Université d' Evry Val d'Essonne & Paris Scalay
Département de Mathématiques

M2 Data Science : Santé Assurance Finance

Data Camp - Kaggle
ASHRAE ENERGY PREDICTION

Auteurs :
NAIT ABDELLA Abdelouahab
YECHCHI Sif-Eddine
EL JAMIY Mohamed

Encadré par :
Mr Simon BUSSY
Mme Agathe GUILLOUX

Année universitaire : 2020-2021

Table des matières

1 Présentation de la compétition

1.1 Contexte et objectif

Un grand pourcentage de la consommation énergétique du secteur du bâtiment (près de 70% en France) émane des besoins en chauffage et climatisation. Ces usages très énergivores sont la source de multiples problématiques environnementales et économiques, ainsi que de notre dépendance aux énergies fossiles. L'un des enjeux majeurs de la transformation des besoins en chauffage et climatisation consiste à réduire leurs émissions de gaz à effet de serre au niveau mondial en développant de nouvelles filières. Rénovations à grande échelle, déploiement des équipements efficaces et bas carbone et flexibilité seront les maîtres mots pour parvenir à une transition écologique, responsable et économiquement viable.

Nous nous intéresserons dans ce rapport à une compétition **Kaggle** qui a été lancée par **ASHRAE** (*American Society of Heating and Air-Conditioning Engineers*), une société américaine innovante en matière des solutions de chauffage et de climatisation des bâtiments.

L'objectif de ce travail est d'évaluer l'efficacité des investissements considérables qui ont été alloués pour la climatisation de grands bâtiments, autrement dit il s'agit de voir si les nouvelles méthodes de chauffage implémentées (qui réduisent les émissions) avec ces investissements nous permettent de réduire significativement la consommation énergétique. Pour se faire, nous allons construire un modèle contre-factuel permettant de prédire le taux de consommation d'énergie avant la rénovation des bâtiments pour pouvoir le comparer avec la consommation observée afin de savoir si l'amélioration apportée a été efficace.

1.2 Base de données

Les données que nous manipulerons durant ce projet proviennent de plus de 1000 bâtiments sur plusieurs sites dans le monde, et qui ont été observées pour une période de trois ans (un enregistrement par heure). Les données fournies nous informent sur les bâtiments concernés par cette étude et les conditions météorologiques au moment de l'observation. Ci-dessous un descriptif détaillé de la base de données mise à notre disposition :

❶ Un fichier **train/test** :

➔ building_id - Identifiant du bâtiment ;

➔ meter - Méthode de climatisation. 0 : électricité, 1 : eau froide, 2 : vapeur, 3 : eau chaude. Pas tous les bâtiments disposent des quatre méthodes !

➔ timestamp - L'instant de la prise des mesures ;

➔ meter_reading (base train seulement) - La variable cible représentant la consommation d'énergie en KWh (elle est en kBTU pour meter = 0 !).

❷ Un fichier **building_meta** :

➔ site_id - Identifiant indiquant le site où se localise le bâtiment ;

➔ building_id - Identifiant du bâtiment ;

➔ primary_use - Indicateur de l'activité principale du bâtiment ;

➔ square_feet - Surface de plancher du bâtiment ;

➔ year_built - Année d'ouverture du bâtiment ;

➔ floor_count - Nombre d'étages dans le bâtiment.

❸ Un fichier **weather_[train/test]** :

➔ air_temperature - Température de l'air en degré Celsius ;

➔ cloud_coverage - Partie du ciel couverte de nuages ;

➔ dew_temperature - Température de rosée ;

➔ precip_depth_1_hr - Précipitations en millimètres,

➔ sea_level_pressure - Pression en Millibars ;

➔ wind_direction - Direction du vent (0-360) ;

➔ wind_speed - Vitesse du vent en m/s.

1.3 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*.

C'est donc la racine carrée de l'erreur logarithmique quadratique moyenne. Elle sera calculée à partir des prédictions soumises par le candidat en utilisant les données de la base test avec les vraies valeurs, qui restent toujours à la disposition de l'organisateur de cette compétition.

