Anticipez les besoins en consommation électrique de bâtiments

...

Juillet 2019

Antoine LEPAGE - yop1001@gmail.com

Démarche

Contexte :

- La ville de Seattle s'intéresse aux émissions des bâtiments non destinés à l'habitation pour atteindre son objectif de ville neutre.
- De coûteux relevés ont été effectués dans de nombreux bâtiments de la ville.

Objectif :

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

Données:

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

Démarche :

- Nettoyage
- Exploration
- Préprocessing
- Recherche de la meilleure modélisation

Nettoyage

...

Nettoyage des données

Description	Nombre d'immeubles	Nombre de variables
Data Source	3 340	42
Filtre sur les bâtiments non résidentiels	1 529	|
Choix des variables	|	20
7 Valeurs manquantes pour *numberofFloors* => Estimations à partir du type de batiment (mediane de la classe)	|	|
1 valeur manquante pour les données Energétique et rejet C02 => Suppression du bâtiment	1 528	|
Regroupement des 24 types de bâtiments (15 types au final) => harmonisation de la casse => regroupement selon les "comportements" identiques	|	21
Segmentation d'un dictionnaire pour récupérer la latitude et la longitude	|	23
Transformation de la date de construction en âge du bâtiment	|	24

Nettoyage des données

Description	Nombre d'immeubles	Nombre de variables
Suppression de la variable *NumberofBuildings* qui n'est pas assez variée		23
Suppression de la variable *CouncilDistrictCode* qui est redondante avec *Neighborhood*		22
Modification des 99 étages d'une église chinoise en 1 étage		
Modification de NumberofFloors de 5 batiments initiallement à 0 étage		
Suppression de 2 bâtiments avec des valeurs =0 pour les variables consommation énergie et rejet CO2	1526	
Suppression de 4 outliers avant la modélisation	1522	22

Premier choix des variables

- Métadonnées :
  - OSEBuildingID
  - PropertyName :
  - ComplianceStatus
  - Outlier

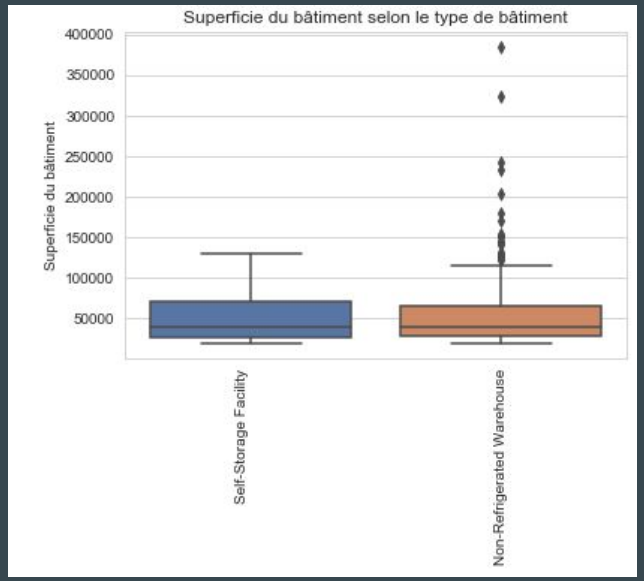
- Variables qualitatives :
  - PrimaryPropertyType
  - Location :
  - CouncilDistrictCode
  - Neighborhood

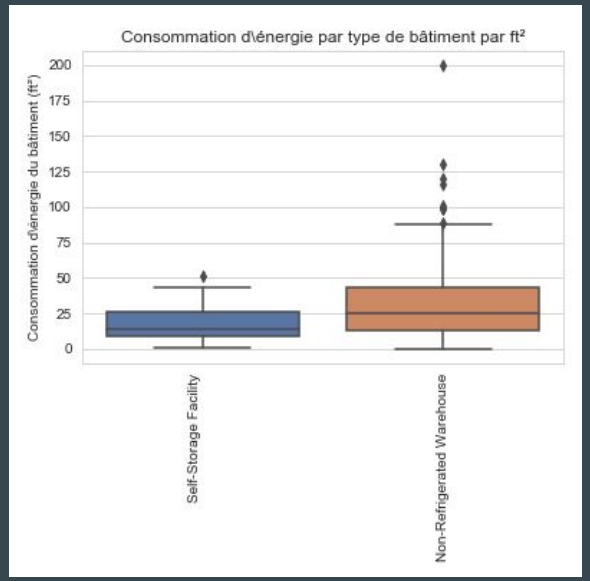
- Variables quantitatives :
  - NumberofFloors : nombre d'étages
  - PropertyGFATotal : surface de plancher Total
  - PropertyGFAParking : surface de plancher parking
  - PropertyGFABuilding(s) : surface de plancher immeuble
  - YearBuilt
  - NumberofBuildings

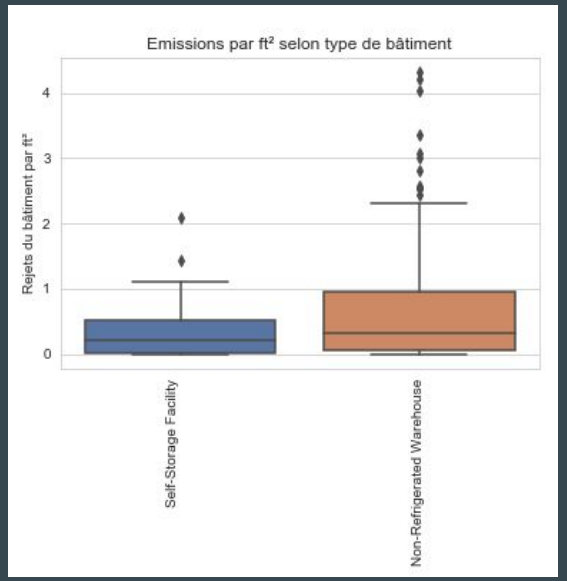
- Energy Star :
  - YearsENERGYSTARCertified
  - ENERGYSTARScore

- Variables quantitatives à expliquer :
  - SiteEnergyUse(kBtu) :
  - SiteEUI(kBtu/sf) :
  - GHGEmissions(MetricTonsCO2e)
  - GHGEmissionsIntensity(kgCO2e/ft2)

