Reconstitution d'un historique de consommation électrique sur 1 an, pour deux sites

Mahamadou Klanan Diarra

Sorbonne Université

10 octobre 2019

SORBONNE UNIVERSITÉ

SORBONNE UNIVERSITÉ

Sommaire

1. Introduction
2. Présentation du module
3. Données
4. Pré-traitement
   - Labellisation
   - Remplacement des nan
5. Construction du modèle
   - Modèle
   - Évaluation
6. Conclusion

Introduction

- **Challenge data** : plateforme en ligne qui donne accès à des données d'apprentissage automatique supervisé à des fins éducatives.
- **Planet OUI** : Fournisseur d'électricité verte, créé en 2007 et passe filiale de **BCM energy** en 2016.
- Projet d'apprentissage automatique : Analysé les données et prédire.
- Données (Planet OUI) : météorologiques et de consommations électriques à télécharger sur la plateforme challenge data
- L'objectif de ce challenge est de construire un modèle de prédiction de la consommation d'électricité sur deux sites (Lille, Aix-en-province)

Présentation

- Big data et ses applications : Présenté par **Georges Uzbelger**
- Les grandes branches du Big Data :
  1. IA (1956) : Modélisation/simulation de comportements cognitifs
  2. Machine learning : ensemble d'outils **Statistique** + **Optimisation**
  3. Data science (2008) : analyse statistique multivariée sur des données de grande taille
- Big Data dans l'entreprise
- Gestion d'un projet Data :
  1. Comprendre les données
  2. Préparation des données
  3. Modélisation
  4. Évaluation
  5. Déploiement

Données

Pour ce projet nous disposons de 3 tableaux de 8760 observations : 2 fichiers de training (input_train, output_train) et 1 fichier de test (input_test).

- **input (15 colonnes)** : température et humidité pour les 2 sites concernés, consommation de 3 sites secondaires, coordonnées géographiques
- **output (2 colonnes)** : consommations pour les 2 sites d'intérêt.

Données

ID	timestamp	temp_1	temp_2	mean_national_temp	humidity_1	humidity_2	loc_1	loc_2	loc_secondary_1	loc_secondary_2	loc_secondary_3	consumption_secondary_1	consumption_secondary_2	consumption_secondary_3
0	2016-11-01T00:00:00.0	8.3		11.1	95.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	143	74	168
1	2016-11-01T01:00:00.0	8.0		11.1	98.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	141	60	162
2	2016-11-01T02:00:00.0	6.8		11.0	97.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	142	60	164
3	2016-11-01T03:00:00.0	7.5		10.9	99.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	139	60	162
4	2016-11-01T04:00:00.0	6.1		10.8	98.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	154	60	164
5	2016-11-01T05:00:00.0	5.4		10.7	98.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	184	62	167
6	2016-11-01T06:00:00.0	5.0		10.5	99.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	182	78	173
7	2016-11-01T07:00:00.0	5.1		10.2	99.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	165	79	175
8	2016-11-01T08:00:00.0	4.7		10.2	99.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	157	80	176
9	2016-11-01T09:00:00.0	5.5		10.3	100.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	153	86	175
10	2016-11-01T10:00:00.0	8.6		10.7	100.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	149	83	177
11	2016-11-01T11:00:00.0	10.4		11.4	98.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	153	82	176
12	2016-11-01T12:00:00.0	12.4		12.3	88.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	137	75	175
13	2016-11-01T13:00:00.0	13.9		13.2	80.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	139	77	176
14	2016-11-01T14:00:00.0	15.7		13.4	70.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	138	79	176
15	2016-11-01T15:00:00.0	15.0		13.7	78.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	139	77	175
16	2016-11-01T16:00:00.0	14.7		13.9	82.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	150	84	175
17	2016-11-01T17:00:00.0	14.1		13.2	81.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	150	84	174
18	2016-11-01T18:00:00.0	13.5		12.4	87.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	162	89	177
19	2016-11-01T19:00:00.0	12.8		11.6	87.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	163	80	176
20	2016-11-01T20:00:00.0	11.7		11.4	95.0		(50.633, 3.067)	(43.530, 5.447)	(44.838, -0.579)	(47.478, -0.563)	(48.867, 2.333)	154	75	171

