Agathe Becquart, Formation Développeur IA, Promotion 4, Septembre 2023

Rapport de Veille : Modèles de Time Series

I. Introduction

a. Contexte :

Nous travaillons pour un fournisseur d'électricité. Notre manager souhaite obtenir des prédictions de la consommation électrique pour la région Hauts-de-France sur la semaine à venir.

b. Besoin :

Un modèle d'intelligence artificielle capable de :

- Faire des prédictions à court terme
- D'intégrer la saisonnalité
- D'intégrer des variables exogènes telles que la température qui influe grandement sur la consommation.

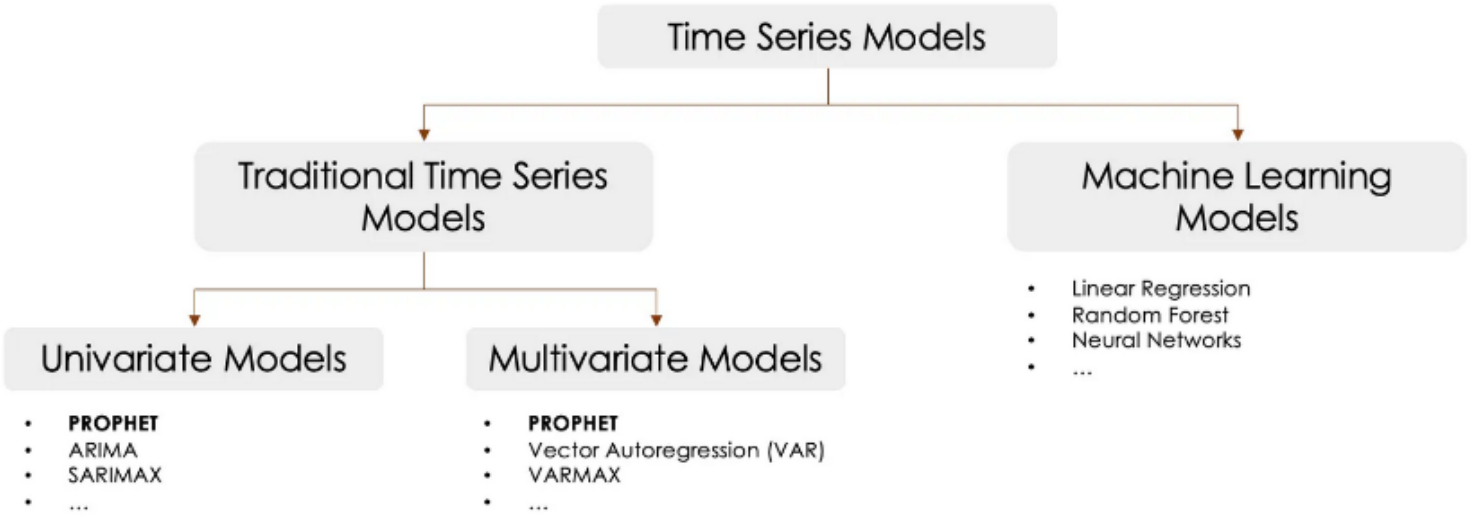
c. Contexte des Séries Temporelles :

Dans le domaine de la prédiction, nous sommes confrontés à un type particulier de données appelé "séries temporelles". Les séries temporelles sont des données qui évoluent dans le temps, où chaque observation est associée à une date.

Dans notre cas, nous sommes en possession de données de consommation électrique qui sont un exemple de séries temporelles, car elles sont enregistrées à intervalles réguliers au fil du temps. Nous orientons donc notre veille sur la recherche d'un modèle qui traite ces séries temporelles et intègre les critères définis précédemment.

II. Modèles de séries temporelles [1]

- **ARMA (AutoRegressive Moving Average)**



ARMA est l'acronyme de Autoregressive Moving Average[2]. Comme son nom l'indique, il s'agit d'une combinaison de deux parties :

- **Autoregressive** : Ce modèle utilise l'historique de la variable que nous essayons de prédire pour estimer sa valeur future.
- **Moving Average** : Ce modèle se concentre sur les erreurs passées, également appelées résidus.