2 Exploration des données

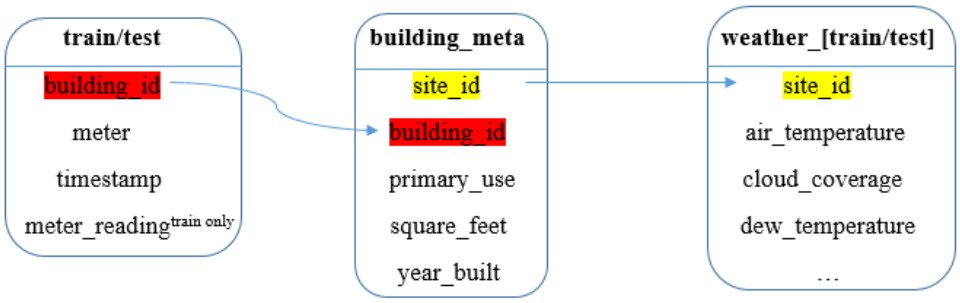
L'exploration des données est une étape cruciale pour bien réussir la phase de modélisation et obtenir un bon modèle prédictif minimisant l'erreur définie précédemment. Elle nous permettra également d'interagir avec les données et de bien comprendre les différents enjeux qu'elles présentent.

2.1 Préparation de la base de données

En parcourant la description des variables, nous avons croisé une remarque notifiant que la consommation énergétique correspondant à l'utilisation de l'électricité est en kBtu (*kilo-British thermal unit*), alors qu'elle est exprimée en KWh pour les 3 autres méthodes de climatisation. Nous avons donc tout mis en même unité (KWh).

2.1.1 Jointure des fichiers

Nous avons réalisé une jointure entre les trois sources de données, ce qui nous a permis d'obtenir pour chaque observation de la consommation énergétique des informations sur le bâtiment concerné et la météo à l'instant de prise de cette observation. Le diagramme ci-dessous montre comment nous avons procédé :



Nous nous retrouvons alors avec un seul fichier train/test rassemblant toutes les variables décrites auparavant.

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

2.1.2 Conversion de la variable temporelle

La variable *timestamp* nous renseigne les date et heure exactes de la capture des observations, sa valeur est de la forme suivante : "2016-01-01 00 :02 :10" (1^{er} janvier 2016 à 2h10).

Nous avons donc procédé à sa conversion, dans les 2 fichiers : train et test, afin d'en extraire des attributs tels que : hour, month, weekday et day, afin de les manipuler et de s'en servir pour la visualisation des données.

2.2 Analyse descriptive de données

Avant toute chose, nous nous intéressons toujours à la distibution de la variable cible en premier. La figure 1 représente la distibution de la variable *meter_reading*. Il semble que la variable soit fortement biaisée de manière positive avec les résultats. Nous avons fixé cela en passant au log(1+.). La figure 2 représente la distribution du log(1+meter_reading).