Regroupement des types de bâtiments



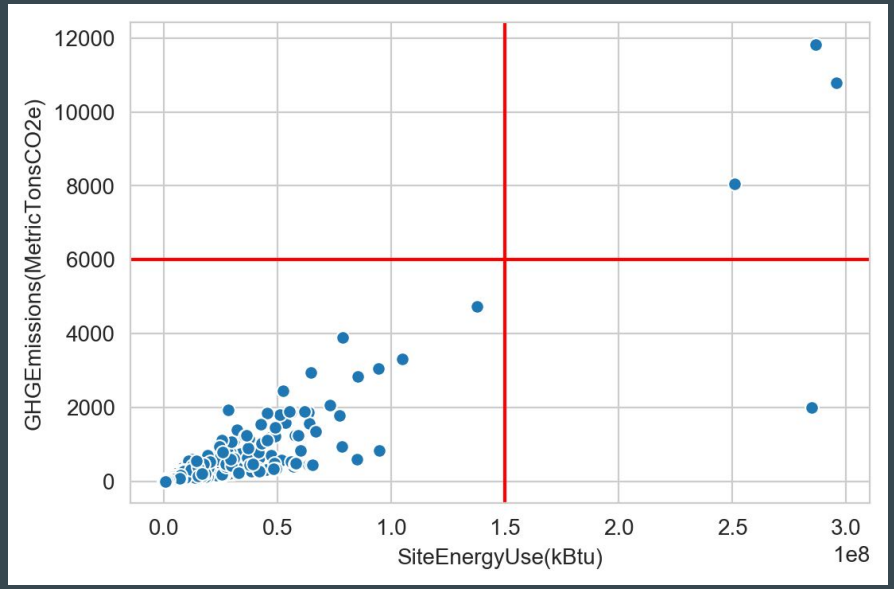




- Regroupements :
  - Self-Storage Facility & Non-Refrigerated Warehouse
  - K-12 School & College/University
  - Worship Facility' & Residence Hall/Dormitory

Suppression des 4 outliers

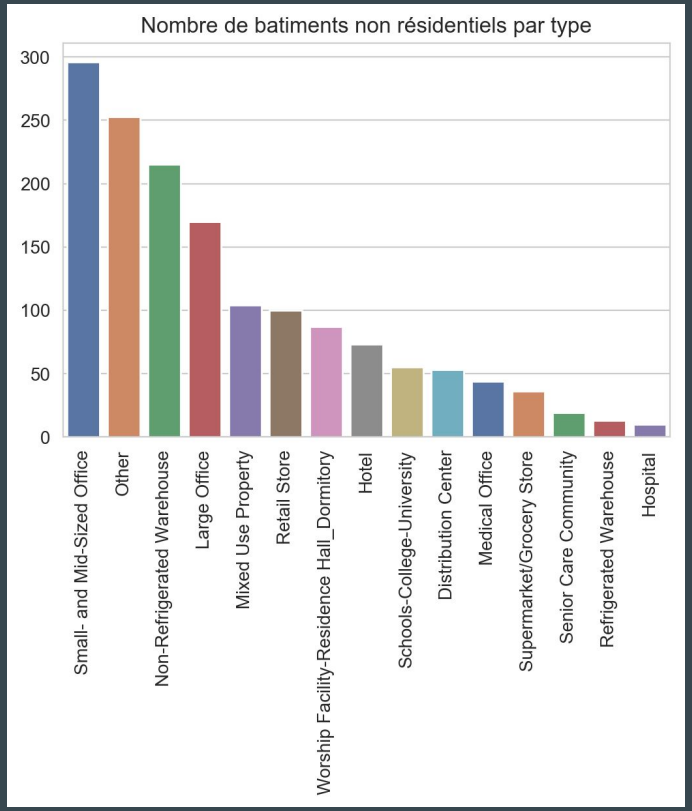
- Les algorithmes de modélisation de données peuvent être très sensibles aux outliers
  - GHGEmissions(MetricTonsCO2e) > 6000
  - SiteEnergyUse(kBtu) > 150 000 000



Exploration

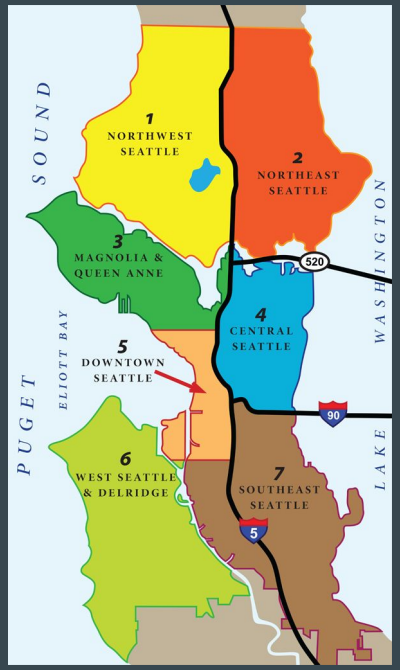
...

