



Projet 4

Anticipez les besoins en consommation électrique de bâtiments

par Rouba Yaacoub

Seattle, leader national contre le changement climatique, a annoncé le maire Jenny A. Durkan en avril 2018 en publiant une mise à jour de la stratégie climatique de Seattle.

Le plan d'action climatique de Seattle, adopté en juin 2013, se concentre sur la réduction de 80 à 100% des émissions de gaz à effet de serre, en 2050, provenant des transports et des bâtiments.

Suite aux enquêtes approfondies menées par nos agents en 2015 et 2016 sur les bâtiments de Seattle, nous allons:

1. prédire les **émissions de CO²**
2. prédire la **consommation totale d'énergie**
3. évaluer l'intérêt de l'**Energy Star Score**

C'est quoi Energy Star Score

Energy Star Score est une note de 1 à 100 calculée par l'EPA, qui évalue la performance énergétique globale d'un bien immobilier.

Un score de 50 correspond à la médiane. Plus bas que cette valeur est plus mauvais, plus haut est meilleur.

TABLE OF CONTENTS

1 Analyse exploratoire & Nettoyage des données

2 Variables catégorielles

3 Transformation normale

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives

A decorative background pattern of light blue and purple circuit lines and nodes, resembling a stylized electronic circuit board, is overlaid on the left side of the slide.

1 Analyse exploratoire & Nettoyage des données

2 Variables catégorielles

3 Transformation normale

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives

Lire les fichiers csv et les nettoyer des valeurs aberrantes:

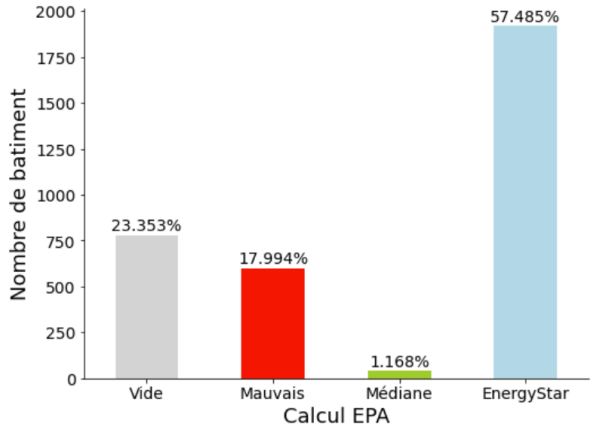
Les deux DataFrames contient des colonnes sur la consommation d'électricité, nombres des étages, "Energy Star Score", etc...

Donc notre stratégie est la suivante:

1. chacune de ces valeur doit être positive
2. les valeurs du "Energy Star Score" doit être dans l'intervalle [1, 100].

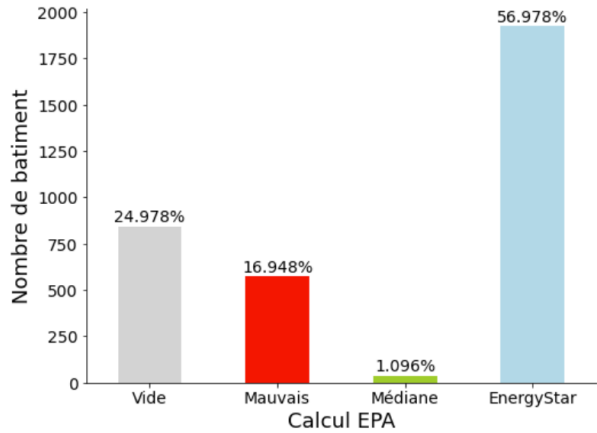
Energy Star Score en 2015 & 2016:

Figure 1: Le nombre de batiments en fonction de leur Calcul EPA en 2015



Energy Star Score en 2015 & 2016:

Figure 2: Le nombre des batiment en fonction de leur Calcul EPA en 2016



Valeurs manquantes: Energy Star Score:

