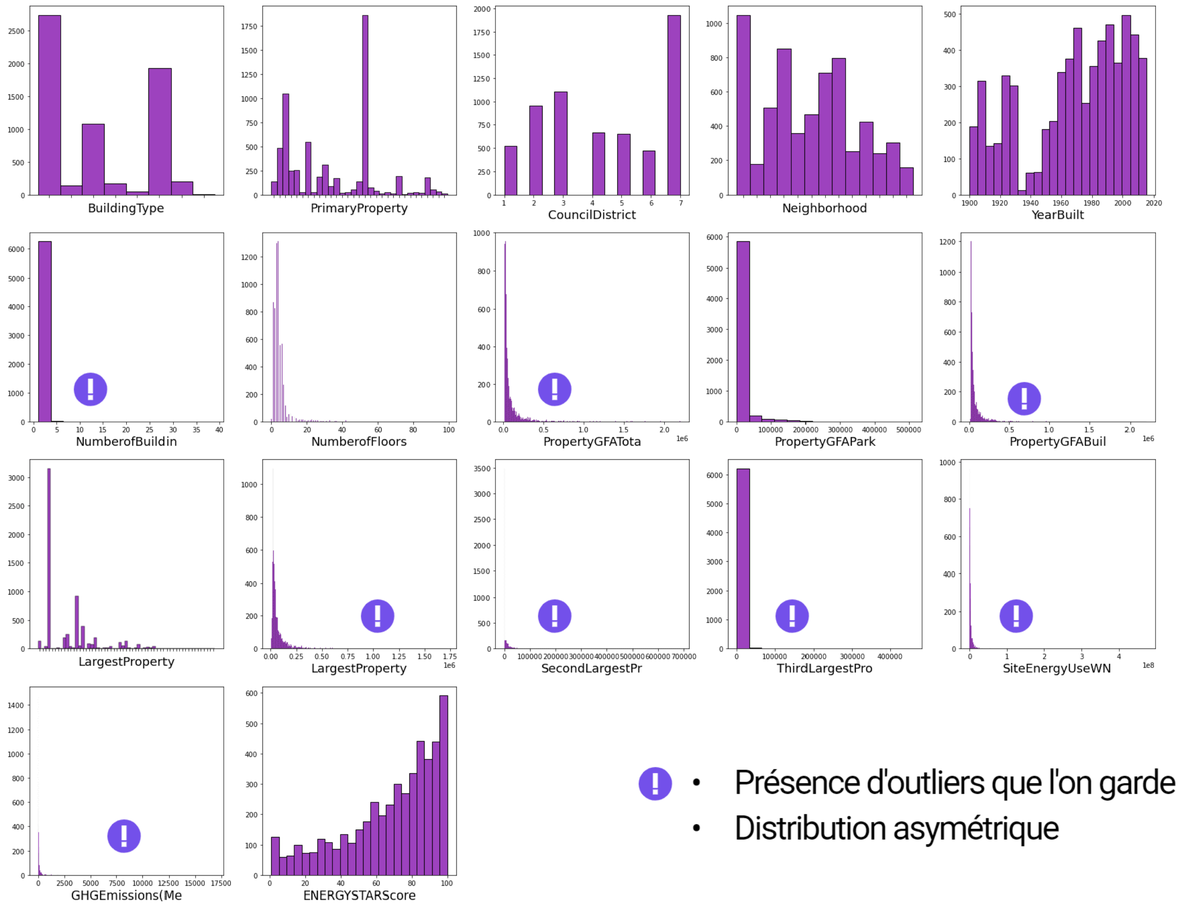
Aprés (6293, 17)

Seattle

2

P03_01_notebook_nettoyage.ipynb

Distribution des valeurs quantitatives et qualitatives avant export



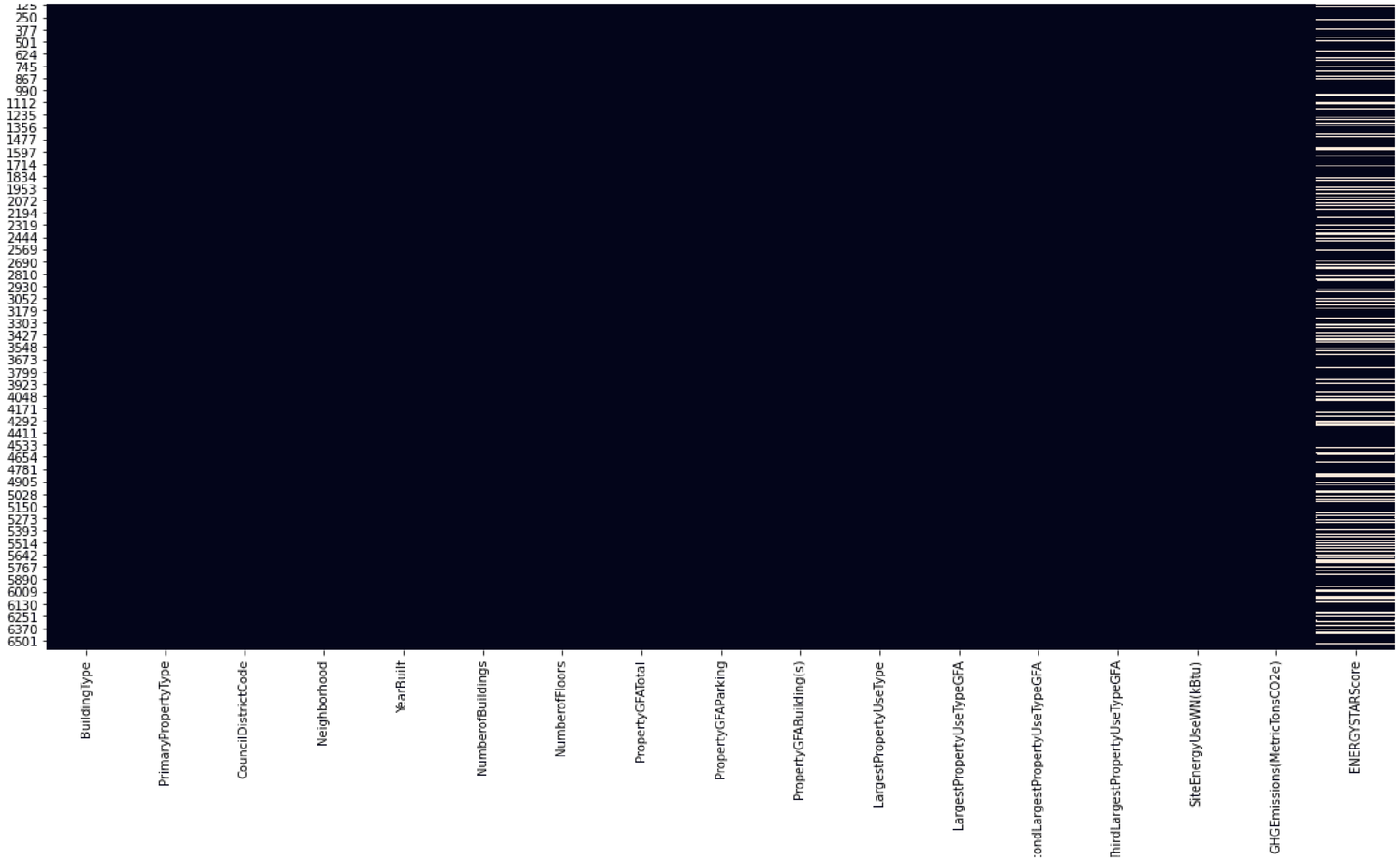
- Présence d'outliers que l'on garde
- Distribution asymétrique