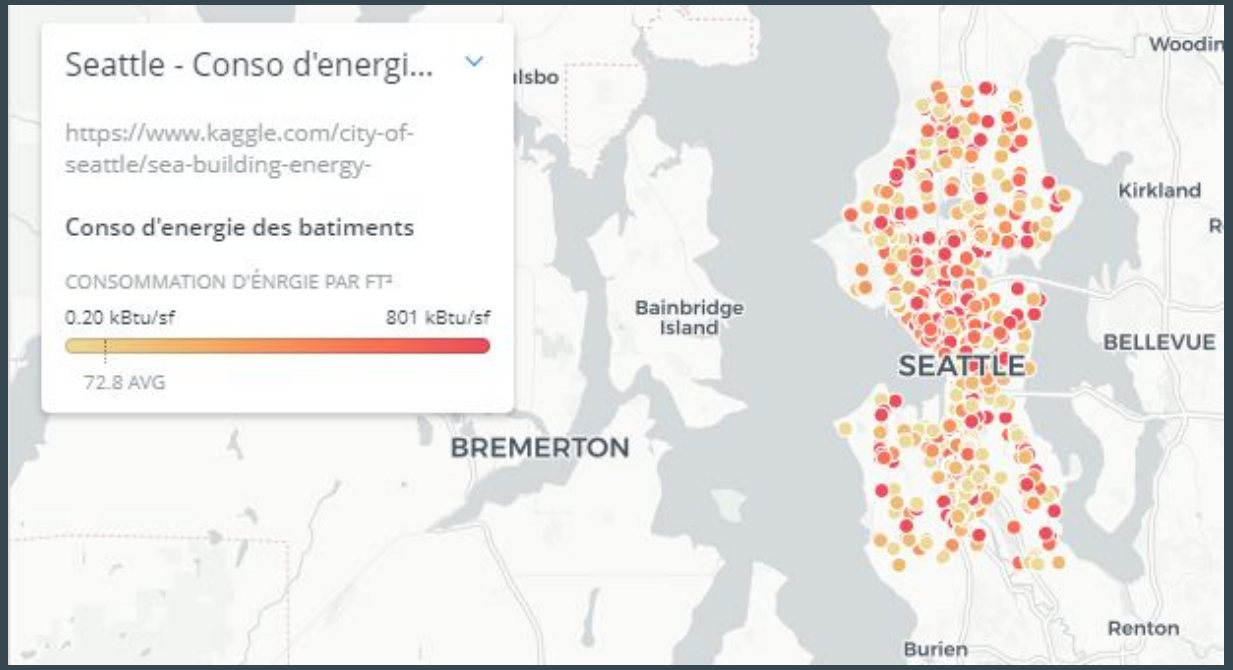
1 522 bâtiments non résidentiels



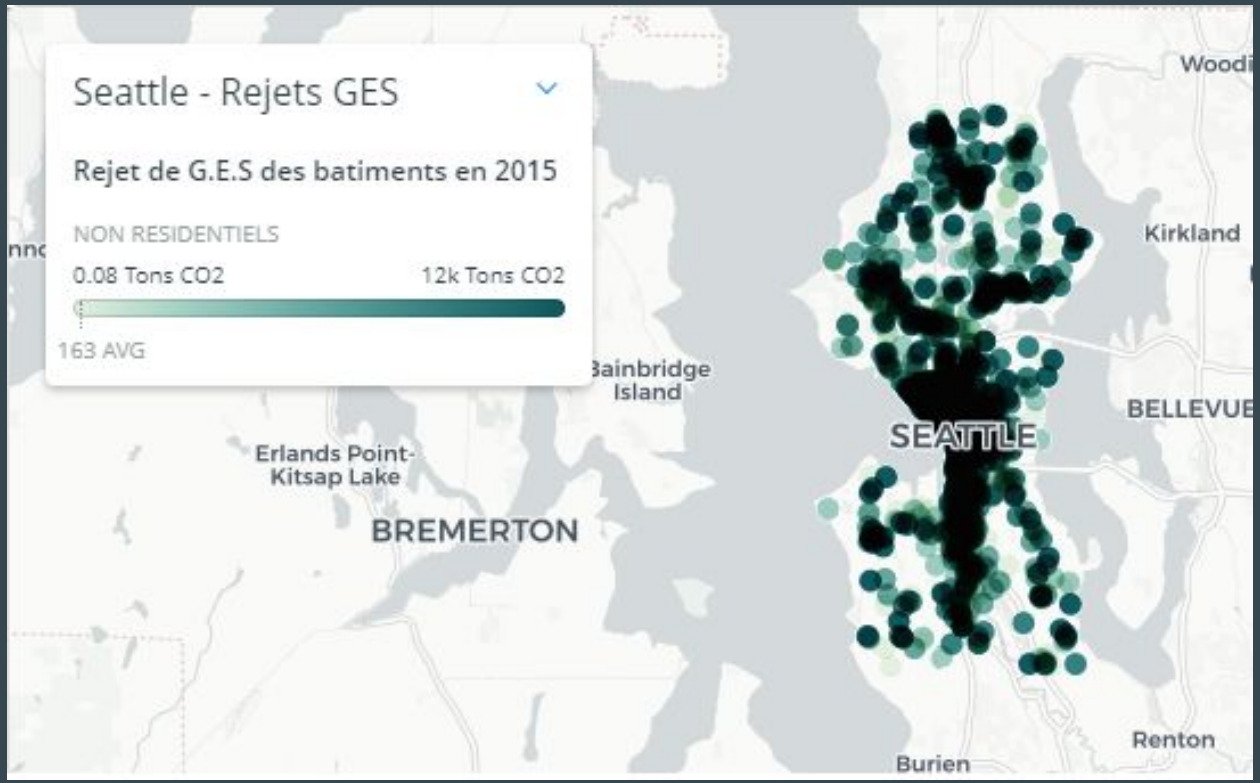
Localisation des bâtiments

- 23 % sont Downtown

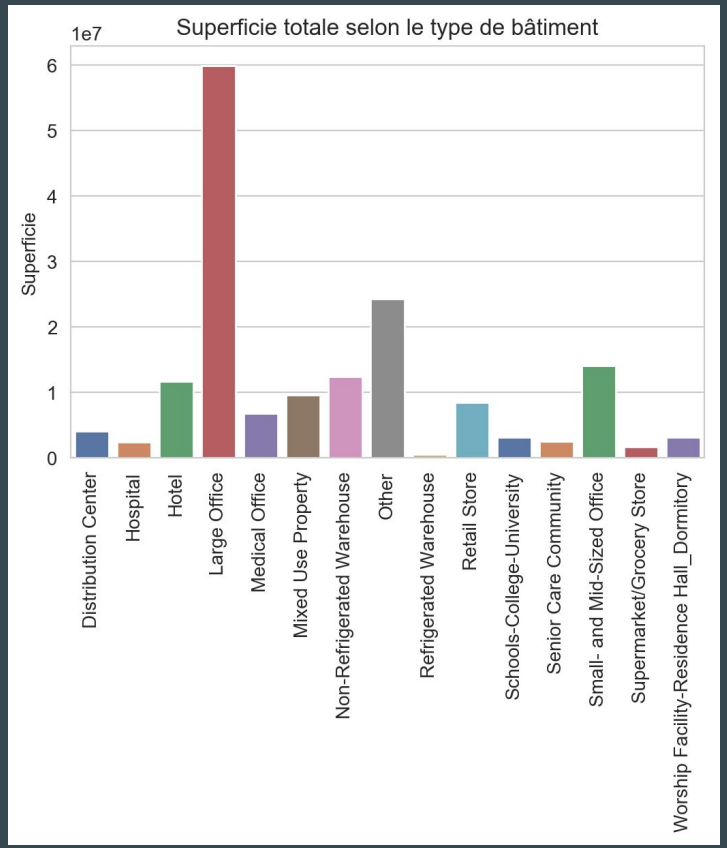


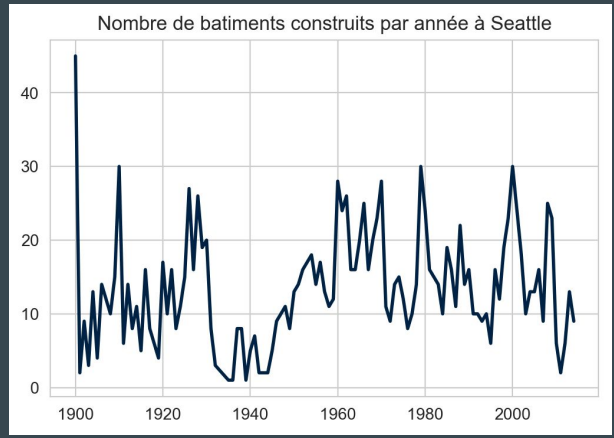


Rejets de CO2

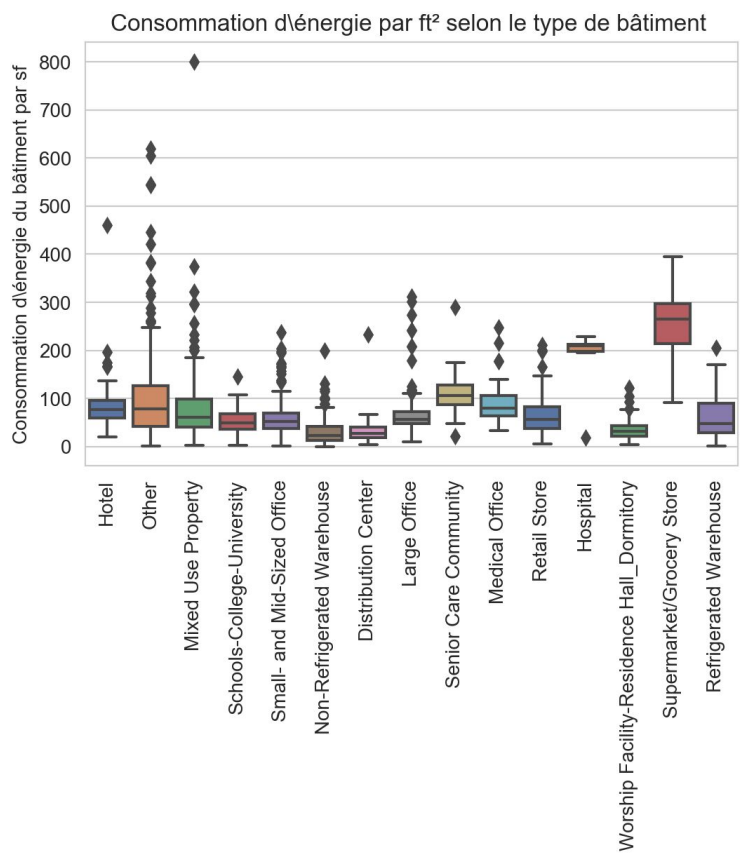


Informations sur les bâtiments



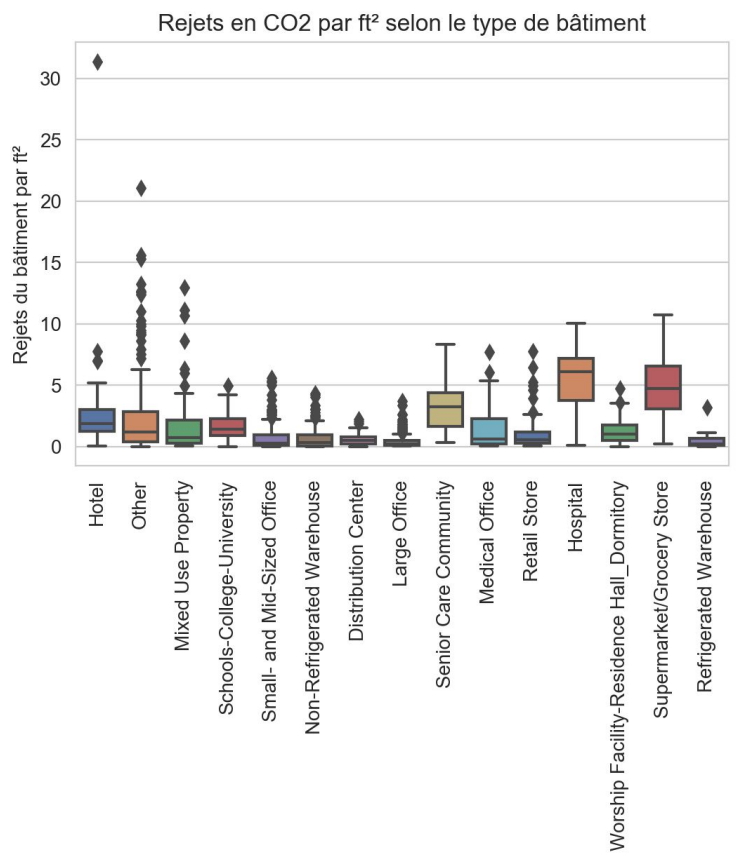


Consommations en énergie

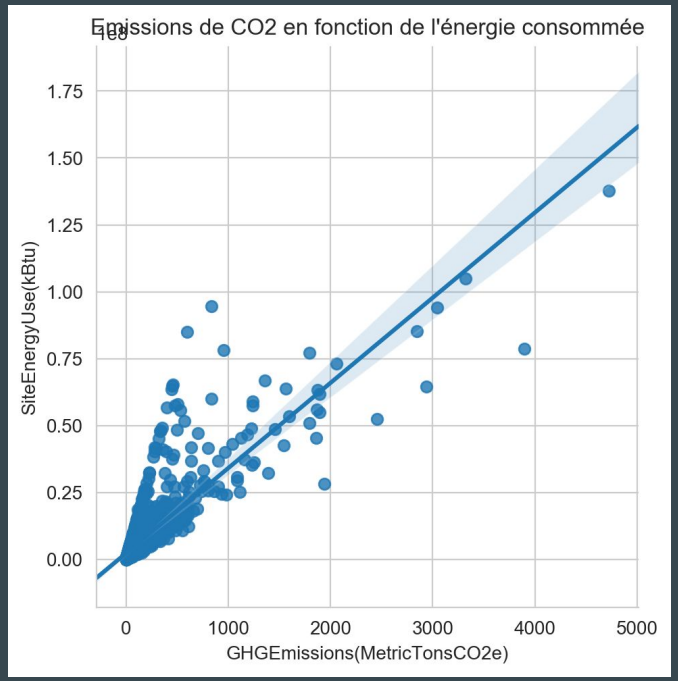


- *Hôpitaux* et *supermarchés* présentent les plus grande consommations en énergie par ft²
- Le double des autres types de bâtiments

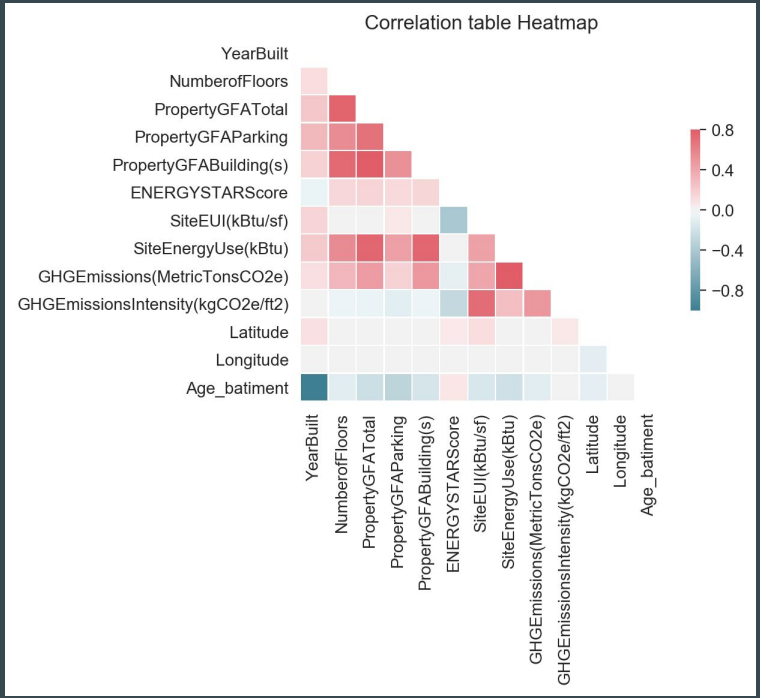
Rejets en CO2



- *Hôpitaux* et *supermarchés* sont les bâtiments qui rejettent le plus de CO2 par ft²