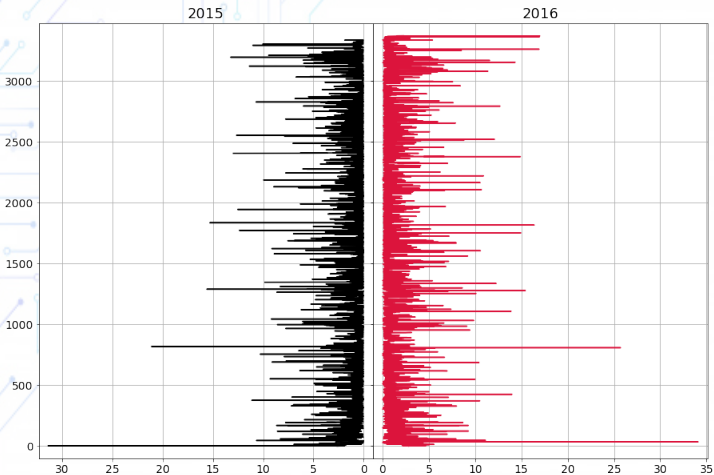
Nous avons trouvé que 25% des valeurs de l'Energy Star Score sont vides.

Nous avons donc deux solutions : soit nous supprimons toutes ces valeurs manquantes, soit nous les imputons.

Dans cette étude, et afin de ne pas biaiser les méthodes de prédiction, nous supprimerons les valeurs manquantes.

Emission totale du CO2 en 2015 & 2016:

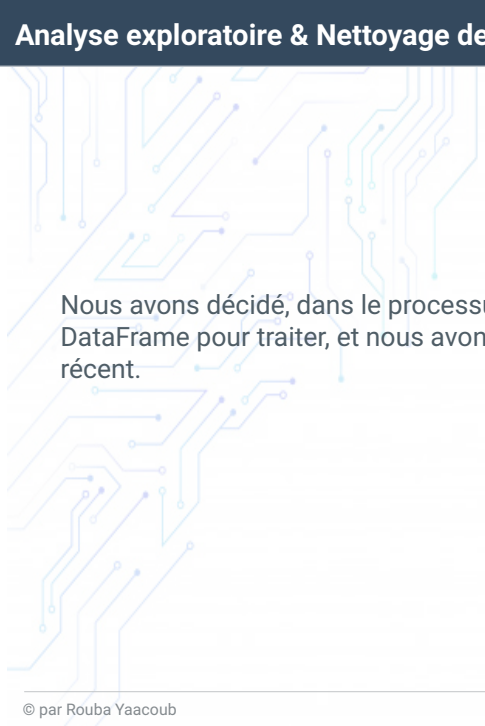
Figure 3: Emission totale de CO2 (kgCO2e/ft2) en 2015 et 2016



Valeurs manquantes: Emission totale du CO2

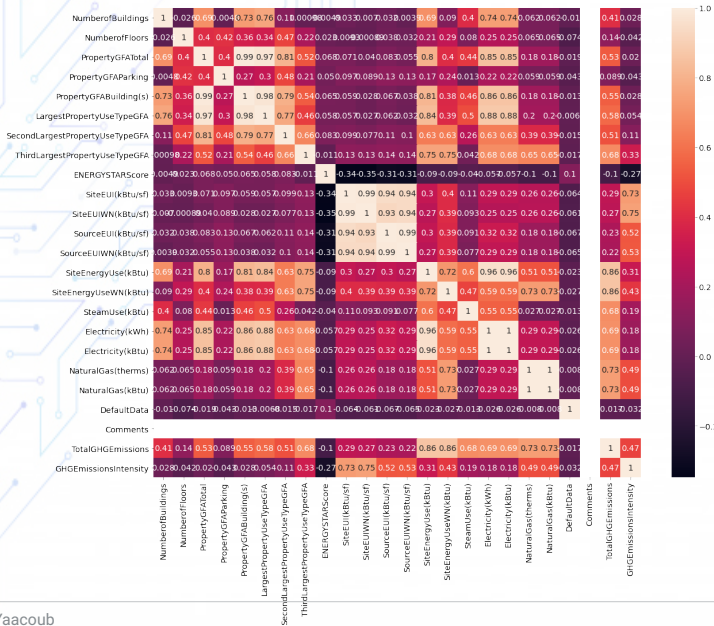
0.299% des valeurs Nan dans le dataframe 2015 et 0.269% dans celle de 2016.