(a)

ID	consumption_1	consumption_2
0	100	93
1	101	94
2	100	96
3	101	95
4	100	100
5	100	110
6	115	110
7	145	108
8	141	110
9	143	109
10	145	106
11	144	105
12	143	107
13	138	107
14	137	102
15	136	100
16	136	100
17	147	103
18	142	115
19	140	116
20	137	118

(b)

Figure – input vs output

Labellisation

But : Découper les observations par saison, période de la semaine.
Avantages : Permet une meilleure exploration et traitement des donnée.

- **Calendrier des saisons** : https://www.kalendrier.com/dates-changements-saisons-2016.html
- `days_label = ["{:%A}".format(x) for x in input_train['timestamp']]`
- **Calendrier de vacances** : https://vacances-scolaires.education/annee-2016-2017.php

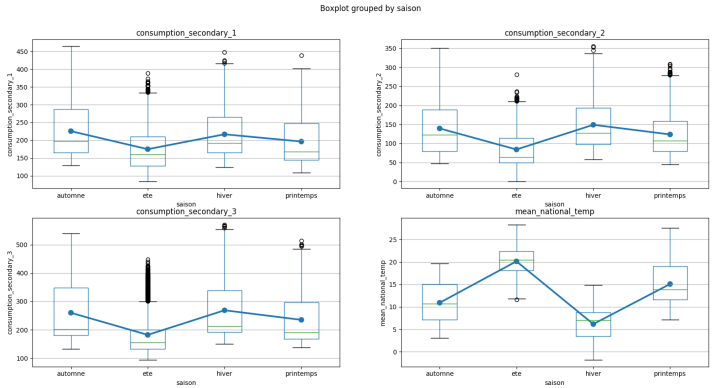
Remplacement des nan

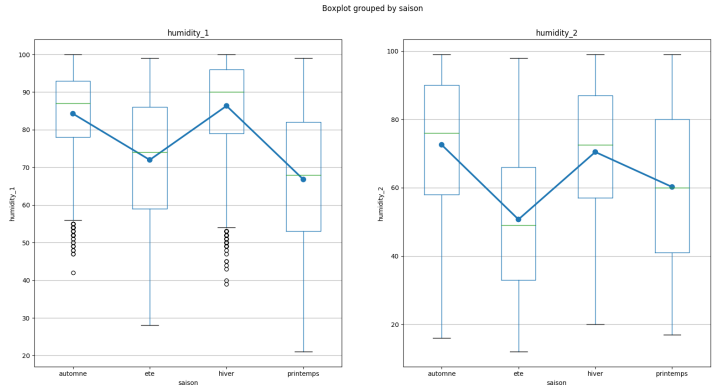
- **3.894 % de nan**
- Supprimés = perte d'informations
- Remplacés par des valeurs adéquates
  1. **K plus proche voisins**
  2. Stratégie : remplacé par la moyenne des K plus proche voisins

