VEILLE SUR MODÉLES

Contexte de travail sur la veille:

Nous sommes dans un contexte d'un poste basé chez un fournisseur d'électricité national, avec une demande de notre manager, il souhaite avoir une prédiction à court terme(une semaine) de la consomation électrique de la région des Hauts de France. De plus il nous indique d'étudier l'incorporation de variables exogénes à nos prédictions.

Afin de répondre au mieux à cette demande, nous allons tout d'abord étudier six des modéles principaux des times séries(ARMA, ARIMA, SARIMA, VARIMAX, Prophet, XGBoost), réaliser une veille dessus puis sélectionner le modèle le plus adapté à ce contexte.

PRICIPALES NOTIONS ABORDÉES

Plusieurs concepts vont revenir parmis les différents modéles pour les définir, dans un premier temps nous allons les définir avant de les utiliser pour décrire les particularités de chaque modèle.

Auto-régression :

>fait référence à une façon de modéliser une variable chronologique en fonction de ses propres valeurs passées, via une focntion linéaire, au lieu de variables indépendantes externes.

La moyenne mobile :

>l'utilité de la moyenne mobile est la réduction du bruit en rapport à la saisonnalité. De cette façon on va définir l'environnement pour le calcul des points x de la série temporelle en fonction de la saisonalité observée de celle-ci.

Par exemple si on observe une saisonalité d'une semaine, prendre les 7 points environants va nous permettre que chaque point absorbe un jour de la semaine, et ne soit pas influencé par des jours à données significatives.

Stationarité :

>va désigner l'absence de trend et de saison.

Pour dire qu’une série est stationaire, on va s'appuyer sur la moyenne et la variance de plusieurs intervalles de temps de notre serie temporelle (chaque intervalle englobant plusieures données). Si cette moyenne et variance restent globalement stable: on dira que la serie stationnaire. De plus pendant les différentes intervalles observées il faut que l’intégralité de ces intervalle soient globalement stable en moyenne et en variance.

Afin de définir la stationnarité de la série on va pouvoir se baser sur un test de stationnarité tel que Dickey-Fuller augmenté (ADF) ou le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

la saisonalité :

>des intervalles de temps qui englobent un même nombre de données et se répètent de manière régulière, se manifeste par une répétition régulière dans la courbe de valeurs.

la tendance :

>c'est l'augmentation ou la diminution des valeurs d'une serie temporelle en fonction du temps, cette variation se caractérisera par du long terme avec une tendance qui persistera dans le temps.

la différenciation :

>Son rôle est de rendre une série stationnaire. Au lieu de prendre comme point de départ la valeure « originale », nous allons définir que pour chaque point temporel T, sa valeur sera égale à la différence équivalente à T – T-1 et ainsi de suite pour chaque point.

Si une série n’est pas stationnaire aprés une première différenciation, je vais pouvoir en effecteur plusieurs jusqu’à ce qu’elle le soit tout en essayant toujours d’avoir un nombre de différénciations minimal, une valeur de différenciation excessive pouvant rendre la série difficile à interpréter et à modéliserOn va se servir de l’autocorélation partielle pour définir le nombre de différenciations à appliquer.

Etude des différents modéles

ARMA

Ce modèle va combiner 2 modéles : l’autorégression et la moyenne mobile. On va se servir des valeures issues de la moyenne mobile pour effectuer l’autorégression de notre série temporelle.

Critéres d'utilisation et adaptation à notre contexte : Arma va être adapté pour les séries temporelles stationnaires, c'est à dire des séries dont la moyenne et la variance vont rester à une certaine stabilité au cours de l'évolution de la série.

Notre série n'étant pas stationnaire et Arma ne proposant pas l'utilisation de série temporelle exogéne, elle ne sera pas adaptée à la demande faîte.

ARIMA

Le modèle Arima va se baser sur les mêmes concepts qu'Arma, à savoir l'autorégression et la moyenne mobile, mais va s'adapter à des séries temporelles non-stationnaires.

Ainsi le « I » de Arima fait référence à la différenciation de la série temporelle que l'on va utiliser pour rendre la série stationnaire.

