Universite Pierre et Marie Curie
M2 de Statistiques

Année 2016 - 2017

Projet de Modèle linéaire : Prévision de la consommation d'énergie

par

Tehe Yannick et Yahia Antony

Introduction au challenge

Nous souhaitons dans ce travail comprendre au cours du temps l'influence de certaines variables sur la dépense en énergie à l'intérieur de quatre bâtiments.
Ces variables d'énergie sont au nombre de cinq. Nous les noterons Y_l avec $l \in \{1, ..., 5\}$ et nous chercherons à les expliquer.

Pour ce faire nous disposons de variables dites explicatives :

l'instant T auquel on souhaite prédire. T est de la forme: mois/jour/heure.
D'autres variables X_i avec $l \in \{1, ..., 5\}$ supposées utiles à la modélisation.

On modélise un individu par le 12-uplet : $(k, t, x_{1,k,t}, x_{2,k,t}, x_{3,k,t}, x_{4,k,t}, x_{5,k,t}, y_{1,k,t}, y_{2,k,t}, y_{3,k,t}, y_{4,k,t}, y_{5,k,t})$ avec :

$k \in \{1, ..., 4\}$ le bâtiment sur lequel on effectue les mesures.
t la date des mesures effectuées.
$x_{i,k,t}$ la mesure effectuée relativement à X_i sur le bâtiment k à la date $T = t$.
$x_{l,k,t}$ la mesure effectuée relativement à Y_l sur le bâtiment k à la date $T = t$.

Dans le langage de l'apprentissage statistiques, l'ensemble de ces 12-uplets ***définissent nos données d'entraînements***.

Nous disposons aussi de données : $(k, t^*, x^*_{1,k,t}, x^*_{2,k,t}, x^*_{3,k,t}, x^*_{4,k,t})$ pour lesquels nous souhaitons prédire les $y^*_{l,k,t}$. C'est à dire d'un ***échantillon test*** sur lequel nous allons effectuer des prédictions après avoir créer de bons modèles.

Les choses sont désormais claires: Il s'agit de faire les prédictions les plus conformes à la réalité. Le juge de paix étant le score ayant pour valeur la moyenne au carré des écarts entre nos prédictions et la réalité. Sera déclaré gagnant du Challenge le binôme ayant le score le plus faible à l'issu du 8 janvier 2017.

Le Fil rouge de l'exposé

Nous devons réaliser des prédictions, lesquelles se basent sur la créations de modèles découverts en cours de Pratique du modèle linéaire, d'autres en cours d'Apprentissage. D'autres idées nous aurons été inspirées par certaines recherches sur internet.
Néanmoins l'expérience nous a appris que "caler des modèles" était insuffisant, il faut au préalable avoir conscience des objets que l'on manipule (les données). D'ou viennent ils ? A quoi sont ils associées ? La base de donnée est elle complète? Y a t'il des liens de colinéarité entre certaines variables ? ***Etude de la base de donnée*** sera la première partie de cet exposé et celle ci nous servira ensuite de socle pour mieux créer nos modèles.

Une fois que nous aurons fait d'avantage connaissance avec nos données nous chercherons à créer des modèles en vu de faire des prédictions. Néanmoins chaque variable de prédiction Y_l a un mode d'explication unique. Dans une partie ***Modèles et prédictions*** nous détaillerons pour chaque variables Y_l l'évolution de nos modèles et des performances obtenues.Cette partie sera donc décomposée en cinq sous parties.

Enfin sera venue l'heure de prendre du recul sur le travail que nous aurons effectués, faire le bilan de notre étude, penser à des pistes d'amélioration et ouvrir d'avantage. la partie ***Bilan de l'étude clôturera*** ce travail.

Etude de la base de donnée

Variables d'entrées X_l
En introduction nous avons introduit cinq variables d'entrées.Afin de faire des modélisations pertinentes nous pensons qu'il serait utile de mieux connaître ces variables : savoir à quoi elles se rapportent.