Seattle

2

P03_01_notebook_nettoyage.ipynb

Valeurs manquantes avant export

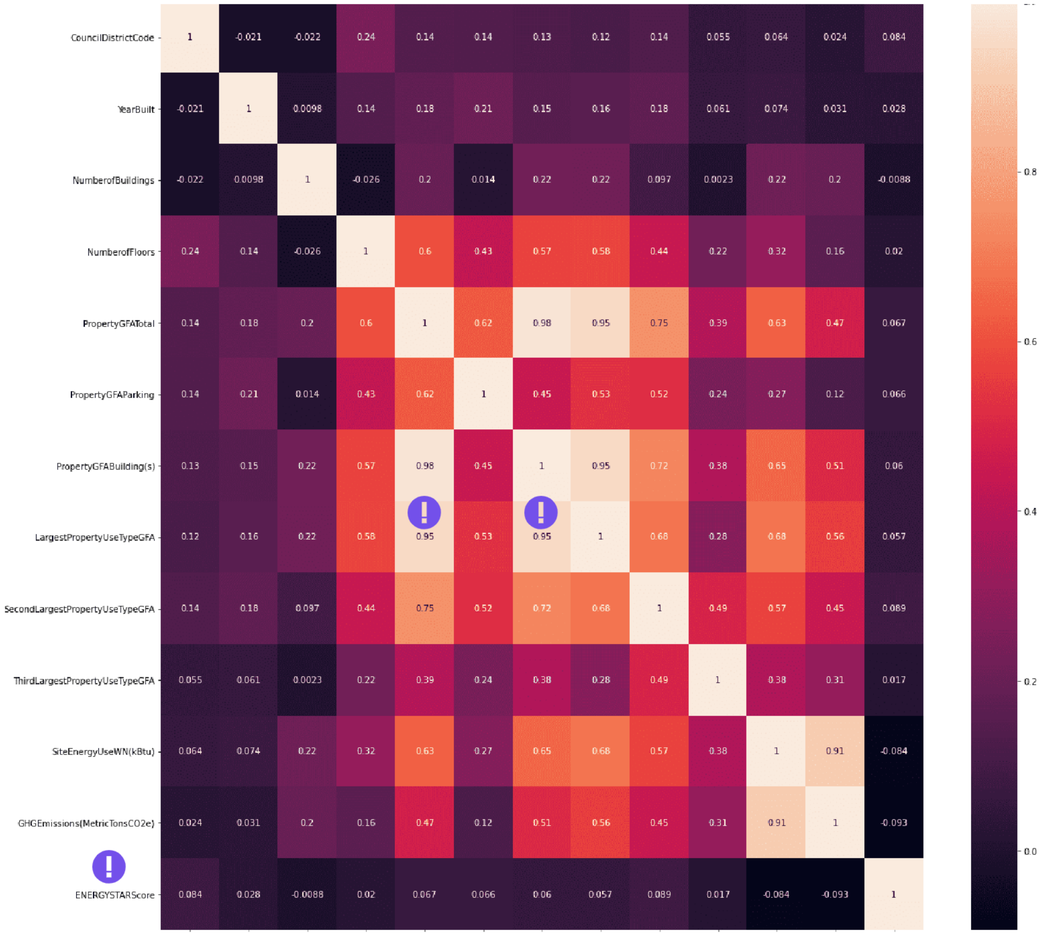


Seattle

2

P03_01_notebook_nettoyage.ipynb

Matrice de corrélation linéaire



- variables indépendantes

Forte corrélation entre :

- PropertyGFABuildings
- PropertyGFATotal
- LargestPropertyUseTypeGFA

Aucune corrélation :

