Modélisation Prédictive Rapport

Valentin Gölz, Laura Fuentes

February 14, 2023

Contents

Introduction **2**

Choix du type de modèle **2**

Linear Models 2

Random Forest 3

Generalized Additif models 3

ARIMA et Kalman Filter 4

Pipeline basée sur le modèle qgam 4

Aggrégation d'experts 4

Introduction

Nous avons un jeu de données regroupant différents variables en rapport avec la consommation énérgétique française pendant la période de 2012 à 2021. Le but ici est de construire un modèle qui permettant de prédire la consommation française en énergie pendant la période du Covid.

Le premier reflèxe est ici de télécharger l'ensemble des packages et divisé le set train en deux pour pouvoir tester nos modèles avant de les soumettre. Nous avons ainsi choisi la période de 2012 - 2019 comme train et 2019-(15/04/2020) comme test.

Choix du type de modèle

Pour comprendre quelles variables sont plus significatives, et argumenter le choix des variables, nous allons effectuer une random forest, et regarder l'importance des variables.

```
### METTRE CODE ICI
```

Nous pouvons ainsi bien remarquer que les variables à plus forte importance sont: Load.1, Load.7, les variables relatives à la temperature, WeekDays, WD, BH toy, Summer_break, DLS and Christmas_break. Nous avons vérifié ce résultat à l'aide de ANOVA. !!!!!!!!!!

Nous avons crée tout d'abord, la variable WeekDays2. Il s'agit d'une version modifiée de la variable WeekDays qui distingue les jours laborales, samedis et dimanches. Nous avons d'autre part recupéré les différents mouvements sociaux et le pourcentage de population mobilisée !!ajouter plus de précisions?!!. Ensuite nous avons créée une variable mesurant la température ressentie. Le problème de cette dernière variable concernait les nombreuses valeurs NA's, ainsi que la représentativité au niveau national des stations météorologiques constituant les données.

Linear Models

Simple linear model

Pour comprendre et apprehender le cadre d'étude on commencera par effectuer un modèle simple. C'est-à-dire un modèle linéaire avec les covariables choisies précédamment. Nous avons considéré que la consommation de la veille changeait en fonction du jour de la semaine. C'est pour cela que nous avons décidé de créer une fonction de la consommation de la veille en fonction de chaque catégorie de WeekDays.

```
mod1 = lm(Load~WeekDays2+Temp, data=Data0[sel_a,])
summary(mod1)
```

```
##
## Call:
## lm(formula = Load ~ WeekDays2 + Temp, data = Data0[sel_a, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16071.4  -3480.3    -21.9   3256.1  18573.0
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       76284.60     294.73  258.825  < 2e-16 ***
```

```
## WeekDays2Monday   -1153.50     324.92   -3.550 0.000391 ***
## WeekDays2Saturday -5298.50     326.07  -16.250  < 2e-16 ***
## WeekDays2Sunday   -8014.38     325.88  -24.593  < 2e-16 ***
## WeekDays2WorkDay    450.09     265.74    1.694 0.090423 .
## Temp              -1572.77      14.36 -109.510  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4708 on 2939 degrees of freedom
## Multiple R-squared:  0.8191, Adjusted R-squared:  0.8188
## F-statistic:  2661 on 5 and 2939 DF,  p-value: < 2.2e-16
```

Polynomial transformations

Nous avons mis en place des transformations polynomiales sur le modèle linéaire comme une première approche de compléxification du modèle.

```
#mettre
```