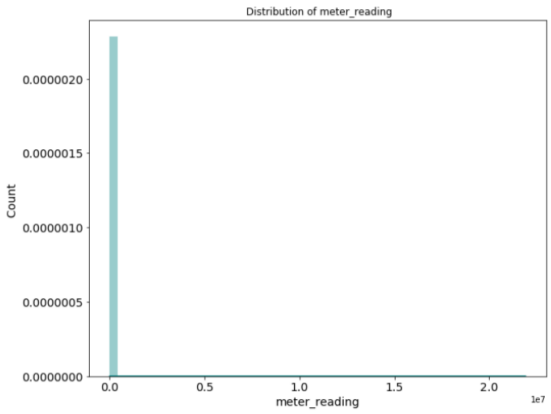


FIGURE 1 – Distribution de la variable cible

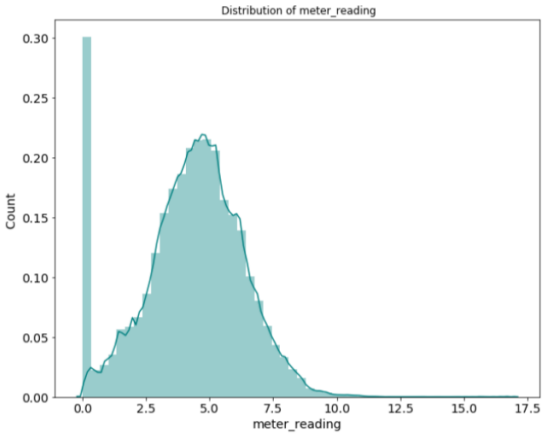
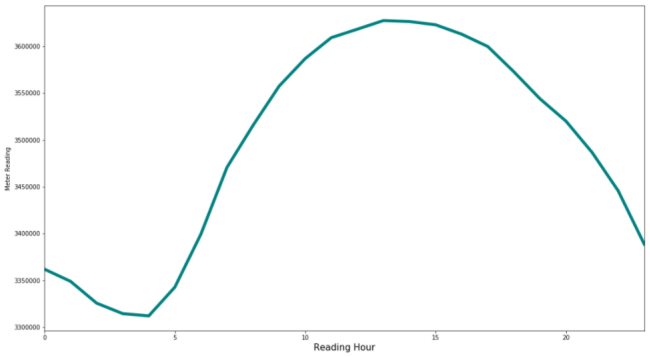


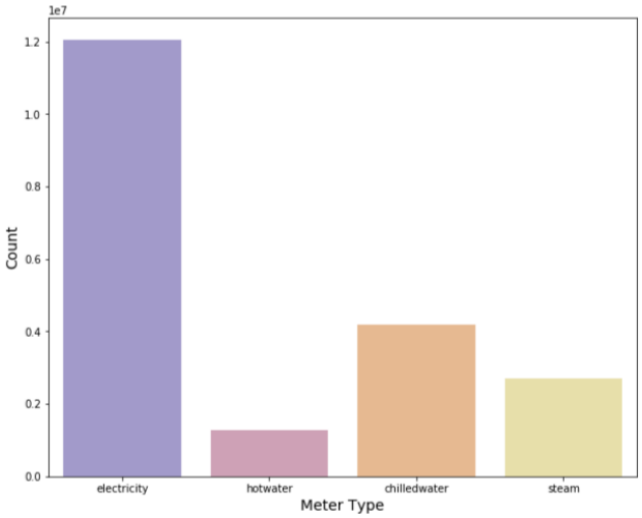
FIGURE 2 – Distibution de la variable cible après la transformation

Pour comprendre la distribution de la variable cible au cours du temps, nous avons tout d'abord tracé la distribution de **meter_reading** en fonction de la variable **hour** :

Nous pouvons observer que la consommation d'énergie est importante entre 9h et 17h, ce qui correspond aux horaires de travail les plus fréquentes.



Nous pouvons constater que l'électricité est le type d'énergie le plus utilisé tandis que l'eau chaude est la moins consommée.



Nous nous concentrerons dans la suite de cette partie sur notre variable cible *meter_reading* en la mettant en relation avec les variables explicatives potentielles :