Le pourcentage de données manquantes dans les colonnes de Total Greenhouse Gas Emissions n'est pas important, nous supprimerons ses lignes

A decorative graphic on the left side of the slide, consisting of a network of thin, light blue lines that resemble a circuit board or data paths. These lines are interconnected with small circles in shades of blue, green, and purple, creating a complex, web-like structure that extends from the top left towards the center of the slide.

Nous avons décidé, dans le processus suivant, de ne choisir qu'un seul DataFrame pour traiter, et nous avons choisi 2016 parce qu'il est plus récent.

Analyse exploratoire & Nettoyage des données



En nous appuyant sur la matrice de corrélation, nous filtrons le DataFrame en gardant les variables pertinentes non corrélées:

A) Les données déclaratives du bâtiment:

1. Building Type
2. Year
3. Number of Buildings, Floors
4. Property GFA total, parking

B) Les différents types de consommation par ces bâtiments:

1. Energy Star Score
2. SourceEUI
3. Site Energy Use
4. Natural gaz
5. Emission CO2

1 Analyse exploratoire & Nettoyage des données

2 **Variables catégorielles**

3 Transformation normale

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives

Variables catégorielles

```
df_with_ESS.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 2532 entries, 0 to 3371
```

```
Data columns (total 12 columns):
```

#	Column	Non-Null Count	Dtype
0	BuildingType	2532 non-null	object
1	YearBuilt	2532 non-null	int64
2	NumberOfBuildings	2532 non-null	float64
3	NumberOfFloors	2532 non-null	int64
4	PropertyGFATotal	2532 non-null	int64
5	PropertyGFAParking	2532 non-null	int64
6	PropertyGFABuilding(s)	2532 non-null	int64
7	ENERGYSTARScore	2532 non-null	float64
8	SourceEUI(kBtu/sf)	2532 non-null	float64
9	SiteEnergyUse(kBtu)	2532 non-null	float64
10	NaturalGas(kBtu)	2532 non-null	float64
11	TotalGHGEmissions	2532 non-null	float64

```
dtypes: float64(6), int64(5), object(1)
```

```
memory usage: 257.2+ KB
```

Variables catégorielles

One Hot Encoded est un processus par lequel des variables catégorielles sont converties en de nouvelles colonnes binaires afin de rendre la prédiction plus efficace.

Il y a plusieurs façon pour faire One Hot Encoded, pour cette étude nous avons choisi `get.dummies()` pour la variable catégorielle "BuildingTypes"

1 Analyse exploratoire & Nettoyage des données

2 Variables catégorielles

3 **Transformation normale**

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives

Transformation normale

Figure 4: Transformation sur le variable SiteEnergyUse

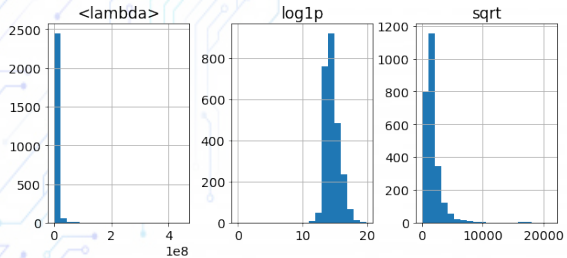
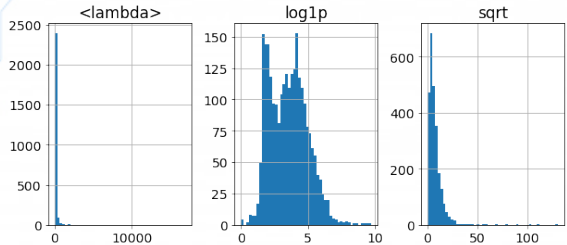


