Data Camp - Kaggle

NAIT ABDELLA Abdelouahab
YECHCHI Sif-Eddine
EL JAMIY Mohamed

Université Paris Saclay, July 5, 2021

Plan

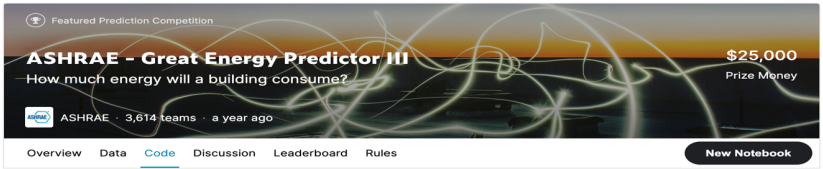
- Présentation de la compétition
- Analyse descriptive
- Pre-processing
- Modeling

ASHRAE ENERGY PREDICTION?

ASHRAE™

kaggle

Objectif de la compétition



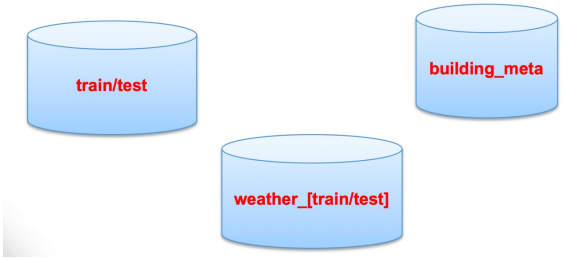
Développer des modèles précis de la consommation d'énergie des bâtiments mesurée par des compteurs dans les domaines suivants : eau glacée, électricité, eau chaude et vapeur.
Les données proviennent de plus de 1 000 bâtiments sur une période de trois ans.

https://www.kaggle.com/c/ashrae-energy-prediction/overview

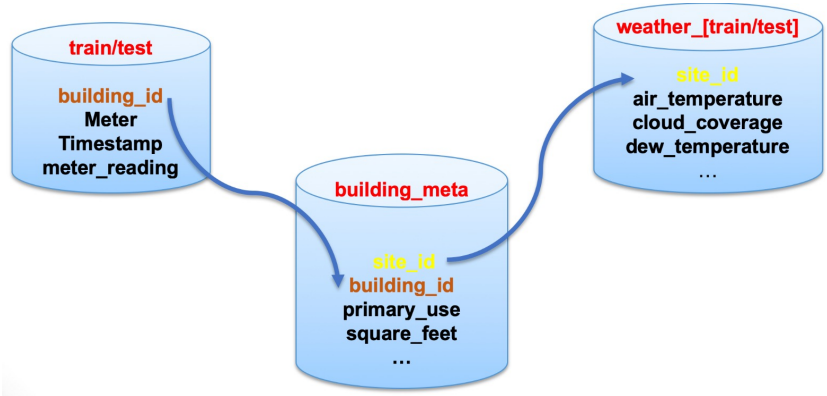
Database

L'ensemble de données comprend trois années de relevés horaires de compteurs provenant de plus d'un millier de bâtiments situés sur plusieurs sites différents dans le monde.

```
train shape is (20216100, 4)
test shape is (41697600, 4)
weather_train shape is (139773, 9)
weather_test shape is (277243, 9)
metadata shape is (1449, 6)
```



Merging Data



- le fichier *train* est de taille 20216100×16.
- le fichier *test* est de taille 41697600×16.