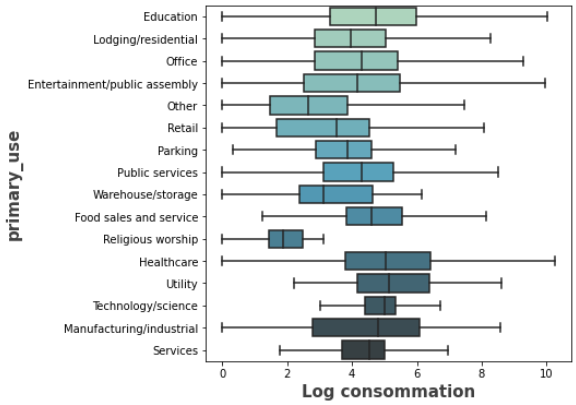


FIGURE 3 – Consommation × *primary_use*

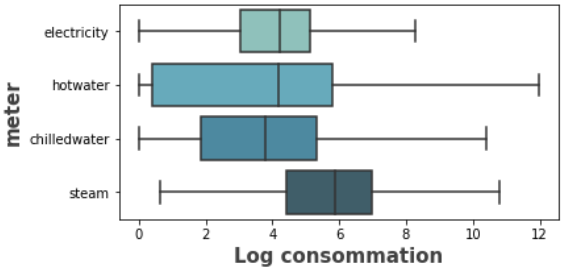


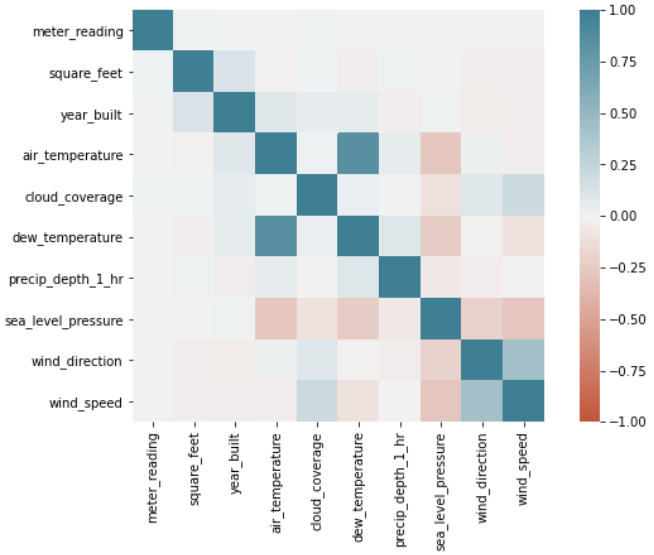
FIGURE 4 – Consommation × *meter_type*

La figure 3 représente la consommation d'énergie en fonction de l'utilisation des bâtiments. Nous pouvons observer que la consommation dépend des activités faites dans les bâtiments. Nous remarquons aussi que les bâtiments religieux gaspillent le moins d'énergie tandis que l'utilisation la plus fréquente se fait pour des motifs éducatifs, de santé ou encore pour des motifs utilitaires.

A partir de la figure 4 représentant des boîtes à moustaches, nous pouvons observer que le type d'énergie *steam* (la vapeur) consomme le plus d'énergie par rapport aux autres types.

Matrice de corrélation des variables continues :

La figure à droite présente la matrice de corrélation entre les variables continues de notre base de données. Nous remarquons qu'il existe des corrélations linéaires entre les variables météorologiques.



Corrélation de Spearman :

La variable *square_feet* mesure la somme des surfaces dans chaque étage et indique la taille des bâtiments. Il semble alors intéressant de voir comment elle interagit avec la consommation. La matrice des corrélations présentée ci-avant a montré qu'il n'y a pas de lien linéaire entre les deux variables, nous avons alors tenté la corrélation de Spearman :

	meter_reading	square_feet
meter_reading	1.000000	0.499369
square_feet	0.499369	1.000000

FIGURE 5 – Corrélation entre la consommation d'énergie et le surface des bâtiments

Nous observons une corrélation positive significative, ceci approuve que la consommation augmente avec la taille (la surface) des bâtiments, chose que nous pouvons constate également sur le graphe suivant :

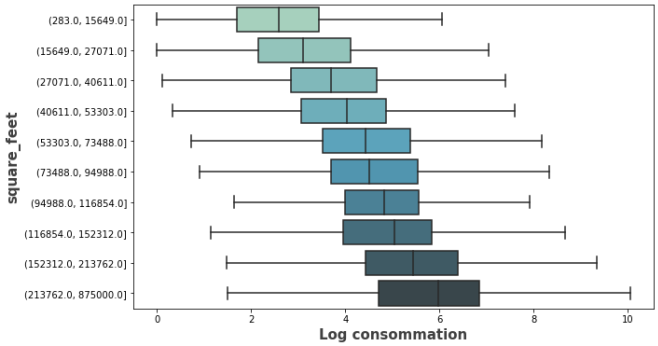
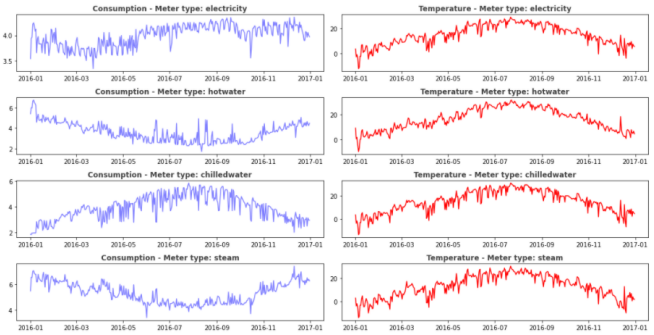