- ENERGYSTARScore

Seattle

P03_02_notebook_modelisation.ipynb

Choix des algorithmes

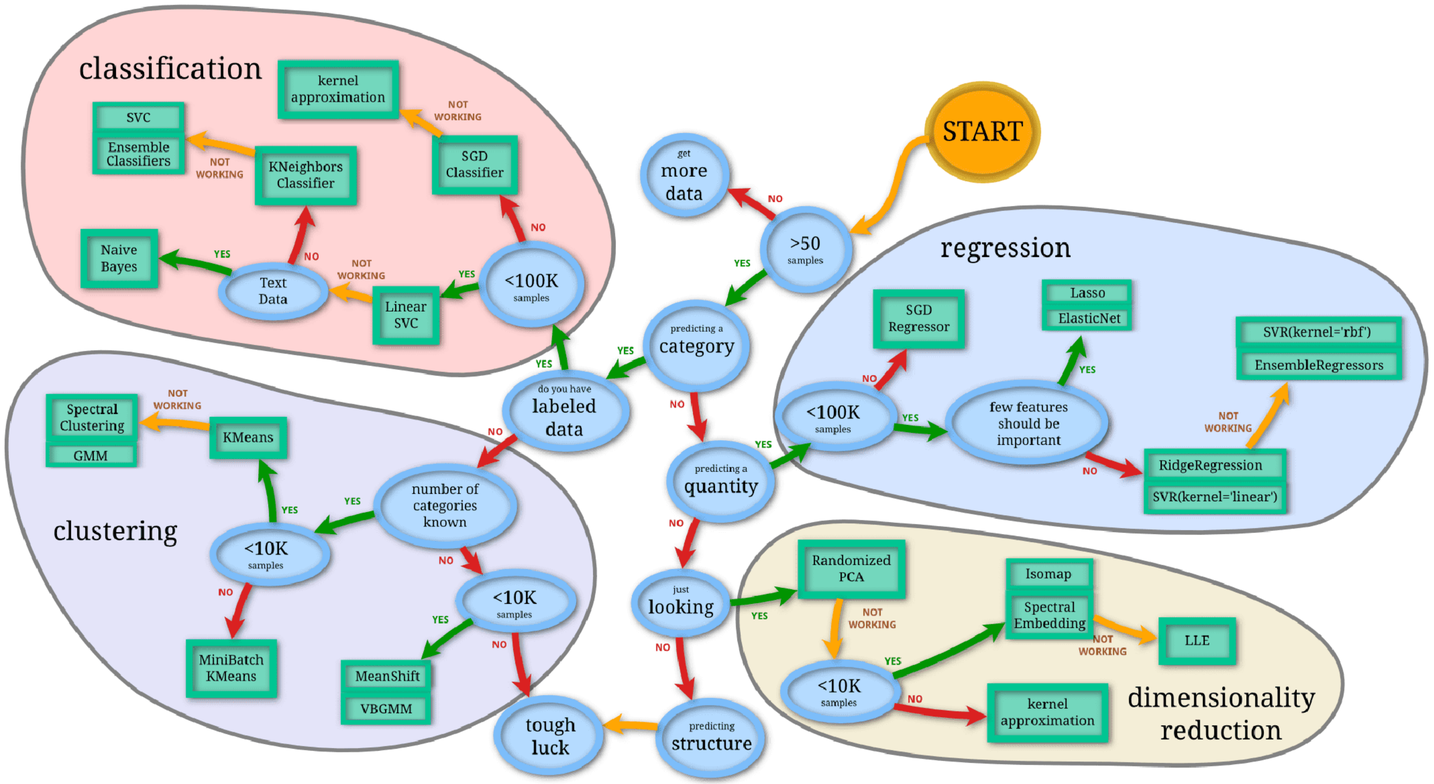
Lasso

ElasticNet

Ridge

SVR

Random Forest



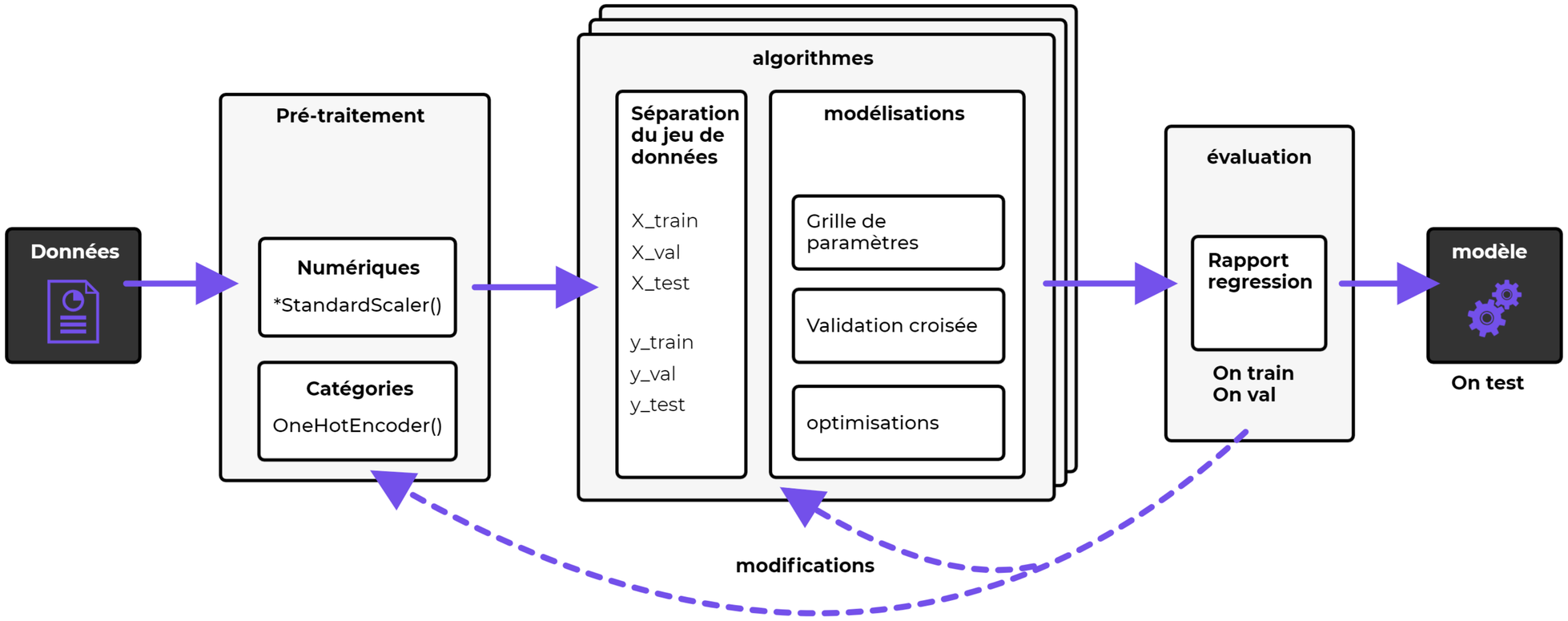
scikit learn

Seattle

3

P03_02_notebook_modelisation.ipynb

Process d'entrainement de nos modèles *(dév)*



Seattle

3

P03_02_notebook_modelisation.ipynb

Grille des paramètres d'optimisations

Ridge

alpha : np.logspace(-2, 8, 100)

Lasso

alpha : np.logspace(-2, 8, 100)
max_iter : [10000]

ElasticNet

alpha : np.logspace(-2, 8, 100)
l1_ratio : np.arange(0.1,1,0.1)
max_iter : [10000]

Support vector regression

gamma : np.logspace(-8, 10, 8)
epsilon : np.logspace(-3, 1, 5)
C : np.logspace(-3, 2, 5)

Random forest

n_estimators : [100,200,300]
max_depth : [None]
min_samples_leaf : [1,3,5,10]
max_features : ['auto', 'sqrt']

Seattle

3

P03_02_notebook_modelisation.ipynb

onehotencoder()

17 columns  55 columns

Entrée [10]:
```
X.sample(10)