Figure 5: Transformation sur le variable émission du CO2



1 Analyse exploratoire & Nettoyage des données

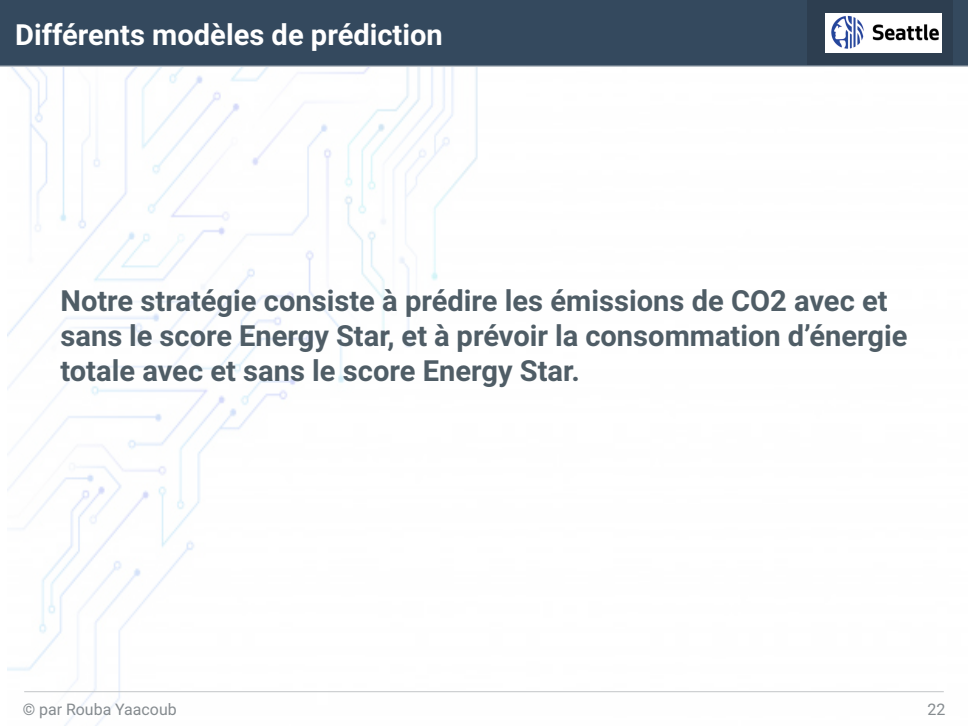
2 Variables catégorielles

3 Transformation normale

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives



Notre stratégie consiste à prédire les émissions de CO₂ avec et sans le score Energy Star, et à prévoir la consommation d'énergie totale avec et sans le score Energy Star.

Baseline + modèles de prédiction:

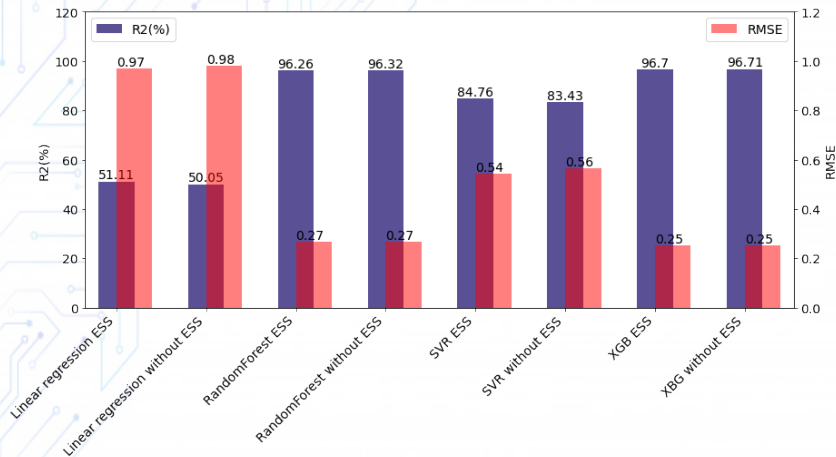
Nous utiliserons, comme baseline, dummy regressor ensuite nous utiliserons **linear regression**, **RandomForest**, **SVR** et **XGBoost**.