Corrélations



Corrélations linéaires positives fortes entre :

- *NumberofFloors* et *PropertyGFATotal* - (0.74)
- SiteEnergyUse(kBtu) et *PropertyGFATotal* - (0.66)
- *GHGEmissions(MetricTonsCO2e)* et *SiteEnergyUse(kBtu)* - (0.87)
- *GHGEmissionsIntensity(kgCO2e/ft2)* et *SiteEUI(kBtu/sf)* - (0.70)

Etonnant :

- Pas de corrélation entre Age_batiment et SiteEUI(kBtu/sf) - (-0.16)

Préprocessing

...

Choix des variables

Variables explicatives

Exprimer la taille du bâtiment
- *NumberofFloors*
- *PropertyGFAParking*
- *PropertyGFABuilding(s)*

Exprimer la location
- *Lattitude*
- *Longitude*

Exprimer l'ancienneté du bâtiment
- *Age_batiment*

Exprimer le type d'utilisation du bâtiment
- *PrimaryPropertyType*

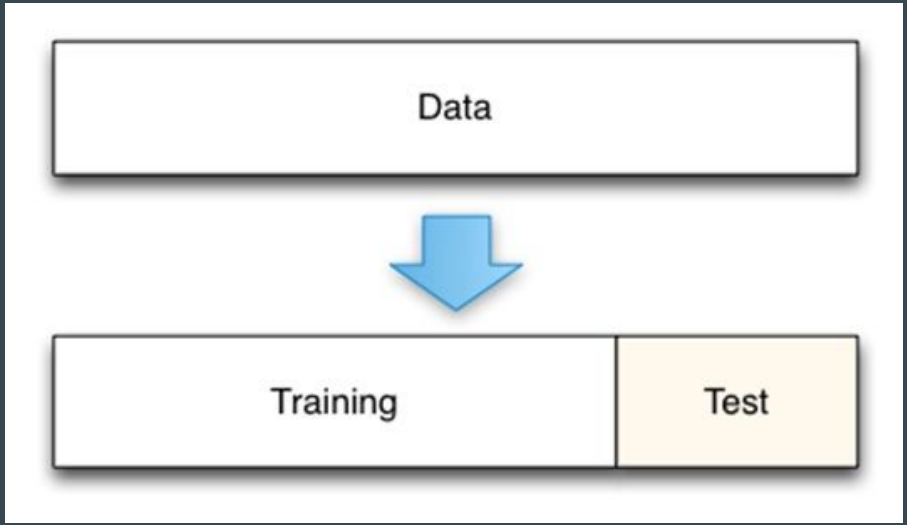
Variables dépendantes

- *SiteEnergyUse(kBtu)*
- *GHGEmissions(MetricTonsCO2e)*

Test/Train sets

Test/train set

- Train set : entrainement du modèle - 75% des données
- Test : évaluation du modèle



Transformations des variables

Transformations essayées:

- Pas de standardisation
- Passage au logarithme de la variable dépendante

Transformations retenues:

- Encodage en variables binaires de *PrimaryPropertyType* (variable qualitative)
- Standardisation de l'ensemble des données (variables explicatives et dépendantes)
    - standardisation de la variable dépendante permet d'avoir des résultats cohérents sur les régressions non linéaires.