```

Out[10]:

	CouncilDistrictCode	YearBuilt	NumberofBuildings	NumberofFloors	PropertyGFATotal	PropertyGFAParking	PropertyGFABuilding(s)	LargestPropertyUseTypeGF
1674	0.262295	0.960845	-0.079863	0.241297	0.085509	-0.279442	0.170509	0.04051
4066	-0.209716	-0.193660	-0.079863	0.058127	-0.402234	-0.279442	-0.386077	-0.3601
5927	1.206318	0.292447	-0.079863	-0.308214	-0.382106	-0.279442	-0.363109	-0.3357
1366	-0.681728	0.505119	-0.079863	-0.308214	-0.306945	-0.278324	-0.277631	-0.2447
5819	-0.681728	-1.712744	-0.079863	-0.125044	-0.443250	-0.279442	-0.432882	-0.4098
1284	-0.209716	1.143135	2.241507	0.241297	-0.194886	-0.279442	-0.149464	-0.1091
3109	1.206318	1.416570	-0.079863	0.058127	-0.143985	-0.279442	-0.091378	-0.2318
2752	1.206318	0.505119	-0.079863	-0.125044	-0.445935	-0.279442	-0.435946	-0.4100
2275	-0.681728	0.019012	-0.079863	3.355191	0.371833	-0.279442	0.497246	0.1497
1381	-1.625750	0.292447	-0.079863	-0.674554	-0.462119	-0.279442	-0.454415	-0.4326