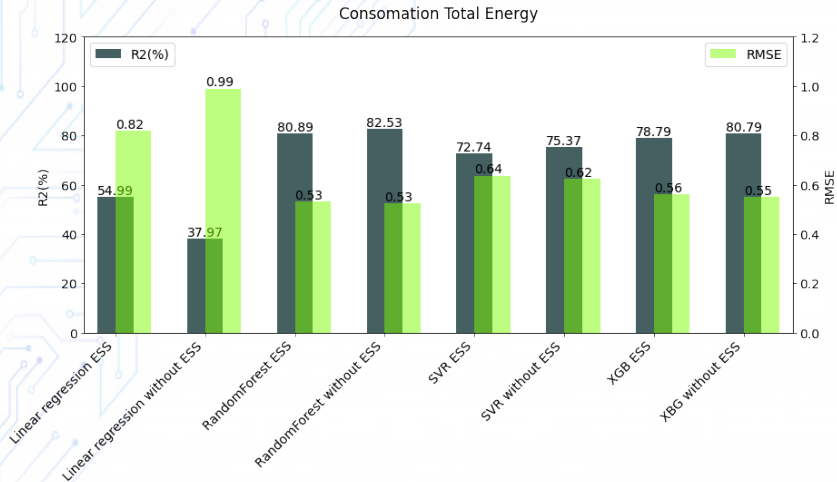
Chaque algorithme (sauf le baseline) est utilisé dans une pipeline avec `StandardScaler()`. Et chacune de ces pipelines (sauf linear regression) est utilisé avec `GridSearchCV` pour paramétrer les hyperparamètres avec cross validation.

Évaluation du performance du modèle:

Lors de chaque algorithmes, nous calculons **R²**, **RMSE** et le **temps de calcul**, puis nous l'affichons avec les meilleurs hyperparamètres choisis.

Emission du CO2





Récapitulation:

1. Les algorithmes maintiennent de bons résultats entre l'inclusion et l'exclusion de l'ESS (avec une légère différence).
2. Dans l'émission de CO2 : RandomForest et XGBoost ont des valeurs très proches et le SVR a aussi de bons résultats.
3. Dans la consommation d'énergie : RandomForest, XGBoost et SVR ont des valeurs proches.

Notre dernier choix sera donc basé sur le temps de calcul de RandomForest, SVR et XGBoost.

1 Analyse exploratoire & Nettoyage des données

2 Variables catégorielles

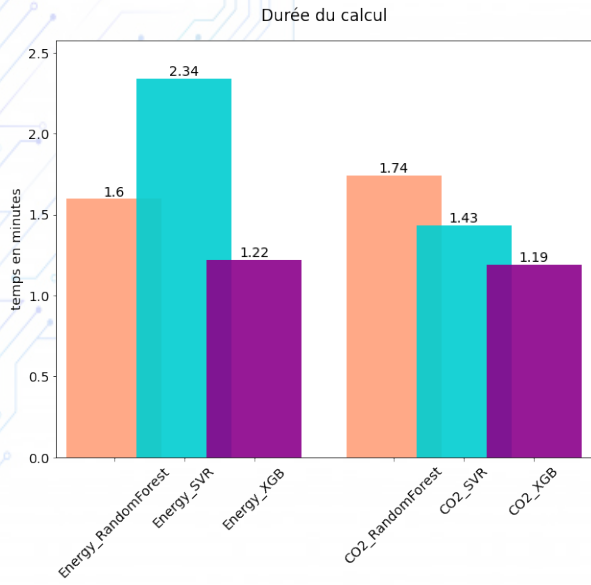
3 Transformation normale

4 Différents modèles de prédiction

5 **Choix du modèle**

6 Conclusion & perspectives

Temps du calcul



1 Analyse exploratoire & Nettoyage des données

2 Variables catégorielles

3 Transformation normale

4 Différents modèles de prédiction

5 Choix du modèle

6 Conclusion & perspectives

Conclusion & perspectives

Dans un premier temps, nous avons:

- testé plusieurs algorithmes pour prédire l'émission du CO2 et la consommation d'énergie avec et sans le score Energy Star
- proposé des algorithmes qui maintiennent de bons résultats entre l'inclusion et l'exclusion de l'ESS (avec une légère différence).
- suggérés des algorithmes qui ont de bons résultats, pas de overfit ou underfit.

Dans un deuxième temps, nous pouvons:

- ajouter plusieurs hyperparamètres.
- réduire l'échelle des hyperparamètres une fois que nous avons choisi la meilleure méthode de régression.