Métrique d'évaluation

Les compétiteurs seront classés selon la mesure d'erreur suivante :

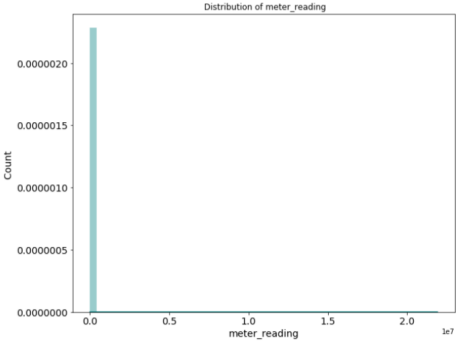
$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i + 1) - log(a_i + 1))^2}$$

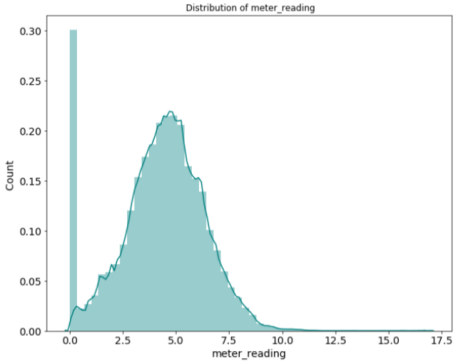
avec:

- ε : le score RMSE
- n: le nombre total d'observations dans le dataset (privé/public).
- p_i les valeurs prédites de la variable cible *meter_reading*.
- a_i les vraies valeurs de la variable cible *meter_reading*.

Analyse descriptive

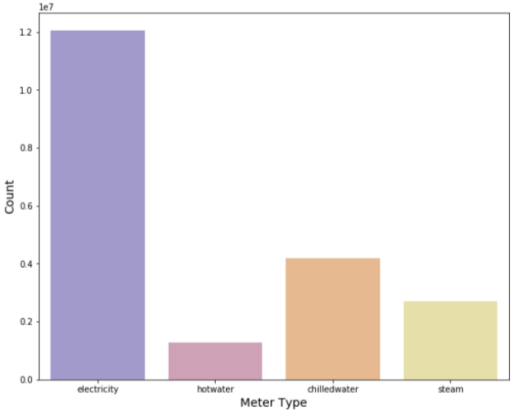
La variable cible:





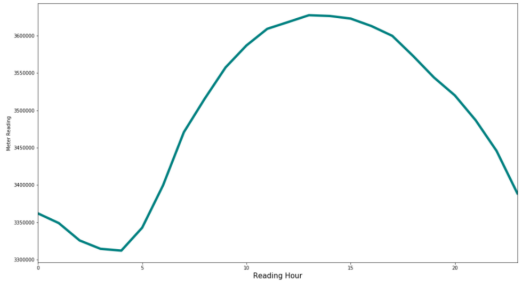
Types d'énergie:

- L'électricité est le type d'énergie le plus utilisé.
- l'eau chaude est la moins consommée.



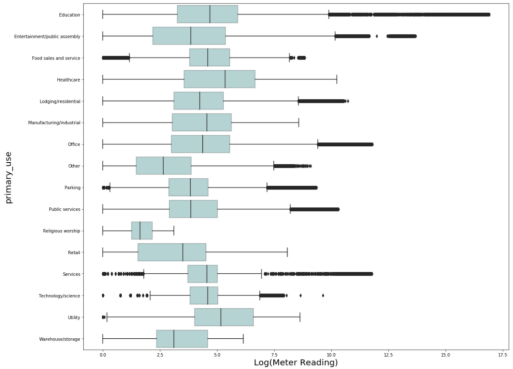
Consommation d'énergie le long d'une journée:

Importante consommation d'énergie entre 9h et 17h, ce qui correspond aux horaires de travail les plus fréquentes.



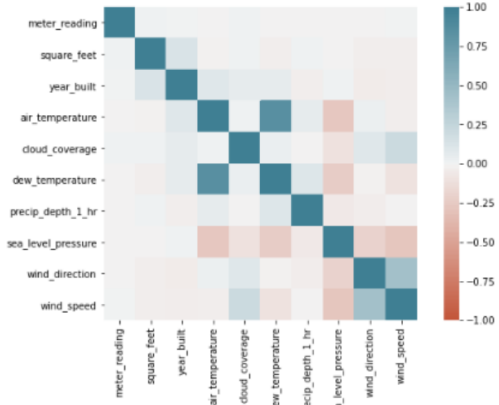
Utilisation d'énergie × consommation:

- Les bâtiments religieux gaspillent le moins d'énergie
- L'utilisation la plus fréquente se fait pour des motifs éducatifs, de santé ou encore pour des motifs utilitaires.