10 rows × 55 columns

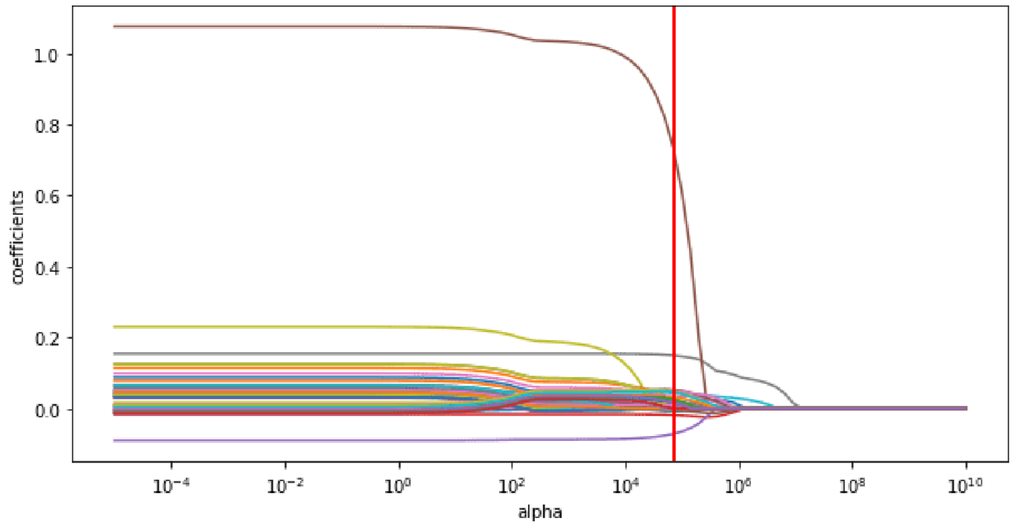
Seattle

3

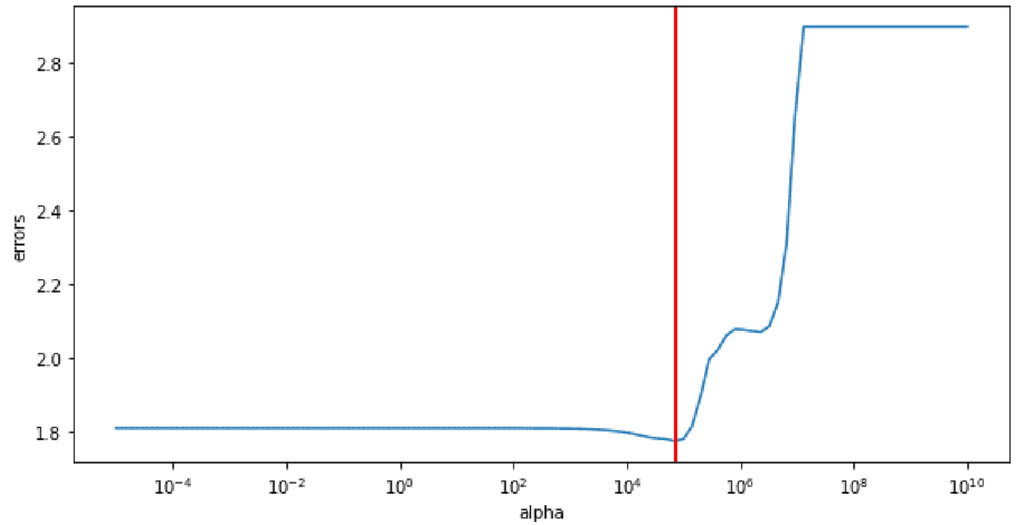
P03_02_notebook_modelisation.ipynb

Lasso régression de SiteEnergyUseWN(kBtu)

Chemin de régularisation



Erreurs en fonction de la pénalité



Seattle

3

Rapport de régression pour la prédiction de SiteEnergyUseWN(kBtu)

	Min erreurs relatives	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_train	0	0.006	0.03	0.071	0.14	0.53	57	2.7e+06	0.97
linear_regression_on_train	0	0.051	0.25	0.57	1.1	3	3.2e+02	8.9e+06	0.69
gscv_ridge_regression_on_train	0	0.052	0.25	0.58	1.1	3	3.1e+02	8.9e+06	0.69
gscv_lasso_regression_on_train	0.001	0.052	0.24	0.52	0.97	2.8	2.9e+02	8.9e+06	0.69
gscv_elasticnet_regression_on_train	0	0.056	0.27	0.6	1.2	3.2	2.8e+02	9e+06	0.69
dummy_regression_on_train	0	0.13	0.62	1.7	4.3	8.6	4.5e+02	1.6e+07	0
gscv_svr_on_train	0	0.064	0.32	0.66	0.97	2.5	1.7e+02	1.6e+07	-0.042

On Train

	Min erreurs relatives	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_val	0.002	0.015	0.078	0.18	0.38	1.1	14	3.8e+06	0.87
gscv_elasticnet_regression_on_val	0.001	0.038	0.27	0.63	1.2	3.1	39	6.4e+06	0.64
gscv_ridge_regression_on_val	0.004	0.048	0.26	0.62	1.2	3.2	41	6.8e+06	0.59
gscv_lasso_regression_on_val	0.001	0.046	0.26	0.55	1	2.9	41	7e+06	0.57
linear_regression_on_val	0.002	0.052	0.27	0.61	1.1	3.2	42	7.1e+06	0.56
dummy_regression_on_val	0.015	0.13	0.6	1.6	4.1	7.7	55	1.1e+07	-0.001
gscv_svr_on_val	0.001	0.054	0.33	0.7	0.95	2.2	19	1.1e+07	-0.11

On Val

Seattle

3

Rapport de régression pour la prédiction de GHGEmissions(MetricTonsCO2e)

	Min erreurs relatives	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_train	0	0.012	0.059	0.14	0.37	1.6	2.2e+02	1.1e+02	0.95
linear_regression_on_train	0	0.097	0.5	1.2	3.5	12	7.1e+02	2.7e+02	0.69
gscv_ridge_regression_on_train	0.001	0.096	0.5	1.3	3.7	12	7.9e+02	2.8e+02	0.69
gscv_lasso_regression_on_train	0	0.089	0.47	1.1	2.8	10	1.1e+03	2.8e+02	0.68
gscv_elasticnet_regression_on_train	0.001	0.1	0.52	1.3	3.8	12	8.8e+02	2.8e+02	0.68
gscv_svr_on_train	0	0.056	0.31	0.62	1.1	4	4.4e+02	4.5e+02	0.16
dummy_regression_on_train	0	0.14	0.65	2.4	11	28	1.4e+03	4.9e+02	0

On Train

	Min erreurs relatives	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_val	0.001	0.026	0.17	0.39	0.89	3.6	46	1.6e+02	0.74
gscv_elasticnet_regression_on_val	0.002	0.078	0.49	1.2	3.5	11	1e+02	2.4e+02	0.42
gscv_ridge_regression_on_val	0	0.071	0.45	1.2	3.4	10	1.1e+02	2.5e+02	0.35
gscv_lasso_regression_on_val	0.002	0.065	0.42	0.94	2.8	9.5	89	2.5e+02	0.35
linear_regression_on_val	0.001	0.078	0.48	1.2	3.3	9.8	1.2e+02	2.7e+02	0.26
gscv_svr_on_val	0.006	0.085	0.34	0.65	1.1	4.3	49	2.8e+02	0.23
dummy_regression_on_val	0.003	0.16	0.62	2	9.6	27	1.4e+02	3.2e+02	-0.002

On Val

Seattle

3

Rapport de régression pour la prédiction de SiteEnergyUseWN(kBtu) avec ENERGYSTARScore