Modélisations

...

Stratégie

Objectif : Modéliser les données afin de prédire les consommations d'énergie et les rejets de CO_2 pour un nouveau bâtiment.

Modèles testés :

- Régressions linéaires : classique, Ridge, Lasso
- Régressions non linéaires : SVR, MLPRegressor
- Méthode ensembliste : Random Forest Regressor

Recherche du meilleur paramètre:

- GridSearch

Stratégie de CrossValidation:

- Crossvalidation sur le training set
- 10 folds

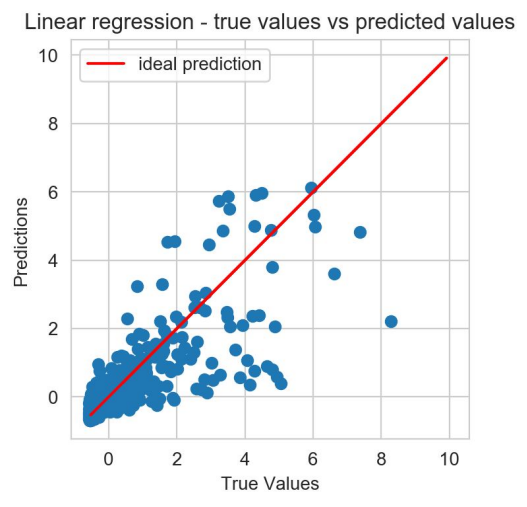
Evaluation de la performance sur le training set

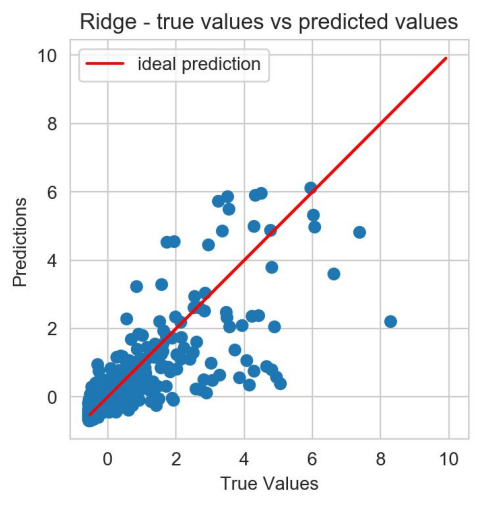
- RMSE
- R^2

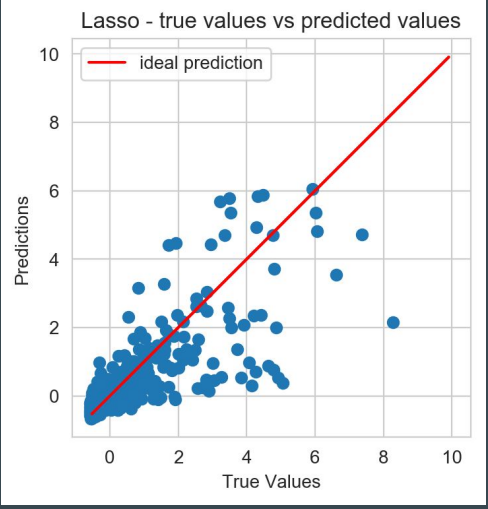
Evaluation de la généralisation sur le test set

Régressions linéaires

Prédictions des consommations d'énergie

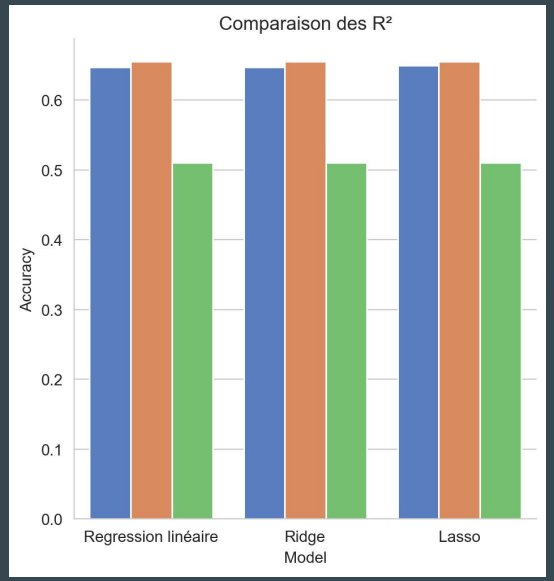


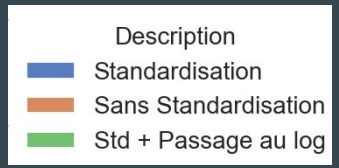


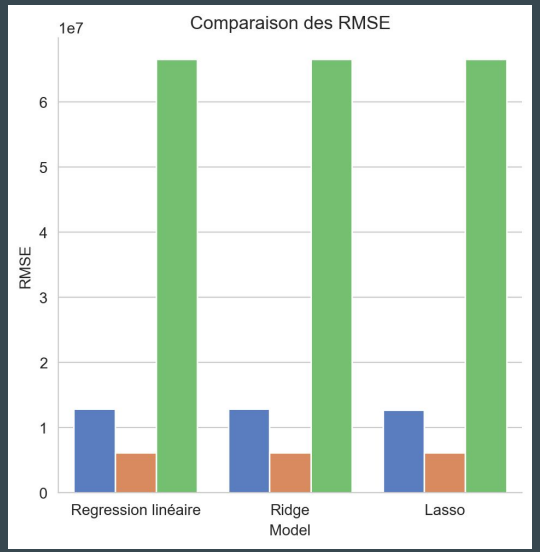


	Classique	Ridge	Lasso
RMSE	0.579	0.579	0.578
R^2	0.647	0.647	0.650

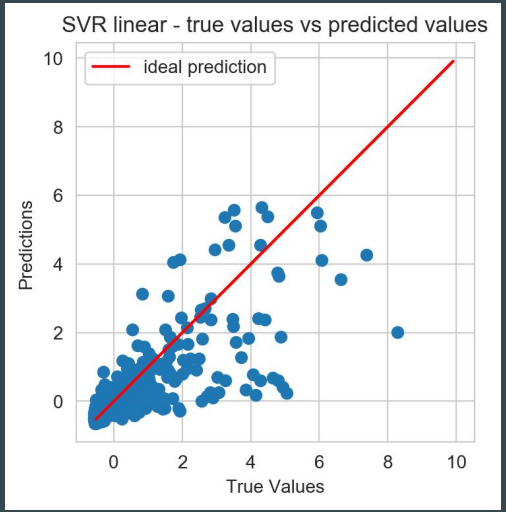
Régressions linéaires avec transformations des variables