L'énoncé du problème indiquait que les variables X_2 et X_5 étaient liées à l'environnement extérieur aux bâtiments. Après recherches nous avons identifié la variable X_2 comme étant la ***température extérieure***(celle ci prenait des valeurs de l'ordre de grandeur d'une température. Durant la période hivernale nous observons même qu'elle prend des valeurs négatives.) La variable X_5 a été plus difficile à identifier car l'énoncé du problème ne donnait aucun indice dessus. Nous sommes allés sur le site de Oze Energie voir si d'éventuels indices s'y trouvaient et nous avons pu identifier cette variable comme étant ***l'irradiance***.La courbe ci joint nous aura permit de faire cette identification.

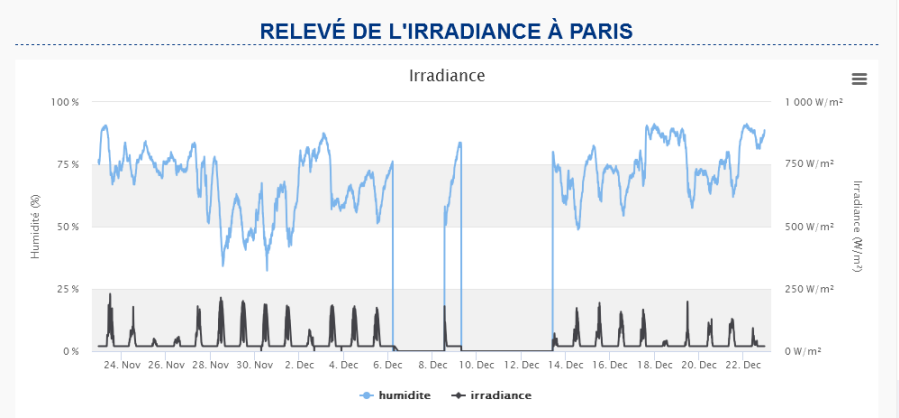

Figure 1: irradiance et humidite au cours du temps pour Paris

Concernant les autres variables, celles ci sont mesurées directement à l'intérieur des bâtiments. X_1 est un ***taux d'occupation***. Cette valeur numérique nous indique comment évolue le nombre de personnes dans le bâtiment. On remarque par exemple que le taux d'occupation est très faible la nuit et se retrouve beaucoup plus élevée entre 7h et 18h (heures de travail). On peut donc penser que les bâtiments étudiés sont des lieux de travail. X_3 est la ***température à l'intérieur*** des bâtiments. D'ailleurs on retrouve respectées les normes d'urbanisme sur la température à l'intérieur des bureaux... X_4 pour le moment n'a pas été identifiée mais l'énoncé laisse à penser que cette variable est liée à ***l'humidité dans l'air***.

Maintenant que nous connaissons mieux nos variables d'entrées nous nous apercevons que celles-ci sont directement ***fonctions de T.***
par exemple l'heure de la journée influe sur le taux d'occupation des bâtiments. L'irradiance est directement liée au fait qu'il fasse nuit ou pas...

On peut penser aussi que certaines variables sont peut être ***liées par des interactions***. Par exemple le taux d'occupation peut avoir une influence sur l'humidité de l'air à l'intérieur du bâtiment à cause de l'eau rejetée lors de la respiration des occupants.

La délicate question du temps T
Le temps est une variable importante de notre problème car celle ci a une influence directe sur nos variables d'entrées et donc par conséquent sur les variables de sorties.
Donnons encore des exemples : le fait d'être en Wee ou pas influe sur le taux d'occupation du bâtiment.
L'heure influe sur la température.
Le mois ou plus globalement la saison agit sur la luminosité et sur la température. . En clair créer des variables dérivées de la variable T. La création des nouvelles variables fera l'objet de la partie qui suit..

Variables supplémentaires
Commençons par expliciter des variables dérivées des X_l que nous utiliserons par la suite dans nos modèles de prédiction :
- ***Les puissances*** de ces dernières : X_l^K pour des modèles de régression polynomiales ou l'entier K désigne la puissance à laquelle on élève X_l.
- ***Les interactions*** entre les X_l que nous détaillerons plus tard à l'intérieur des modèles.