Random Forest

Generalized Additif models

Choix de la partie linéaire et spline

Dans la suite, nous avons ainsi considéré de mettre en place des Modèles Additifs généralisés. Pour cela, nous avons d'abord distingué les variables à mettre dans la partie linéaire du modèle, puis dans la partie spline. Nous avons intégré ainsi les variables qualitatives ainsi que la consommation de la veille en fonction du jours de la semaine dans la partie linéaire. On a ajouté ainsi dans la partie spline les variables ayant une notion de temporalité comme la consommation de la semaine ou les températures. Nous avons également regroupé dans un même spline des variables ayant une relation logique, comme c'est le cas de la température et le Temps ou les Temperatures_s99 min et max. Nous avons également crée une fonction spline pour chaque jour de la semaine pour la variable toy pour ne pas négliger l'effet des jours de la semaine sur la consommation annuelle.

```
#mettre
```

Pour améliorer le rendement du modèle, nous avons tenté de comprendre l'origine des erreurs à partir des courbes de consommation. Nous avons constaté que les erreurs commencent à s'accentuer au niveau du mois de mars 2020, juste au niveau du début de la période Covid. Ceci s'explique du fait que la variable GovernementResponseIndex comprends des valeurs nulles pendant des années, et celles-ci explosent dans une courte période d'un mois, laissant peu de temps d'entrainement sur la pandémie. Pour simuler le comportement de la population pendant le confinement avec les données que l'on avait déjà, nous avons pensé aux samedis. En effet, nous avons mis l'hypothèse qu'un jour de confinement était comparable en termes de consommation à un jour de weekend comme un samedi. Dans cet esprit, nous avons créé la variable WD, qui modifie le jour de la semaine à samedi s'il y a confinement (la GovernementResponseIndex>=70), et maintien des jours de la semaine sinon.

Nous avons également utilisé la fonction gam.check pour améliorer le rendement du modèle gam. Celle-ci nous a permis d'ajuster la dimension des bases spline. Nous avons ainsi incrémenté les valeurs de k quand la p-value était très petite. Pour la variable toy, on a !!!!!! Nous avons également vérifié que les résidus étaient bien gaussiens à chaque fois à partir de l'histogramme issu du plot.

```
#mettre code
```

GAM et régréssion quantile: qgam

Dans la suite on utilisera le package qgam, et en particulier la fonction qgam. Celle-ci ajuste un modèle additif ainsi qu'une régréssion quantile sur un unique quantile. On utilise ici la même equation qu'auparavant, il suffit juste d'ajuster la variable "qu", correspondant au quantile. Après plusieurs essais, nous avons remarqué qu'on obtenait des meilleurs résultats avec "qu" autuours de 0.4. En effet, en fixant le quantile à 0.4, on change la fonction de perte. On introduit ainsi un biais, qui permet de s'ajuster mieux aux données lors de la période du covid.

```
#mettre
```

ARIMA et Kalman Filter

Pipeline basée sur le modèle qgam

Étant arrivés au bout des amélioration de qgam, nous avons considérer d'autres modèles vus en cours pour comparer les performances. On a ainsi décidé de garder notre équation sur la qgam et de l'implémenter ensuite sur d'autres modèles. On a ainsi d'étudier les résidus. Ceci va ainsi nous permettre de ???? \ Nous avons ainsi décidé de tester les forêts aléatoires sur les résidus de qgam. Après avoir appliqué l'effet des forêts aléatoires sur le modèle, nous avons amélioré davantage la performance à l'aide de Arima.

```
#mettre
```

Aggrégation d'experts

Comme dernière méthode, nous avons décidé de mettre en place un aggrégation d'experts pour extraire une combinaison de prédicteurs qui puissent améliorer davantage la performance du modèle. Pour cela, nous avons regroupé les différents prédicteurs dans une variable experts. Dans cette aggrégation d'experts, nous avons utilisé les différents modèles qgam (avec et sans arima), une forêt aléatoire comprenant toutes les variables, ainsi qu'un filtre kalman. Nous avions déjà essayer un tel élément, mais les résultats n'étaient pas assez satisfaisants.
Le modèle obtenu par combinaison des différents prédicteurs obtient une performance bien meilleure que celle obtenue auparavant.

%%% A Rajouter les critères utilisés pour des bons modèles.