FIGURE 6 – Le logarithme de la consommation × la sufrace des bâtiments

La consommation d'énergie due à l'utilisation de l'eau chaude augmente lorsque la température baisse et vice-versa. Nous retrouvons la même tendance pour la méthode vapeur.
Lorsque l'eau froide est utilisée pour refroidir les bâtiments lors des fortes chaleurs, le niveau de consommation d'énergie augmente.



3 Pre-Processing

3.1 Valeurs manquantes

Python, comme tout outil de modélisation, ignore les lignes qui contiennent au moins une valeur manquante trouvée dans la base de données qu'on lui fournit, ce qui peut nous induire à un modèle biaisé. Nous allons localiser dans cette partie l'ensemble de valeurs éventuellement manquantes tout en proposant des méthodes pour les imputer. La figure suivante montre le nombre et le pourcentage des valeurs manquantes pour chaque variable :

	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 7 – Le nombre des valeurs manquantes et leur pourcentage

Variables liées aux bâtiments :

• Les valeurs manquantes dans la variable *year_built* ont été imputées par l'année la plus fréquente (le mode correspondant) en 1976.

• Plus de 82% des valeurs de la variable indiquant le nombre d'étages dans les bâtiment sont manquantes, soit 75% des bâtiments ciblés par cette étude. Nous avons alors pris la décision de supprimer cette variable.

Variables météorologiques :

Il existe deux sources de valeurs manquantes pour les variables météorologiques : des valeurs manquantes dans le fichier *weather*, et des dates d'enregistrement qui sont présentes dans la base *train* et absentes dans la base *weather*.

Les bâtiments sont présents à différents endroits géographiques, ce qui entraîne des variations dans les conditions météorologiques.

Compte tenu de la structure de ces données, au lieu d'imputer toutes les valeurs manquantes par une seule valeur, nous avons choisi de les remplir par la moyenne tout en regroupant par l'identifiant du site, le mois et le jour, c'est à dire que nous avons imputé chaque valeur manquante par la valeur moyenne dans le même jour et dans le site où se localise le bâtiment. Cette méthode semble plus précise, tout en gardant une variabilité dans les données imputées.

3.2 Valeurs aberrantes :

Les valeurs aberrantes ont toujours posé des problèmes de distinction entre les points extrêmes (valeurs observées) et les valeurs non fiables (valeurs fausses dues à des fautes de saisie par exemple). Ce paragraphe est consacré à la détection et le traitement de ces valeurs.

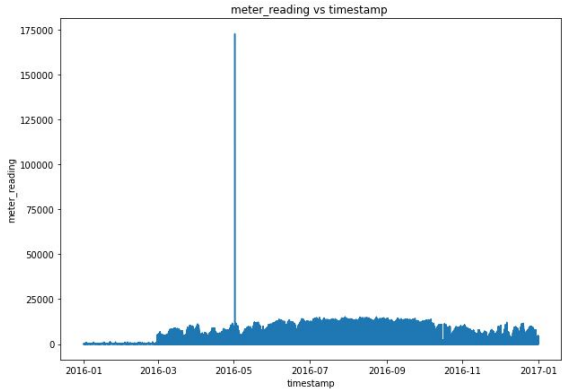


FIGURE 8 – La consommation dans le site 0 en fonction de timestamp

FIGURE 9 – La consommation d'électricité dans le bâtiment 0 de site 0

La figure à gauche représente l'évolution dans le temps de la consommation d'énergie, associée aux bâtiments du site 0 :

- Nous remarquons que la consommation s'annule pour tous les bâtiments de ce site dans les trois premiers mois.
- Nous observons une valeur anormale de consommation vers mai 2016.