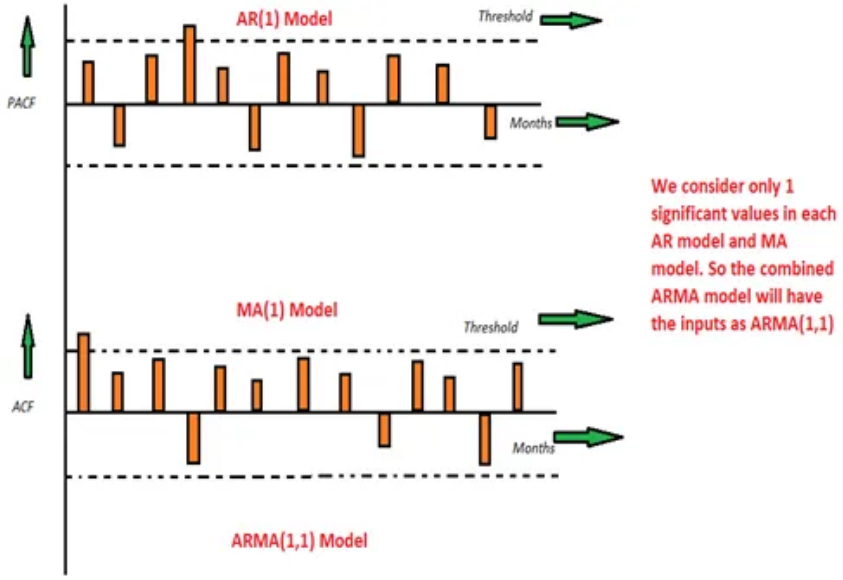
Le modèle ARMA prédit les valeurs futures basées à la fois sur les valeurs et les erreurs précédentes. Ainsi, ARMA a de meilleures performances que les modèles AR et MA seuls[3]. [6]

$$ARMA(p,q) : Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

Les modèles ARMA ont deux paramètres principaux : p et q : [4]

- Le paramètre "p" (Ordre de l'AutoRégressif) représente le nombre de pas en arrière dans le temps que nous regardons pour faire nos prédictions.
- Le paramètre “q” (Ordre de la Moving Average) représente le nombre de valeurs passées de l'erreur que nous utilisons pour prédire la valeur actuelle de la série temporelle.

Déterminer p et q : [2]



- Avantages :
  - Capture les tendances et les cycles
  - Facilité de compréhension et de mise en oeuvre
- Inconvénients :
  - Ne gère pas les données non stationnaires
  - Pas de possibilité d'intégrer des variables exogènes telles que la température.

- **ARIMA (AutoRegressive Integrated Moving Average)**

La différence entre ARMA et ARIMA est la partie intégration [3]. Le I représente le nombre de différences de décalage pour rendre les séries temporelles stationnaires. Les modèles ARIMA sont largement utilisés pour l'analyse des séries chronologiques car la plupart des données des séries de temporelles sont non stationnaires et doivent être différenciées.

Aux paramètres déjà évoqués avec ARMA s'ajoute le paramètre d: [4]

- "d" le nombre de différenciation nécessaire pour rendre la série stationnaire

Déterminer d :

Une des méthodes pour la détermination du paramètre d est basée sur une évaluation visuelle de la série, en examinant si les tendances ou les modèles sont éliminés avec un certain degré de différenciation. [5]

Si la série montre une tendance linéaire, vous pouvez choisir d=1.

Si la série montre une tendance courbée (parabole), vous pouvez choisir d=2.

- Avantages :
  - Comprend la différenciation pour rendre les données stationnaires.
- Inconvénients :
  - Ne tient pas compte des variables exogènes.

- **SARIMA (Seasonal ARIMA)**

Les modèles SARIMA permettent de modéliser des séries qui présentent une saisonnalité. [7]

En amont de la différenciation déjà appliquée pour le modèle ARIMA classique, nous procédons à une décomposition saisonnière.