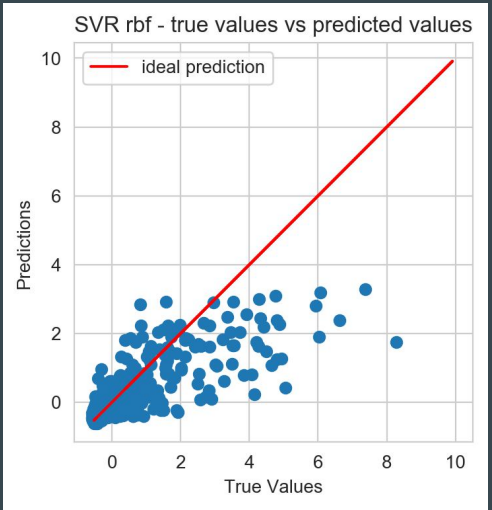


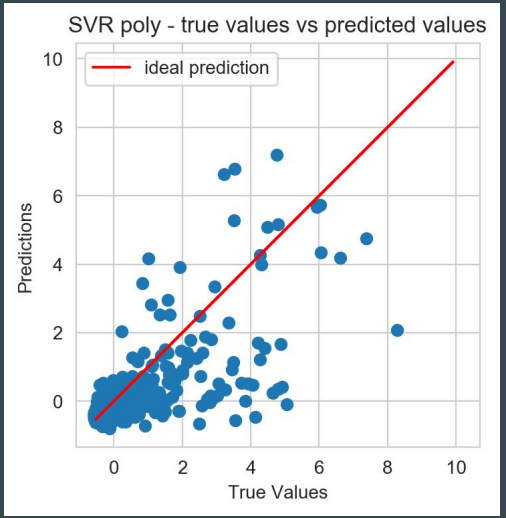




Régressions non linéaires - SVR







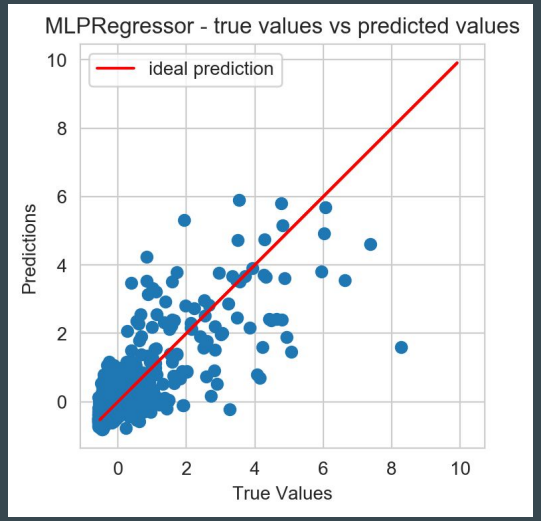
	SVR Linear	SVR rbf	SVR poly
RMSE	0.582	0.597	0.676
R^2	0.646	0.633	0.509

Régressions non linéaires - Réseau de neurones

Plusieurs fonctions d'activations sont possibles
- La fonction *tanh* est la meilleure

Performances
- RMSE : 0.576
- R^2 : 0.642

Pas d'amélioration notable par rapport aux précédents modèles



Méthode ensembliste - Random Forest Regressor

Utilisation de GridSearchCV pour trouver les meilleurs paramètres

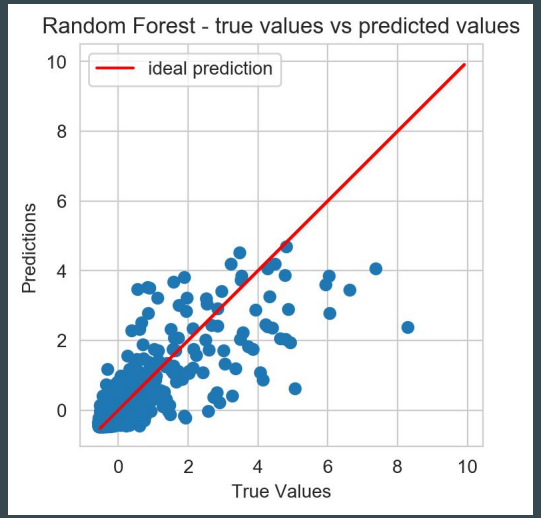
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV

# Perform Grid-Search to find best parameters
gsc = GridSearchCV(estimator=RandomForestRegressor(), param_grid={'max_depth': range(3,7), 'n_estimators': (10, 50, 100, 1000)},
```