Nous avons zoomé dans la deuxième figure sur un bâtiment(*building_id*=0 qui appartient au même site. Nous retrouvons la même remarque sur la tendance que notre variable cible (la variable cible est nulle jusqu'au 20 mai 2016), nous soupçonnons un problème dans les enregistrements de consommation d'énergie dans ce site à cette période. Nous supprimons donc ces observations de notre base de données.

Ce n'est pas le cas uniquement pour ce building mais également pour plusieurs sur le site 0 jusqu'au bâtiment 104. Les relevés d'électricité étaient nuls pendant 5 mois à partir de janvier, ce qui indique que la consommation a commencé ou que les relevés d'électricité ont été effectués à partir du 20 mai. Il se peut que le site n'ait pas été utilisé avant cette date ou qu'il avait des problèmes au niveau des compteurs ou pour d'autres raisons que nous ignorons. Nous avons donc enlevé les zéros lors l'apprentissage de notre modèle.

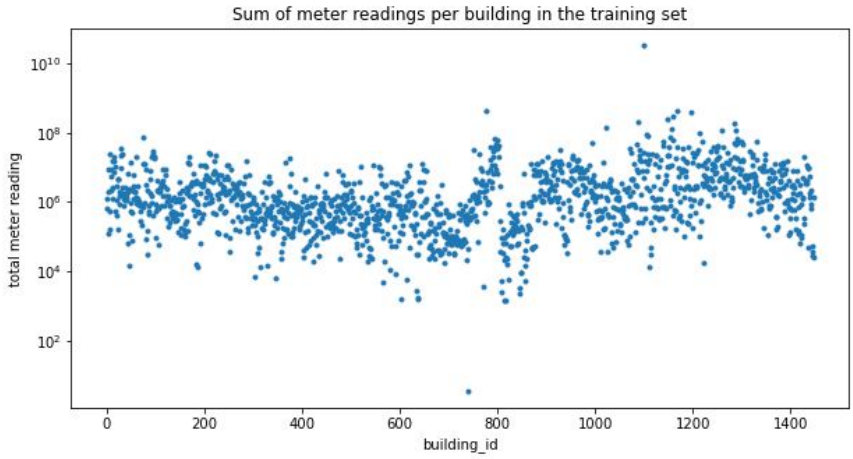


FIGURE 10 – Le totale de la consommation dans chaque bâtiment

Nous avons présenté dans cette figure la consommation totale de chaque bâtiment durant la période d'observation. Deux bâtiments s'éloignent beaucoup du nuage de points.

FIGURE 11 – la consommation dans le bâtiment 1099

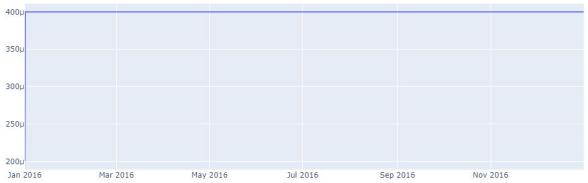


FIGURE 12 – la consommation dans le bâtiment 740

La figure à gauche montre la variation de la consommation du bâtiment 1099. Nous voyons clairement qu'elle atteint des valeurs très élevées (atteint 20 millions KWh) par rapport à la consommation des autres bâtiments qui s'élève à 80000 KWh.

La figure à droite représente la consommation dans le bâtiment 704 où nous observons, tout au long de l'étude, que la consommation d'énergie est presque nulle (environ 0.0004 KWh). Nous avons donc décidé de supprimer toutes les lignes correspondantes à ce bâtiment de notre jeu de données.

Notons qu'il ne s'agit pas des seuls exemples où l'on observe des valeurs aberrantes (dans d'autres bâtiments) et qui doivent être filtrées également.