Corrélation entre les variables continues:

Corrélations linéaires entre les variables météorologiques.



Corrélation de Spearman:

	meter_reading	square_feet
meter_reading	1.000000	0.499369
square_feet	0.499369	1.000000

Figure: Corrélation entre la consommation d'énergie et le surface des bâtiments

square_feet $\times$ consommation:

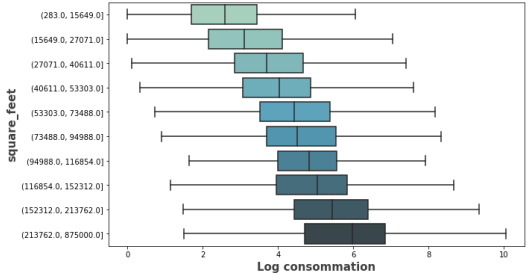


Figure: Le logarithme de la consommation $\times$ la sufrace des bâtiments

Valeurs manquantes

	COLUMN NAME	MISSING VALUES	TOTAL ROWS	% MISSING
7	year_built	12127645	20216100	59.99
8	floor_count	16709167	20216100	82.65
9	air_temperature	96658	20216100	0.48
10	cloud_coverage	8825365	20216100	43.66
11	dew_temperature	100140	20216100	0.50
12	precip_depth_1_hr	3749023	20216100	18.54
13	sea_level_pressure	1231669	20216100	6.09
14	wind_direction	1449048	20216100	7.17
15	wind_speed	143676	20216100	0.71