Pour ce qui est du temps T nous allons créer des variables dichotomiques dérivées de celui-çi:
-***24 variables horaires*** : $H_0; H_1; ..; H_i, ...; H_{23}$ prenant pour valeur 1 lorsque l'on se trouve à l'heure i et 0 sinon.
exemple : soit un individu prit à la date $T = 01/01/5heure$. Pour cet individu on a :

$$\begin{cases} H_i = 1 & \text{si } i = 5 \\ H_i = 0 & \text{sinon.} \end{cases}$$

-Nous créons aussi une variable dichotomique W prenant pour valeur 1 lorsque l'individu est en ***période de Wee-end*** et 0 sinon.
-Nous créons une variable P dichotomique de ***présence ou d'absence des gens au travail***. Cette valeur sera prise à 1 durant la période de travail et 0 en période de repos. Nous définissons la plage horaire de travail entre 7h et 18h.
-Enfin deux variables dichotomiques de ***saison*** Hiver et Printemps afin de prendre en considération l'impact saisonnier.

Nous prenons en compte la ***spécificité de chaque bâtiment*** avec quatre variables dichotomiques A,B,C,D.

Dans un problème touchant à la consommation d'énergie, créer toutes ces variables n'est pas un luxe et comme nous le verrons plus tard elles améliorent notre score. ***On apporte de l'information pertinente.***

Le jeu de données est un gruyère
Nous nous rendons compte que les données d'entraînements sont incomplètes.
En fait il y a des valeurs manquantes dans les individus que nous considérons et avant de passer à la création de modèles de prédiction nous devons ***remplacer les*** ? ***par des valeurs réalistes*** qui ne fausseront pas notre modèle. Cette étape est cruciale pour la suite.
Plusieurs solutions se sont offertes à nous:
la première, naïve consistait à remplacer les ? par des moyennes ou bien médianes. De par la disparité des données cette méthode ne donne pas de bons résultats. Les scores obtenus nous l'ont fait abandonner.
Une seconde méthode plus élaborée consistait à prédire les données manquantes en utilisant comme données d'entraînements les individus sans valeurs manquantes par une régression linéaire. Néanmoins certains individus présentent plusieurs inconnues rendant ainsi impossible notre précédé, même par régressions successives.
L'intuition nous conduisait à remplacer ces valeurs manquantes par celles des individus ayant des caractéristiques proches de celles ci. ***Un système à question*** nous paraissait alors idéal. Nous avons fait le rapprochement avec la prédiction par arbres puis de manière ***plus puissante par Random Forest*** comme conseillé en cours d'Apprentissage. Nous avons **utilisé le package missForest du logiciel R**.
Les résulats ont été satisfaisants mais nous nous sommes rendus compte que nous pouvions faire mieux...

En régressant sur les variables dichotomiques décrites plus haut de manière successives de tel sorte que l'enchaînement ait un sens :
Nous avons commencé par une régression sur X_4 qui contenait le plus petit nombre de valeurs manquantes qui après avoir été prédites donnaient plus de sens aux variables d'interactions entre X_4 et les autres variables en ne gardant les prédictions qu'aux endroits à données manquantes.
Par suite nous avons utilisés la variable d'interaction entre X_4 et X_2 ainsi que les variables dichotomiques pour prédire X_2. **N*ous avons itéré ce procédé*** jusqu'à obtenir un jeu de données sans trous.

QUID de la multicolinéarité entre les X_l ?
Nous devons nous assurer qu'il n'y a ***pas de liens de multicolinéarité*** entre les variables explicatives. En clair que l'espace qu'elles engendrent est de dimension le nombre de variables explicatives.