FIGURE 13 – la consommation dans le bâtiment 1288

Dans l'exemple de ce bâtiment, la consommation chute brusquement vers 0, puis elle revient à sa tendance normale, ces observations ont été retirées.

3.3 Feature Engineering

3.3.1 Ajout des variables :

- Ajout des variables temporelles telles que l'heure, le jour, le mois et le week-end en transformant la variable **timestamp**
- Les besoins énergétiques peuvent être différents pendant les jours de vacances/ jours fériés par rapport aux autres jours, l'inclusion de cette caractéristique peut être une bonne piste afin d'améliorer notre modèle. Pour ce, on ajoute la variable nommée **is_holiday**.

3.3.2 Encodage des variables

Nous avons encodé la variable *primary_use* représentant l'usage des bâtiments afin d'avoir 16 modalités codées 0 ou 1.

4 Modeling :

Une fois notre base de données prête, nous avons attaqué la partie modélisation. Nous avons commencé dans un premier temps par des modèles simples d'arbres de décision, nous avons lancé par la suite des modèles plus complexes comme les forêts aléatoires et XGboost, qui prennent beaucoup de temps lors de la compilation ($\simeq 10h$) vu le volume de données manipulées, les scores obtenus n'étant pas satisfaisants.

Nous présentons ci-dessous le modèle LightGBM (LGBM), le modèle qui nous a permis d'obtenir le meilleur score.

4.1 LightGBM

LightGBM, introduit par Microsoft, est un cadre de renforcement de gradient qui utilise un algorithme d'apprentissage basé sur un arbre introduit en 2017. On dit qu'il est plus rapide que les autres méthodes basées sur un arbre, car il fait grandir l'arbre verticalement (profondeur). Il choisira une feuille à pousser (croissance par feuille) avec une perte maximale de delta.

En outre, il est populaire car il se concentre sur la précision des résultats et traite de grandes quantités de données mieux que les autres méthodes. LightGBM introduit deux techniques qui lui donnent la vitesse et la légèreté : l'échantillonnage unilatéral basé sur **le dégradé (GOSS)** et **le regroupement de fonctionnalités exclusives (EFB)**.

En effet, Le principal avantage du modèle LGBM est qu'il choisit la feuille dont la perte delta est la plus élevée pour croître. Ainsi, l'utilisation du LGBM peut nous aider à réduire les pertes dans une plus large mesure que n'importe quel algorithme par niveau.

Le schéma ci-dessous explique la mise en œuvre de LightGBM :

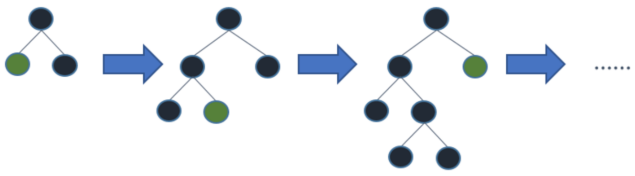


FIGURE 14 – le fonctionnement du LightGBM

4.2 Entraînement du modèle :

Nous avons divisé les données en 80 % pour l'entraînement et 20 % pour la validation afin d'entraîner et valider notre modèle.

Les hyperparamètres ont été choisi aléatoirement (sur la base des valeurs recommandées). Nous pensons que LightGBM sera suffisamment performant, sans qu'il soit nécessaire de faire le tunning des hyperparamètres.

hyperparamètres :

- **objective :**'regression'
- **boosting_type :**'gbdt'

- **metric :**'rmse'
- **n_jobs :**-1
- **learning_rate :**0.07
- **num_leaves :**2**8
- **max_depth :**-1
- **tree_learner :**'serial'
- **colsample_bytree :**0.7
- **subsample_freq :**1
- **subsample :**0.5
- **n_estimators :**8500
- **max_bin :**255
- **verbose :**1
- **seed :**SEED
- **early_stopping_rounds :**3500

Le schéma suivant illustre les étapes du travail effectué :