	Min	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score	
gscv_rfr_on_train	0	0.004	0.022	0.05	0.1	0.31	6.9	3.2e+06	0.96	On Train
linear_regression_on_train	0	0.051	0.26	0.61	1.3	3.2	45	9e+06	0.7	
gscv_ridge_regression_on_train	0	0.058	0.27	0.67	1.4	3.5	35	9.1e+06	0.69	
gscv_elasticnet_regression_on_train	0	0.056	0.27	0.67	1.4	3.5	35	9.1e+06	0.69	
gscv_lasso_regression_on_train	0	0.057	0.28	0.62	1.3	3	38	9.1e+06	0.69	
dummy_regression_on_train	0.001	0.14	0.62	1.7	4	7.5	62	1.6e+07	0	
gscv_svr_on_train	0	0.061	0.31	0.65	0.94	2.2	22	1.7e+07	-0.037	

Le cleaning dropna() pour ENERGYSTARScore a un effet sur les données / 4802 rows

	Min	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score	
gscv_rfr_on_val	0.001	0.017	0.068	0.15	0.29	0.69	3.5	2.4e+06	0.93	On Val
gscv_lasso_regression_on_val	0.002	0.046	0.26	0.57	1.3	2.9	15	5.9e+06	0.56	
gscv_ridge_regression_on_val	0.001	0.05	0.26	0.56	1.4	3.3	18	5.9e+06	0.56	
gscv_elasticnet_regression_on_val	0	0.055	0.26	0.56	1.4	3.3	18	5.9e+06	0.56	
linear_regression_on_val	0	0.051	0.24	0.53	1.2	3.2	22	6.1e+06	0.54	
dummy_regression_on_val	0.002	0.13	0.66	1.5	3.6	7.6	27	8.9e+06	-0	
gscv_svr_on_val	0.005	0.054	0.27	0.59	0.92	2.2	9.5	9.4e+06	-0.12	

Seattle

3

Rapport de régression pour la prédiction de SiteEnergyUseWN(kBtu) sans ENERGYSTARScore mais même cleaning

	Min	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_train	0	0.005	0.026	0.058	0.12	0.36	10	3.1e+06	0.96
linear_regression_on_train	0.001	0.048	0.24	0.55	1.1	2.5	47	9.1e+06	0.69
gscv_ridge_regression_on_train	0	0.048	0.26	0.57	1.1	2.6	39	9.2e+06	0.69
gscv_elasticnet_regression_on_train	0	0.05	0.26	0.58	1.1	2.6	39	9.2e+06	0.69
gscv_lasso_regression_on_train	0	0.048	0.22	0.47	0.84	2.2	41	9.3e+06	0.68
dummy_regression_on_train	0.001	0.14	0.62	1.7	4	7.5	62	1.6e+07	0
gscv_svr_on_train	0	0.062	0.31	0.65	0.94	2.2	22	1.7e+07	-0.037

On Train

	Min	5%	25%	50%	75%	95%	Max	root_mean_squared_error	r2_score
gscv_rfr_on_val	0	0.013	0.065	0.15	0.31	0.73	5	2.6e+06	0.92
gscv_lasso_regression_on_val	0.003	0.039	0.21	0.44	0.76	1.8	14	5.9e+06	0.56
gscv_ridge_regression_on_val	0.002	0.051	0.23	0.48	0.92	2.5	17	5.9e+06	0.56
gscv_elasticnet_regression_on_val	0.003	0.05	0.24	0.49	0.92	2.5	16	5.9e+06	0.56
linear_regression_on_val	0.002	0.048	0.23	0.47	0.85	2.3	20	6.1e+06	0.54
dummy_regression_on_val	0.002	0.13	0.66	1.5	3.6	7.6	27	8.9e+06	-0
gscv_svr_on_val	0.003	0.055	0.27	0.59	0.92	2.2	9.5	9.4e+06	-0.12

On Val

Seattle

3

P03_02_notebook_modelisation.ipynb

intérêt de ENERGYSTARScore

Le rapport de regression montre que l'ENERGYScore est quasi sans intéret sur la performance. Une légere modification dans la distribution des erreurs relatives.

Cette comparaison a mis en avant l'intêret du cleaning effectué pour ce test, une meilleure performance sur le random forest. Il faut analyser en profondeur à nouveau les données pour mieux comprendre si elles sont représentatives.

Seattle

3

P03_02_notebook_modelisation.ipynb

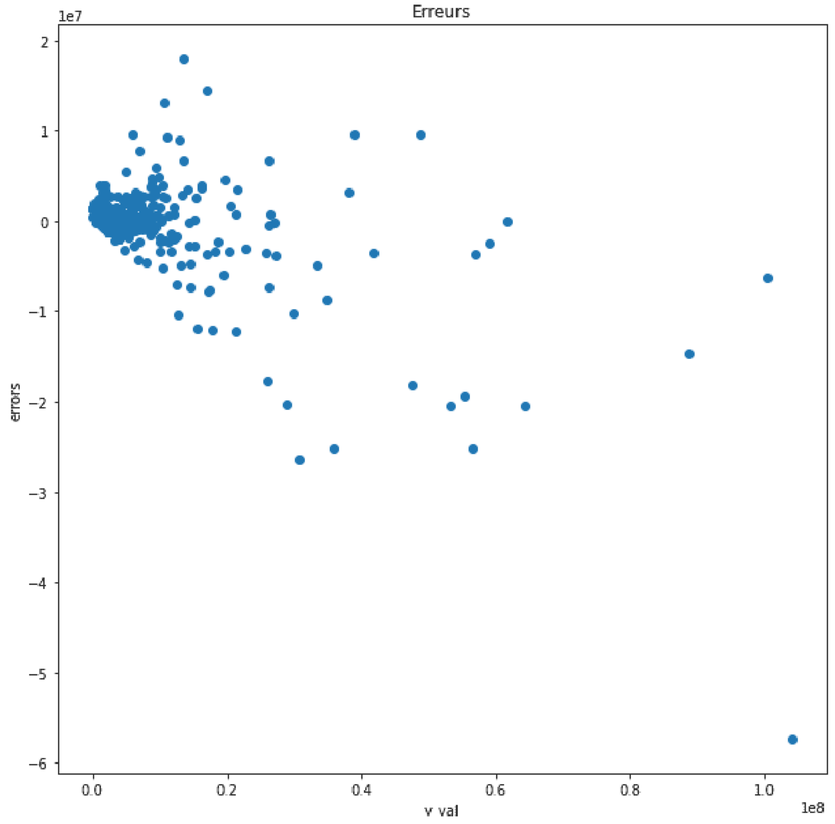
modèle sélectionné	**performance**	**contraintes**
RandomForestRegressor()	RMSE	inconnus