```
grid_result = gsc.fit(X_train, y_train)
```

```
best_params = grid_result.best_params_
best_params
```

```
{'max_depth': 6, 'n_estimators': 100}
```



Performances

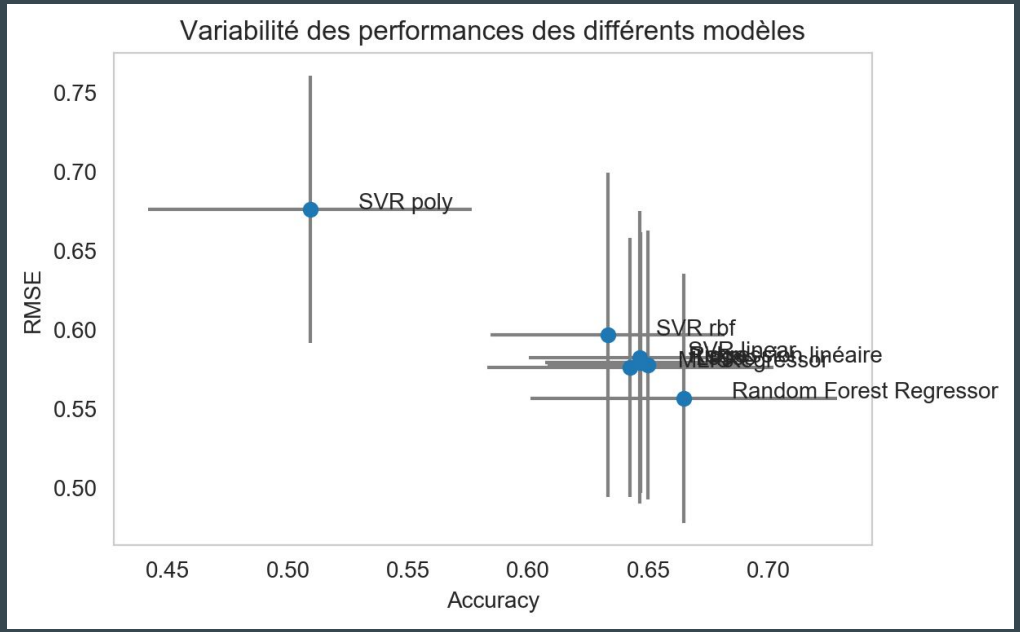
- RMSE : 0.557
- R^2 : 0.665

Meilleur modèle a priori

Choix du meilleur modèle

Grande variabilité des résultats (quand on relance le code)

- Mesure des intervalles de confiance sur les 10 résultats de la CV pour les metriques de performance



modèle minimisant les erreurs

modèle maximisant la qualité de la prédiction

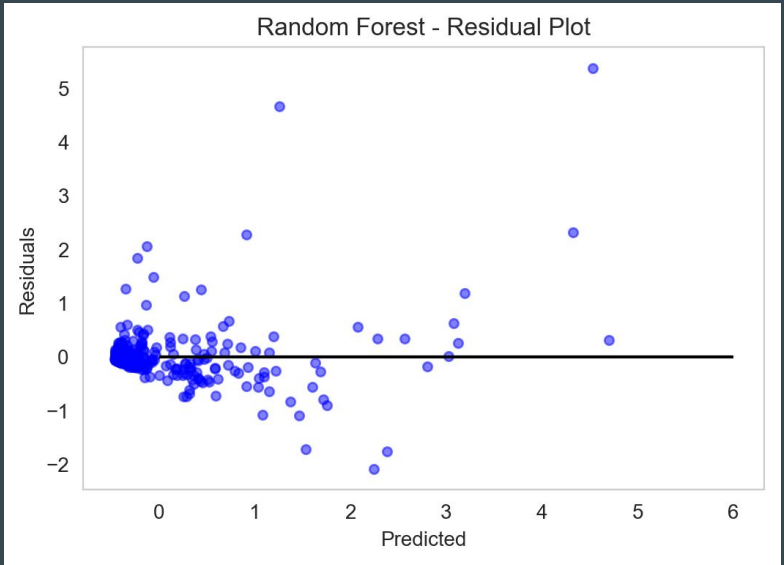
Généralisation du meilleur modèle

On applique le modèle entraîné au test set pour mesurer la performance de la généralisation du modèle

- Mesure des intervalles de confiance sur les 10 résultats de la CV pour les métriques de performance

Performances

- RMSE : 0.533
- R^2 : 0.715



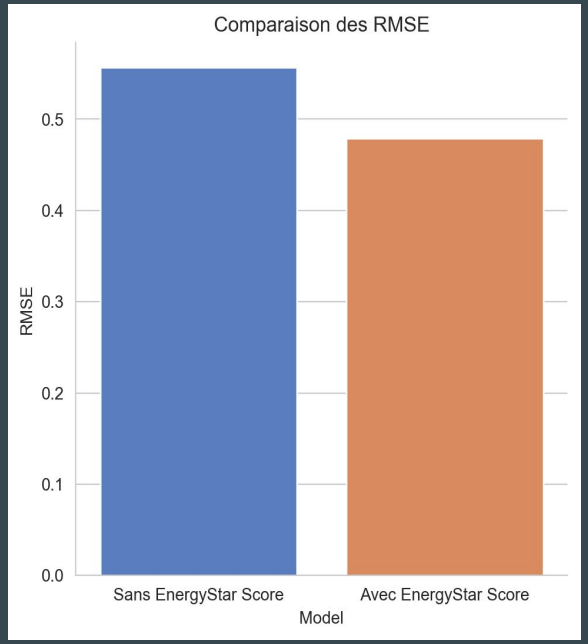
Visualisation des erreurs

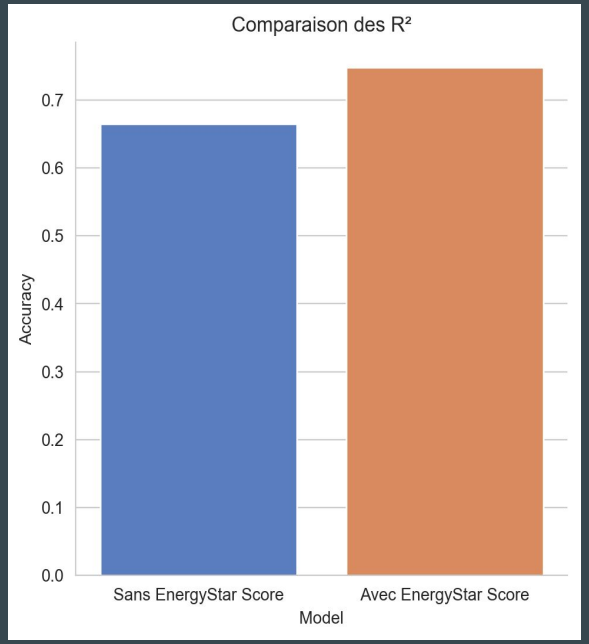
- données autour de 0 de manière aléatoire
- pas de structure

Bonne généralisation du modèle Random Forest regressor

Intérêt de la variable EnergyStar Score

On applique le meilleur modèle aux données incluant la variable *EnergyStarScore*.





Malgré son coût à calculer, connaitre la variable EnergyStarScore améliore sensiblement la qualité du modèle ainsi que ses prédictions.

Conclusion

Pour aller plus loin :

- essayer la transformation au logarithme des variables explicatives avec le RandomForest
- Appliquer le modèle choisi aux données de 2016 pour voir son comportement