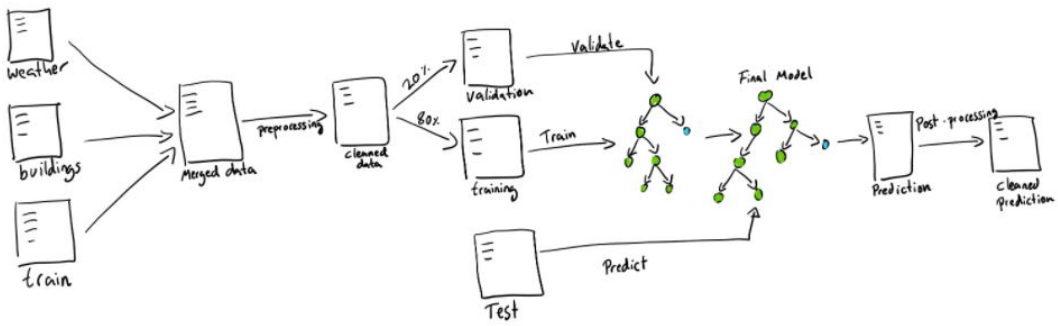


FIGURE 15 – Résume

4.3 Importance des variables

Le graphique ci-dessous montre l'importance des variables explicatives dans la prédiction de notre variable cible via le modèle entraîné. Nous retrouvons les résultats de la partie descriptive : la surface des bâtiments et la température contribuent le plus dans l'explication de la consommation énergétique.

Des valeurs ont été prédites négatives donc nous leur avons donné la valeur 0.

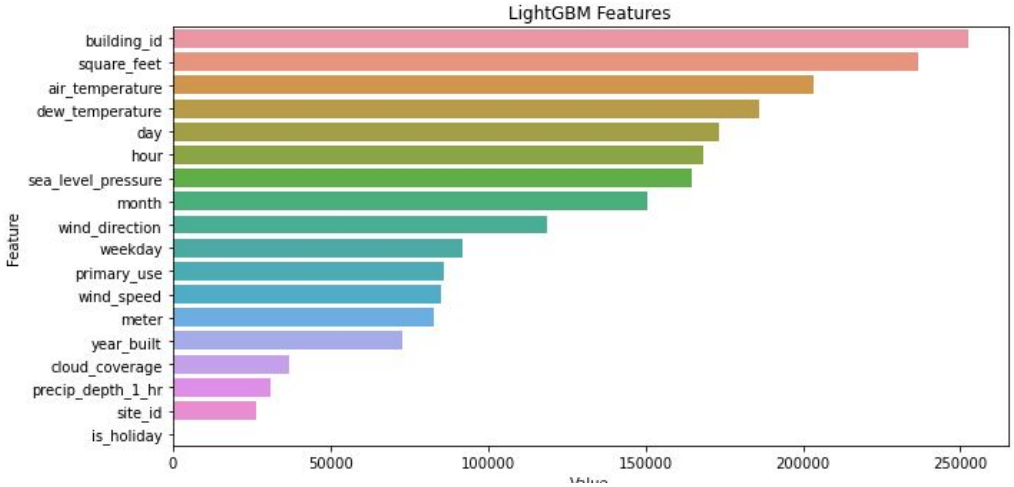


FIGURE 16 – L'importance des variables

4.4 Résultats du modèle :

Après avoir entraîné le modèle, les résultats étaient bien améliorés par rapport à avant et nous nous retrouvons avec un **RMSLE de validation de 0.18478**. De plus, après avoir soumis les résultats, le score obtenu est de 1,190.

FIGURE 17 – Meilleur score obtenu

5 Conclusion

Ce projet était l'occasion de mettre en pratique l'ensemble de nos connaissances en machine learning sur une problématique réelle et avec des données très volumineuses. La plus grande difficulté à laquelle nous avons dû faire face a été la phase de traitement des données ainsi que la partie preprocessing. De plus, nous avons été confrontés à des problèmes de mémoire de stockage, ce qui nous a empêché de tester plus de modèles et d'optimiser les hyperparamètres de chacun.