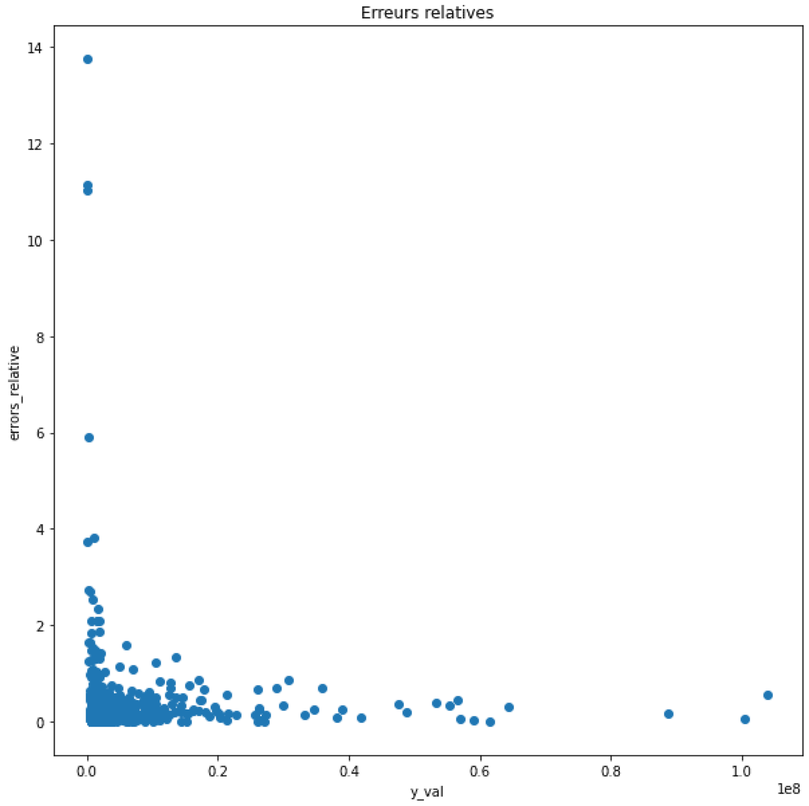
Seattle

3

P03_02_notebook_modelisation.ipynb

Analyse des erreurs SiteEnergyUseWN(kBtu)