On effectue une différenciation en "saisonnalité" si des comportements similaires sont observés de manière périodique. Par exemple, si nos données mensuelles de consommation électrique présentent un pic annuel, on utilise une différenciation en "saisonnalité" avec s=12.

  - Avantages :
    - Gère les données saisonnières ce qui se prête très bien à nos données de par les schémas de consommations (augmentation de la consommation en hiver ou encore baisse de la consommation le dimanche).
  - Inconvénients :
    - Ne tient pas compte des variables exogènes.

- **VARIMAX (Vector AutoRegressive Moving Average with Exogenous variables)**

Le modèle VARIMAX est un modèle créé pour combiner trois concepts et connaissances, à savoir le modèle ARIMA mélangé avec le modèle VAR et une variable exogène utilisée dans le modèle. Ce modèle peut être utilisé à la fois pour des prévisions à court terme et à long terme. [8]

VAR, ou "Vector AutoRegressive," est un terme utilisé en économétrie et en modélisation statistique pour décrire un modèle qui traite de multiples séries temporelles simultanées et qui permet de modéliser les interactions et les dépendances entre ces séries temporelles, en examinant comment chaque variable dépend de ses valeurs passées ainsi que des valeurs passées des autres variables endogènes.

Les variables exogènes sont ajoutées dans le modèle VARIMAX sous forme de covariables, c'est-à-dire des variables supplémentaires qui peuvent influencer les variables endogènes. Ces covariables sont ajoutées aux équations du modèle pour prendre en compte leur impact potentiel sur les variables endogènes.

  - Avantages :
    - Possibilité d’ajouter des variables exogènes comme la température.
    - Prévisions à court et long terme
    - Capacité à analyser et modéliser plusieurs séries temporelles en même temps.
  - Inconvénients :
    - Complexité

III. Modèle Prophet (Facebook) :

Prophet est un algorithme de prévision de séries chronologiques développé par l'équipe de Core Data Science de Facebook. Il est conçu pour effectuer des prévisions basées sur un modèle additif, capable de capturer des tendances non linéaires ainsi que des effets saisonniers quotidiens, hebdomadaires et annuels, en plus des effets liés aux jours fériés. [9]

Selon Facebook, Prophet “fonctionne mieux avec des séries chronologiques qui ont de forts effets saisonniers et plusieurs saisons de données historiques et est robuste aux valeurs aberrantes et aux changements de tendance”.

Le Modèle additif se présente ainsi : $y(t)=g(t)+s(t)+\varepsilon(t)$

où y(t) correspond à la modélisation de la série temporelle
g(t) la tendance
s(t) la composante saisonnière
ε(t) la composante aléatoire ou erreur

Prophet ajoute une nouvelle composante qui correspond à l’impact des congés/vacances sur le modèle.

Ainsi, le modèle de décomposition de Prophet est le suivant : $y(t)=g(t)+h(t)+s(t)+\varepsilon(t)$
où h(t) correspond à l’effet vacances (h comme holidays).

Mise en oeuvre :

Les analystes doivent préparer l'ensemble de données et créer un bloc de données avec deux colonnes principales: “horodatage” (au format datetime) et “y” ou la mesure de prévision qui doit être en valeurs numériques, ils ont également la possibilité d’ajouter des variables exogènes qui influent sur le “y”. Ensuite, les analystes doivent créer un objet à partir de la classe Prophet(), dans lequel le DataFrame est intégré à l'objet. Après cela, les analystes peuvent choisir la période souhaitée pour la prévision, puis procéder à la prévision. Le résultat de la prévision comportera plusieurs colonnes dont la colonne “horodatage” et “yhat”. “yhat” est une colonne contenant les résultats prévus de “y” à partir des données historiques. “horodatage”' et “yhat” peuvent être tracés pour montrer des caractéristiques telles que la tendance future ou la saisonnalité. [10]