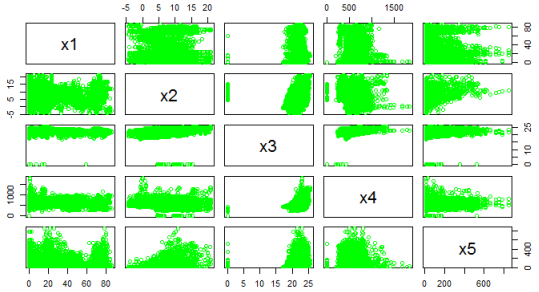

Figure 2: scatterplot des variables X_l

Graphiquement on voit que la non-linéaité entre les X_l est actée comme le montre le scaterplot
Avec les outils vus en cours pour diagnostiquer une éventuelle multicolinéarité (***comme le VIF***) on peut supposer que notre matrice du plan d'expérience est de rang plein.
De toute façon l'objectif ici est d'avantage un objectif de prédiction plutôt que d'explication. On conclue qu'il n'y a pas de multicolinéarité entre les variables explicatives.

Modèles et prédictions : $Y1$ & $Y2$

Méthodes de prédictions utilisées

a) La première méthode que nous avons utilisé est une ***régression linéaire simple*** sur les variables : X1 , X2 , X3 , X4 , A , B , C , D , H0 , ... , H23 , P , W , saison, les variables d'intéractions $X_l * X_{l'}$.
les individus aberrants et influents n'ont pas été pris en compte.

b) la méthode précédente a été modifiée en faisant le ***diagnostique des individus aberrants et influents*** via le critère des résidus studentisés et la distance de Cook.

c) Une troisième approche consistait en l'utilisation d'un modèle de ***régression type polynomiale*** en prenant en compte les puissances des variables X_l^2 et en supprimant les individus identifiés en b).

d) Une technique d'Apprentissage supervisée a aussi été testée : ***la Random Forest***.

e) Enfin nous avons essayé une ***méthode de Boosting*** via le package xgboost pour lequel nous avons "tuné" les paramètres.

f) la meilleure méthode de $\{a, b, c, d, e\}$ utilisée seulement sur ***le mois d'Avril***.

Méthodes mises au banc d'essai

bâtiments \ méthodes	a	b	c	d	e	f
1	155,295	155,29	151,858	135,39	23,45	21,9
2	89,229	89,229	85,812	62,12	18,84	14,63
3	NA	NA	NA	NA	NA	NA
4	176,46	176,46	152,764	126,098	24,36	14,2

Figure 3: performances des méthodes de prédiction pour Y_1 sur l'échantillon d'apprentissage

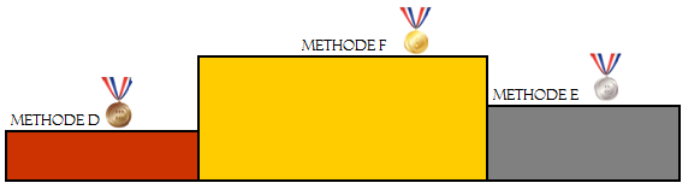

Figure 4: Podium des méthodes de prédiction pour Y_1 sur l'échantillon test

bâtiments \ méthodes	a	b	c	d	e	f
1	NA	NA	NA	NA	NA	NA
2	293,71	293,711	289,04	261,67	189,32	64,81
3	NA	NA	NA	NA	NA	NA
4	2709,862	2330,48	2226,73	1891,61	397,11	107,9

Figure 5: performances des méthodes de prédiction pour Y_2 sur l'échantillon d'apprentissage

Figure 6: Podium des méthodes de prédiction pour Y_2 sur l'échantillon test

Modèles et prédictions : $Y3$

La spécificité de Y3 est que l'on ne prédit uniquement que sur le bâtiment 3.

Méthodes de prédictions utilisées

a) Nous commenons avec un ***modèle linéaire simple*** prenant en compte toutes les variables définies dans le a) de la partie précédente et l'on ***supprime tous les individus pathologiques***.

b) La seconde méthode consistait á s'appuyer sur le modèle précédent en introduisant une nouvelle variable dichotomique ***prenant en compte les très faibles valeurs de*** $Y3$. En effet les très faibles valeurs de Y3 (< 1) n'étaient pas suffisamment prises en compte par le modèle précédent.

c) Nous ajoutons au modèle précédent une nouvelle variable dichotomique permettant de ***capter les individus ayant leur valeur en Y1 nulle***. En effet nous avions remarqué un nombre non négligeable de valeurs nulles en Y1 et nous nous sommes demandés si cela pouvait influer sur Y3.