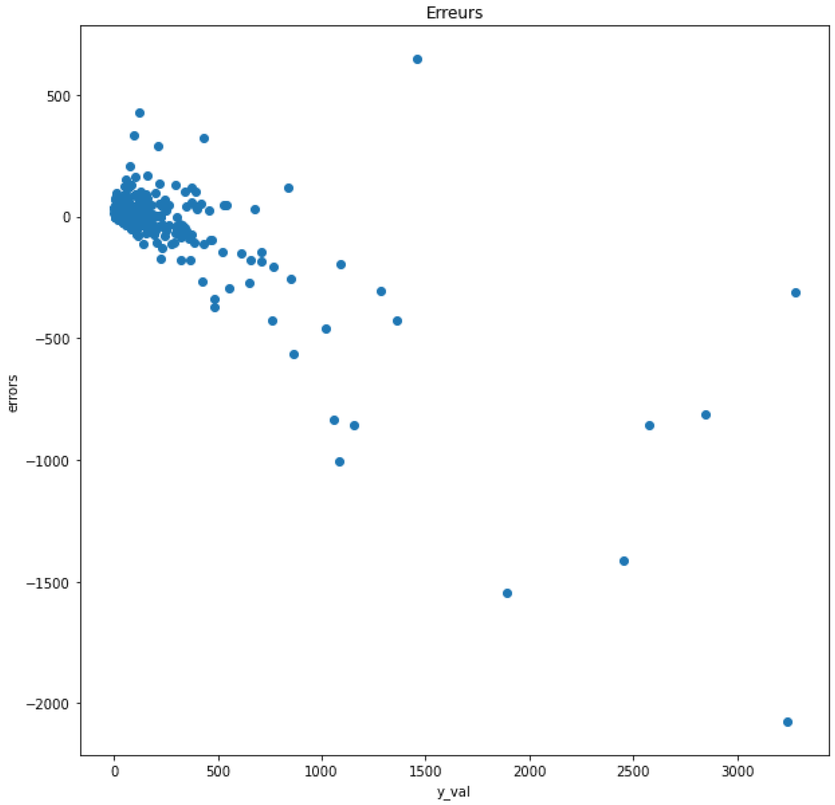
On validation

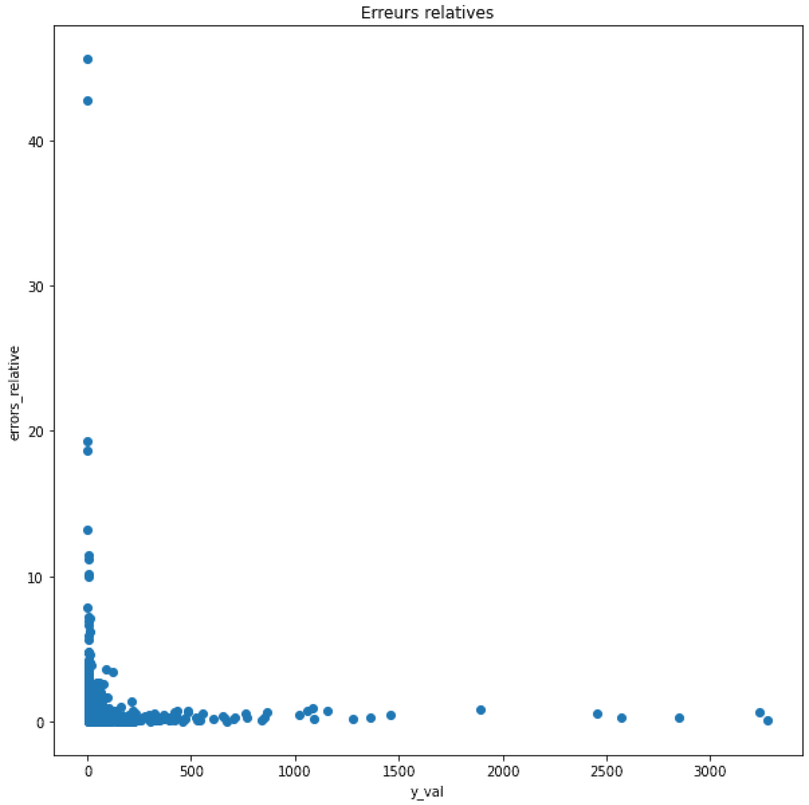
Seattle

3

P03_02_notebook_modelisation.ipynb

Analyse des erreurs GHGEmissions(MetricTonsCO2e)





On validation

Seattle

3

P03_02_notebook_modelisation.ipynb

a) Optimisation : selection variable

Pour peut être ameliorer la performance et réduire le temps de calcul, nous pouvons utiliser un algorithme de selection de variable dit **RFE** (Recursive Feature Elimination)

Les variables non informatives peuvent avoir un effet néfaste sur la précision du modèle. [...]

Des simulations numériques confirment d'une part les résultats theoriques et indiquent d'autre part que l'algorithme RFE tend a selectionner un faible nombre de variables avec une bonne erreur de prédiction.[...]

Le choix des variables prédictives ne peut se faire uniquement grace à la seule évaluation de la mesure d'importance [...]

RFE [...] corrige les effets de la corrélation

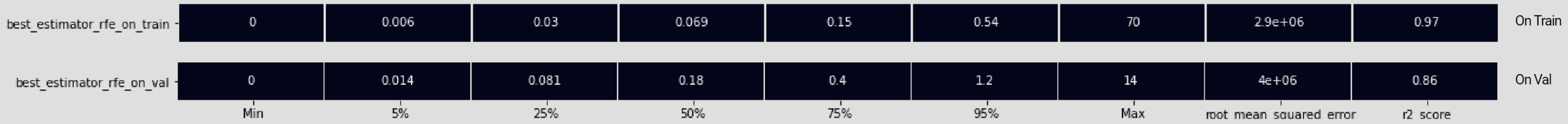
B.Gregorutti, B.Michel, P Saint-Pierre - Correlation et importance des variables dans les forets aléatoires
Univ. Pierre et Marie Curie (papersjds14.sfds.asso.fr/submission_48.pdf)

Seattle

3

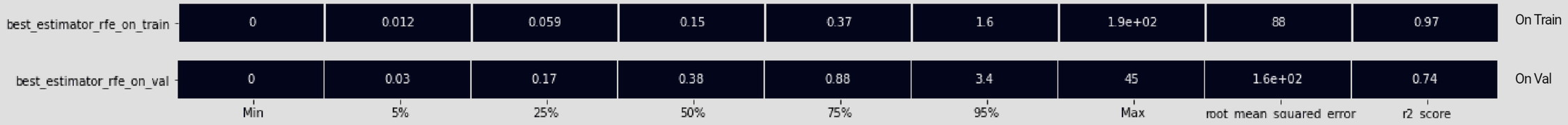
Rapport de régression avec RFE pour la prédiction de SiteEnergyUseWN(kBtu)

Optimum number of features: 44



Rapport de régression avec RFE pour la prédiction de GHGEmissions(MetricTonsCO2e)

Optimum number of features: 46



Seattle

P03_02_notebook_modelisation.ipynb

non réalisé

b) Optimisation : spécificité

Le meilleur modèle "overfit" durant la cross validation, ce qui s'aperçoit également sur notre jeu de validation, nous devons comprendre en profondeur l'algorithme afin de tester des modifications. Ici la profondeur des arbres n'a pas été testée par manque de temps d'entrainement et la stratégie de comparaison des modéles.

max_depth represents the depth of each tree in the forest. The deeper the tree, the more splits it has and it captures more information about the data. [...]

We see that our model overfits for large depth values.

Mohtadi Ben Fraj - In Depth: Parameter tuning for Random Forest
Article
medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d

P03_02_notebook_modelisation.ipynb

non réalisé

c) Optimisation : splitting strategy

Un changement dans le random_state ne devrait pas faire varier drastiquement la performance.

Reduce imbalance/randomness in data split
One of the ways to reduce the effect of a random split on your data is to make sure the split does not affect the composition of the data too much.

Ler Wei Han - Manipulating machine learning results with random state
Article
towardsdatascience.com/manipulating-machine-learning-results-with-random-state-2a6f49b31081

Projet 3

Thank you !

Seattle