Dés lors de la stationnarité de la série, la modélisation de la série va pouvoir être effectuée via une modélisation, qu'on réadaptera à nos valeures initiales pour avoir des prévisions correctes.

Critéres d'utilisation et adaptation à notre contexte: l'avantage d'Arima va donc consister à la non impérativité d'avoir une série stationnaire pour être appliqué, n'est aps adapté à des séries avec saisonalité ou un usage nécessitant une série exogéne.

Ce modèle va s'adapter davantage à notre contexte car accepte notre série non-stationnaire, mais le modèle est limité pour accepter des séries exogénes.

SARIMA

Sarima va se baser sur le modèle Arima en incorporant en plus de passer outre la saisonalité des saisons.

La moyenne mobile va nous permettre de parer à la saisonalité, mais si la saisonalité est complexe ou varie d'une année sur l'autre, la prise en compte de la saisonalité par le modèle Sarima sera nécessaire, il apportera une prise en compte robuste de la saisonnalité.

Les différents hyperparamétres de Sarima :

-p :Combien d'observations temporelles précédentes à T0 je dois inclure pour que la composante autorégressive de mon modéle fonctionne.

-d(*ordre de la différence non-saisonnière)* :C'est le nombre de différenciation que je vais exécuter, le nombre de niveau de différenciation. On va pouvoir le définir par rapport à la forme de la courbe de notre série temporelle(linéaire, parabolique etc).

-**q**(*ordre de la moyenne mobile non saisonnière*) :q représente le nombre d'observations temporelles à inclure dans le modèle pour la composante de moyenne mobile (. Cela capture les erreurs résiduelles de la série après avoir tenu compte des lags

de l'autorégression. L’examen de la fonction d'autocorrélation (ACF) peut aider à déterminer la valeur appropriée de q, en s’arrêtant quand la donnée temporelle n’est plus significativement supérieure à 0.

-P (*ordre de l'autorégression saisonnière*): Nombre de valeurs nécessaires pr modéliser la saisonalité, les observations que l’on va prendre en compte pour observer une éventuelle saisonalité dans notre série temporelle.

-D(*ordre de la différence saisonnière*) :nombre de différences saisonnières nécessaires pour rendre la série chronologique stationnaire.

-Q (*ordre de la moyenne mobile saisonière*) :le nombre de valeurs qui vont definir la saisonalité observée. On va pouvoir déterminer ce nombre grâce à l’autocoréaltion partielle. Celle- qui va avoir comme apport par rapport à une autocorélation classique de faire abstraction des corélations entre les Tx-x entre eux, pour évaluer leur corrélation à Tx.

-m(*période saisonnière*) : représente le nombre de pas identifié entre chaque période saisonnière. C’est le nombre d’observations dans une saison. Ce paramétre va influencer les autres paramétres.

Critéres d’utilisation et adaptation à notre contexte:

Etant donné notre série temporelle qui contient de la de la non-stationarité et de la saisonalité,, ce sera adapté à notre contexte, cependant il n’accepte pas de séries exogénes.

VARMAX

(*Vector Autoregression Moving-Average with Exogenous Regressors*)

Le Varmax va permetre d’inclure des données exogénes à notre série temporelle. Ainsi ce modèle va aider la modélisation de la série temporelle actuelle en vectorisant chaque donnée avec des données d’une autre série temporelle, à même date observées.

Varmax va donc traiter plusieurs séries temporelles comme un systéme de vecteurs. Chaque série temporelle est considérée comme une composante d’un vecteur.

Le modèle VARMAX prend en compte à la fois les relations autoregressives (l'influence de chaque série temporelle sur elle-même à différents retards) et les relations de régression croisée (l'influence de chaque série sur les autres séries) dans le système.

Critéres d'utilisation et adaptation à notre contexte:

Les atouts de Varmax sont la prise en compte de variable exogéne, ainsi son principal avantage pour notre contexte va être de pouvoir inclure les températures par exemple.

PROPHET

Le modèle Prophet est un modèle de séries temporelles développé par Facebook conçu pour gérer des séries temporelles avec des saisons et des vacances.