ID	timestamp	temp_1	temp_2	mean_national_temp	humidity_1	humidity_2	loc_1	loc_2	loc_secondary_1	loc_secondary_2	loc_secondary_3	nption_secon	nption_secon	nption_secon	saison	days	periode
0	2016-11-01 …	8.3	15.82	11.1	95	85.6	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	143	74	168	autonne	week	vacances
1	2016-11-01 …	8	18.46	11.1	98	65.4	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	141	60	162	autonne	week	vacances
2	2016-11-01 …	6.8	16.24	11	97	71.8	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	142	60	164	autonne	week	vacances
3	2016-11-01 …	7.5	19.08	10.9	99	66.6	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	139	60	162	autonne	week	vacances
4	2016-11-01 …	6.1	19.96	10.8	98	57.8	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	154	60	164	autonne	week	vacances
5	2016-11-01 …	5.4	8.06	10.7	98	74	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	184	62	167	autonne	week	vacances
6	2016-11-01 …	5	12.18	10.5	99	59.4	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	182	78	173	autonne	week	vacances
7	2016-11-01 …	5.1	6.24	10.2	99	85.4	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	165	79	175	autonne	week	vacances
8	2016-11-01 …	4.7	7.88	10.2	99	82.8	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	157	80	176	autonne	week	vacances
9	2016-11-01 …	5.5	8.08	10.3	100	79.2	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	153	86	175	autonne	week	vacances
10	2016-11-01 …	8.6	6.96	10.7	100	79	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	149	83	177	autonne	week	vacances
11	2016-11-01 …	10.4	8.86	11.4	98	76.2	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	153	82	176	autonne	week	vacances
12	2016-11-01 …	12.4	13.56	12.3	88	56.6	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	137	75	175	autonne	week	vacances
13	2016-11-01 …	13.9	14.1	13.2	80	66	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	139	77	176	autonne	week	vacances
14	2016-11-01 …	15.7	13.76	13.4	70	68.8	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	138	79	176	autonne	week	vacances
15	2016-11-01 …	15	15.26	13.7	78	72	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	139	77	175	autonne	week	vacances
16	2016-11-01 …	14.7	10.7	13.9	82	77.4	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	150	84	175	autonne	week	vacances
17	2016-11-01 …	14.1	10.42	13.2	81	78.6	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	150	84	174	autonne	week	vacances
18	2016-11-01 …	13.5	7.5	12.4	87	94	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	162	89	177	autonne	week	vacances
19	2016-11-01 …	12.8	9.18	11.6	87	85.4	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	163	80	176	autonne	week	vacances
20	2016-11-01 …	11.7	8.34	11.4	95	84.8	(50.633, …	(43.530…	(44.838, -0.5…	(47.478, -0.56…	(48.867, 2.333)	154	75	171	autonne	week	vacances

Figure – Tableau complet

Modèle





- Utilisation de l'algorithme du **RandomForests** dans la bibliothèque **sklearn**.
- Entraînement par saison et par tranche de valeur :

Entraînement spécialisé

Combinaison de prédicteurs spécialisés par saison et périodes de faibles consommations.

Random Forest

- Algorithme proposé par **Léo Briemen** en 2001
- La méthode s'appuie sur deux techniques principales :
  1. **Bagging : Boostrap Aggregating**
  2. **CART : Classification and Regression Trees**
- Algorithme très efficace : non paramétrique (s'adapte aux données), prédicteur plus stable (Bagging)

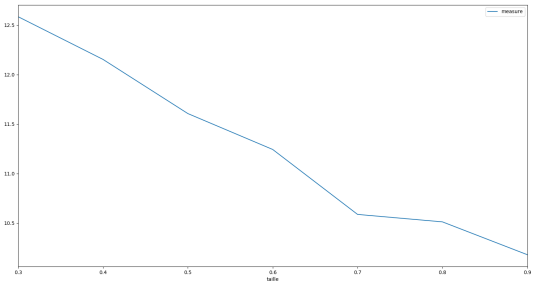


Figure – moyenne des écarts absolues vs taile de l'echantillon de training

- Benchmark **24.33**
- Taille en pourcentage
- **MAE** sur les données d'entraînement est à 12.48 (échantillon d'entraînement : 30%)

Taille vs mae

taille	0.30000000000000004	0.4	0.5	0.6	0.7	0.8	0.9
measure	12.484250520407485	11.976345718731892	11.52014927180635	11.029604356595687	10.43060365714876	10.234570138013474	9.968446641171903

Conclusion

Plusieurs points de conclusion importants :

- Tenue des étapes de suivie d'un projet,
- Une analyse de données cohérente,
- Meilleure utilisation des algorithmes de machine leaning (KNN),
- Une meilleure comprehension du **Random Forest**,
- Un modèle avec de meilleurs résultats

Merci pour votre attention.

Des questions ?