d) Nous améliorons encore le modèle c) en introduisant un ***modèle linéaire polynomiale*** en allant jusqu'á l'ordre 2. En effet nous avons essayé d'avantage de puissances mais un ***test de Fisher*** nous a dissuadé d'aller plus loin.

e)***La methode de boosting*** a elle aussi été utilisée car celle ci nous avait apporté de précieux points lors de la prédiction de Y1.Neanmoins la recherche de bons paramètres a été longue, très longue(8essais pour faire mieux que d))

f) la meilleure méthode de $\{a, b, c, d, e\}$ utilisée seulement sur ***le mois d'Avril***.

Méthodes mises au banc d'essai

bâtiments \ méthodes	a	b	c	d	e	f
1	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA	NA
3	48,321	46,88	43,189	41,325	42,078	39,74
4	NA	NA	NA	NA	NA	NA

Figure 7: performances des méthodes de prédiction pour Y_3 sur l'échantillon d'apprentissage

Figure 8: performances des méthodes de prédiction pour Y_3 sur l'échantillon Test

Modèles et prédictions : $Y4$

La Spécificité de $Y4$ est l'aspect entier des valeurs qu'il faut prédire.

Méthodes de prédictions utilisées

a) Nous avons tout d'abord pensé á utiliser un ***modèle linéaire généralisé Poissonien*** en prenant en compte toutes la variables dichotomiques habituelles.

b) Au vu des prédictions obtenues par a) l'idée que nous avons eu est de rajouter une nouvelle contrainte au modèle afin que celui ci soit mieux calibré.
contrainte utilisée : ***Fixer au maximum de l'échantillon d'apprentissage les valeurs obtenues en a) supérieures á celui-ci.***

c) la meilleure méthode de $\{a, b\}$ utilisée seulement sur ***le mois d'Avril***.

Méthodes mises au banc d'essai

bâtiments \ méthodes	a	b	c
1	955,846	719,35	874,902
2	NA	NA	NA
3	NA	NA	NA
4	638,21	771,98	489,2

Figure 9: performances des méthodes de prédiction pour Y_4 sur l'échantillon d'apprentissage

Figure 10: performances des méthodes de prédiction pour Y_4 sur l'échantillon Test

Modèles et prédictions : $Y5$

a) Nous avons effectués une ***régression simple*** avec toutes les nouvelles variables supplémentaires utilisées jusqu'à présent.

b) Nous sommes restés sur l'idée a) en effectuant une ***recherche d'individus aberrants*** et en les supprimant.

c) Approche ***Random Forest***

d) Approche ***Gradient boosting***.

e) la meilleure méthode de $\{a, b, c, d\}$ utilisée ***seulement sur le mois d'Avril***.

Méthodes mises au banc d'essai

bâtiments \ méthodes	a	b	c	d	e
1	410,77	385,2	503,67	276,63	245,7
2	2307,75	2294,03	2393	344,84	302,17
3	2767,128	2610,89	2689,27	512,45	482,5
4	249,42	201,35	291,33	115,1	86,83

Figure 11: performances des méthodes de prédiction pour Y_5 sur l'échantillon d'apprentissage

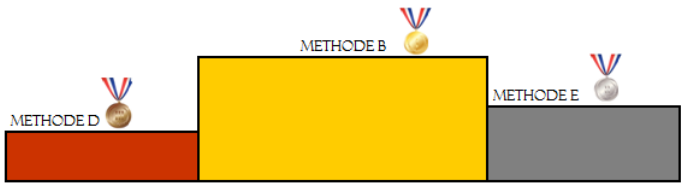

Figure 12: performances des méthodes de prédiction pour Y_5 sur l'échantillon test

Bilan de l'étude

Ce projet a été l'occasion de se pencher sur une réelle problématique type machine learning appliquée à une problématique concrête.