Figure: Les valeures manquantes

Détection des outliers

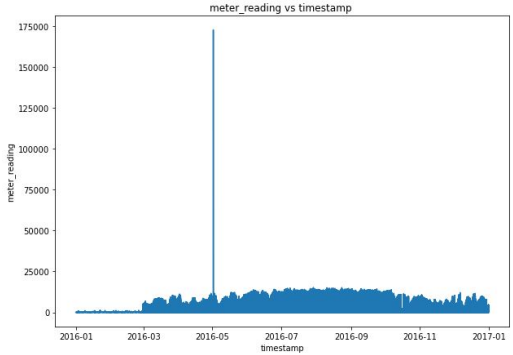


Figure: La consommation dans le site 0 en fonction de timestamp

Détection des outliers

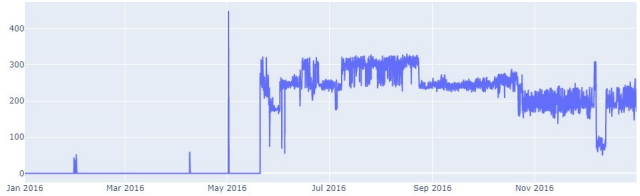


Figure: La consommation dans le bâtiment 0

Figure: La consommation dans le bâtiment 1288

Détection des outliers

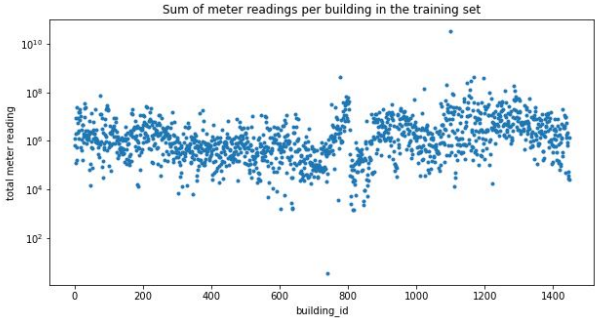


Figure: La consommation dans le site 0 en fonction de timestamp

Détection des outliers

Figure: La consommation dans le bâtiment 1099

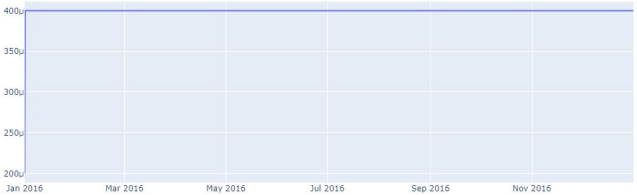


Figure: La consommation dans le bâtiment 740

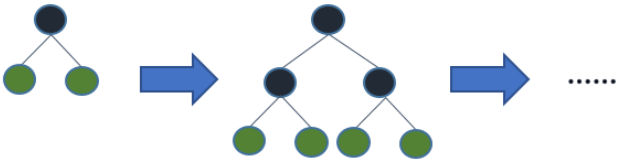
Feature Engineering

- Ajout des variables temporelles telles que l'heure, le jour, le mois et le week-end en transformant la variable timestamp
- Ajouter la variable is_holiday.
- Encodage de la variable primary_use
- La conversion log1p sur la variable meter_reading

Light Gradient Boosting

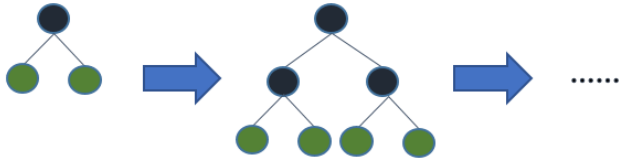
- Light GBM est un cadre de gradient boosting basé les arbres de décision.
- La construction des arbres se fait par feuille plutôt que par niveau de profondeur.
- Quelques avantages:
  - Gérer des grands volumes de données
  - Gérer des grands volumes de données et plus efficace
  - Réduire l'utilisation de mémoire
  - Supporte le calcul parallèle GPU

Principe

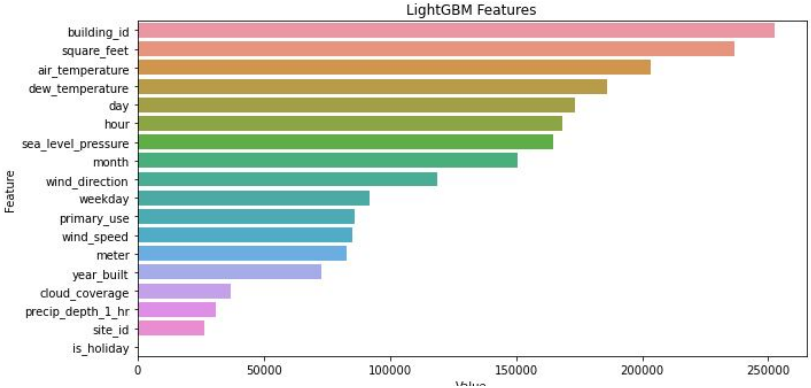


Level-wise tree growth

Principe



Importance des variables pour ce modèle

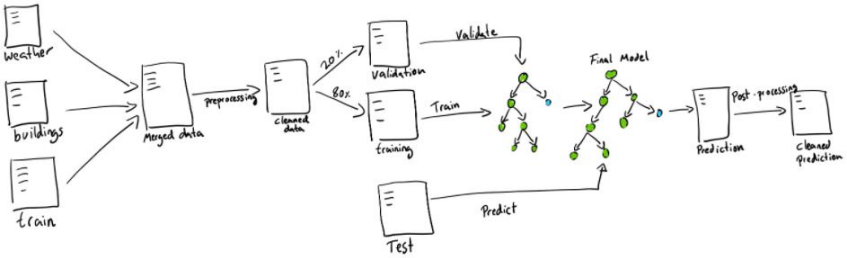


Soumissions et meilleur score

Name	Submitted	Wait time	Execution time	Score
submission_L6.csv	3 minutes ago	1 seconds	175 seconds	1.190

Complete

Resumé


weather
buildings
train
Merged data
preprocessing
cleaned data
20%
80%
Validation
training
Validate
Train
Final Model
Test
Predict
Prediction
Post-processing
cleaned prediction

Thank You !