L'une des particularités sur des séries temporelles saisonniére est que celles-ci, outre les saisons vont souvent être confrontée à des vacances, c'est à dire des événements particuliers qui vont influer directement sur l'intervalle observé. En résumé une mini saisonalité à l'intérieur de la saisonalité observée, mais cette mini-saisonalité ne se définit pas nécessairement à l'intérieur de l'espace temps de la saisonalité définie.
En outre, il peut gérer des données manquantes, des points aberrants et des tendances non linéaires. Il permet également la prise en compte se séries exogénes.
Ce modèle est simple à utiliser et à comprendre, ce qui le rend adapté pour un besoin de prévisions rapides et précises pour des horizons à court terme.
Il prend moins en compte les prévisions à long terme.

Critéres d'utilisation et adaptation à notre contexte:

En définitive, par rapport à notre contexte qui veut une prévision sur une intervalle d'une semaine, Prophet va être adapté à cela par rapport à sa prise en compte de vacances, qui va le rendre plus précis , ainsi que de la saisonalité et de séries exogénes. Aussi sa caractéristique à être particuliérement fiable pour des prévisions court termistes va être adaptée ici.

XGBOOST

XGBOOST est un alorythme de Machine Learning mais qui va pouvoir être aussi utiliser pour prévoir séries temporelles avec l'utilisation de variables exogénes.

De plus il va être très adapté pour gérer des quantiter de données importantes en capturant des modéles complexes. Pour l'utilisation de ce modèle il va être essentiel de le

transformer en un probléme d'apprentissage supervisé, et donc de transformer nos données en fonction.

On va donc utiliser les données d'un nombre de jours J précédents pour prédire notre J0.Cette technique d'apprentissage dans XGBOOST est nommée la validation pas à pas, avec une évaluation de notre modèle de façon séquentielle. Hors contexte des séries temporelles l'usage de la validation croisée peut être adaptée, mais ici faire usage de la validation croisée ne va pas permettre de prendre en compte la structure chronologique de nos données dans les times séries.

L'entrainement initial va consister à entrainer le modèle sur les données les plus anciennes qui existent(dans le temps).)P uis on va utiliser ce modèle initial pour prédire une période future des données à disposition, donc mettre en place un label Y, pour la mise en place de cet apprentissage supervisé. Ensuite on évalue et on met à jour le modèle en focntion du résultat en construisant notre modèle sur d'autres intervalles de données, on « régle » le modèle.

Critéres d'utilisation et adaptation à notre contexte:

L'inconvénient de Xgboost va consister par rapport aux autres modéles de sa complexité d'implémentation, mais va prendre en compte tous les critéres nécessaires au traitement de la demande de notre contexte. De cette façon dans notre contexte, nous avons une grande historicité de données, l'utisation d'une variable exogéne avec la température, pour prédire la consommation d'électricité va faire d'XGboost un modèle adapté.

Conclusion aprés étude des différents modéles

En définitive les principaux paramétres de choix du modèle utilisé par rapport à notre contexte vont être les modéles qui intégrent la possibilité d'intégrer une série temporelle non-stationnaire, avec tendance, qui a de la saisonalité, et capable d'intégrer des variables exogénes.

Ainsi parmis les modéles étudiés on retiendra ici Varmax, Prophet, Xgboost. On va pouvoir tester chacun d'eux sur des données antérieures et sélectionner le modèle qui présentera les meilleures performances pour faire une prédiction la semaine prochaine.

SOURCES

-https://towardsdatascience.com/time-series-models-d9266f8ac7b0

> site reconnu pour ses articles écrits par des experts en datascience

> permet d’avoir une introduction de présentation de modéles de time séries

-https://support.sas.com/resources/papers/proceedings20/4306-2020.pdf

>article issue d’une conférence scientifique, écrit par un économétricien

> donne des compléments d’information sur le modèle Varmax

-https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/

>site avec articles rédigé par un expert en apprentissage automatique et en science des données

> fais une revue des prinicipales méthodes des times séries

-https://machinelearningmastery.com/xgboost-for-time-series-forecasting/

>article abordant le modèle XGBoost

-https://www.youtube.com/watch?v=UQQHSbeIaB0

> vidéo abordant de concept de vecteur d’auto-régression

> la source est de qualité, reconnue avec + de 100K abonnés pour des vidéos consacrées au machine learning et la data science.