Nous avons tout d'abord appris à prendre ce type de problèmes par le bon bout en effectuant un travail important et minutieux sur la base de données et la compréhension des différentes variables mises eu jeu. Tout d'abord nous avons éclaté la variable T en un grand nombre de variables dérivées ce qui permettait de tenir compte de l'influence du temps sur toutes les variables de l'étude. Nous avons aussi créé des variables d'intéractions sur les X_l.
Au début nous avions rajoutés d'autres variables mais une étude sur la multicolinéarité nous a fait seulement garder celles que nous avons présentés au début du rapport. En effet disposer de variables présentant des liens de multicolinéarité n'apportait pas de bons résultats. Pour résumer nous avons apporté de nouvelles variables contenant le plus d'information viable sans faire de redondance.

A partir de cette base de donnée de "bonne qualité" nous avons pu essayer plusieurs méthodes de prédiction confrontant deux écoles : les méthodes associées au modèle linéaire et d'autres issues du machine learning. Ce travail nous a permis de comprendre les avantages et inconvénients de chacune des approches.

En effet l'approche modèle linéaire s'est trouvée plus explicative et plus facilement maniable. On note aussi un gain de temps significatif dans sa mise en oeuvre. Concernant leur puissance prédictive ces méthodes sont plus difficilement améliorables et nécessitent d'avantage de prises d'initiatives de la part de l'utilisateur. On remarque aussi que les méthodes les plus élaborées ne donnent pas toujours de bons résultats de prédictions (la suppression d'individus aberrants n'a pas toujours été favorable pour la progression du score, les individus influents constituent des données dont nous ne pouvons pas nous passer dans cette étude car ils apportent de l'information précieuse). Concernant la sélection de variable nous avons essayé d'en faire au tout début du projet en utilisant les critères AIC et BIC mais la suppression de variables réduisait fortement les capacités prédictives de nos modèles. Nous en avons conclu que les variables mises à notre disposition dans le challenge avaient déjà fait le fruit d'une sélection par les ingénieurs thermiciens de l'entreprise. Au contraire l'ajout de variables pertinentes était une des clés de ce challenge. Nous pouvons même penser qu'il puisse y avoir encore d'autres variables à ajouter.

Concernant les deux méthodes de machine learning utilisées (Random Forest et gradient boosting) elles ont fournies de très bons résultats de prédiction sur certaines variables mais moins bonnes sur les autres comparativement au modèle linéaire. Les performances de ces méthodes sont très hétérogènes et leur qualité dépend du Y_l que nous souhaitons prédire. Un élément fondamental à prendre en compte lorsque l'on utilise ces méthodes (et surtout le gradient boosting) est l'importance du paramètrage. Pour paramétrer optimalement nous avons effectués une cross validation par variables mais aussi par bâtiments dans les proportions suivantes : 70% de l'échantillon input-train pour entraner notre modèle et les 30% restants pour l'évaluation du modèle. La cross validation nous a permis d'obtenir des modèles de prédiction optimaux tout en évitant de sur-apprendre. Le prix de ce paramétrage correct a été le nombre d'essais et le temps passé sur celui-ci. Néanmoins avec l'expérience nous pouvons espérer être plus rapides à l'avenir. Un des désavantage de ces méthodes est leur aspect "boite-noire" (ils ne sont pas très explicatifs et se révèlent difficiles à être expliqués aux non-spécialistes).

Pour la progression en terme de scores ces deux type d'approches ont été complémentaires. En fonction de la variable Y_l sur laquelle nous souhaitions prédire l'une des deux approches s'est avérée meilleure. Si nous devions désigner la méthode de prédiction qui a donné dans l'ensemble les meilleurs résultats nous dirions que c'est le gradient boosting. Pour mieux considérer la variable T nous pouvons nous interroger sur la pertinence de l'utilisation des séries temporelles pour mieux modéliser notre modèle. Les séries temporelles peuvent elles apporter de l'information non détectées jusqu'à présent ? C'est une des raisons qui nous a motivé à choisir le cours de Séries temporelles au prochain semestre. Il serait intéressant de poursuivre le travail effectué jusqu'à présent en utilisant de nouvelles connaissances.