- Avantages :
  - Facilité d’utilisation, pas besoin de connaissances en statistiques.
  - Gestion avancée des effets saisonniers et tendances ce qui est crucial pour la consommation éléctrique.
  - Intégration de variables exogènes pour améliorer la prédiction.
- Inconvénients :
  - Données de qualité nécessaires
  - Boîte noire (Compréhension très limitée de ce que fait le modèle)

IV. Modèle XGBoost :

- **XGBoost (Extreme Gradient Boosting)**

L'idée de base de l'amélioration des modèles d'apprentissage automatique (boosting) est de combiner des milliers de modèles de prédiction de faible précision en un modèle de haute précision. Avec des paramètres raisonnables, cela nécessite souvent de combiner un certain nombre de modèles pour obtenir une précision de prédiction satisfaisante. Si l'ensemble de données est volumineux ou complexe, le modèle devra peut-être être itéré des milliers de fois, pour obtenir une précision satisfaisante ; le modèle XGBoost peut mieux résoudre ce problème. Le modèle XG Boost a été proposé pour la première fois par Chen Tianqi et Carlos Gestrin en 2011 et a été continuellement optimisé et perfectionné dans le cadre de recherches ultérieures menées par de nombreux scientifiques. XGBoost est une variable efficace et évolutive de la machine d'amplification de gradient.

La fonction objective du modèle L'algorithme du modèle XGBoost est :

$$Obj_m = \sum_{i=1}^{n} l\left(\left(y_i, {y_i}^{m-1}\right) + f_m\left(x_i\right)\right) + \Omega\left(f_m\right)$$

où n représente la taille de l'échantillon,
m représente le nombre d'itérations,
fm représente l'erreur dans les m itérations
l représente la fonction de coût utilisée pour mesurer la différence entre l'étiquette et la prédiction de la dernière étape, ainsi que le résultat du nouvel arbre
Ω est le terme de régularisation qui punit la complexité du nouvel arbre

Le modèle XGBoost est une puissante technique d'apprentissage automatique qui peut être utilisée pour résoudre des problèmes de séries temporelles. Lors de la mise en place d'un modèle XGBoost pour traiter des données temporelles, on utilise les valeurs décalées de la série chronologique en tant qu'entrées pour prédire les valeurs futures.

Puisque de nombreuses séries temporelles présentent des tendances saisonnières, il est souvent judicieux d'inclure des décalages saisonniers dans votre modèle. Par exemple, si votre série temporelle suit une saisonnalité mensuelle, vous pourriez inclure les 12 dernières valeurs pour capturer les tendances saisonnières sur une année.

De plus, il offre la possibilité d'ajuster divers paramètres tels que le nombre d'itérations (nrounds), les méthodes de régularisation (lambda), et les méthodes pour gérer les tendances saisonnières (seas_method) et les tendances temporelles (Trend_method).

- Avantages :
  - Performant pour la prédiction.
  - Gère les variables exogènes et les données non linéaires.
  - Peut capturer des relations complexes.
- Inconvénients :
  - Requiert une configuration des hyperparamètres.
  - Complexité

V. Conclusion / Choix du modèle

- Si la qualité des données est bonne et que la facilité d'utilisation est un critère important, Prophet de Facebook semble être la meilleure option. Il peut gérer efficacement les tendances saisonnières et l'intégration de variables exogènes, ce qui est essentiel pour la prévision de la consommation électrique.
- Si la complexité n'est pas un obstacle et que nous souhaitons une solution plus transparente avec une intégration de variables exogènes, VARIMAX est une alternative viable.
- Si nous avons accès à un ensemble de données volumineux, XGBoost pourrait être envisagé, mais cela nécessite une expertise en configuration des hyperparamètres.

VII. Sources

1. https://towardsdatascience.com/time-series-forecasting-with-facebooks-prophet-in-10-minutes-958bd1caff3f, **Time Series Forecasting with Facebook's Prophet in 10 Minutes** - Guillaume Weingertner, *Mars 2023*
2. https://towardsdatascience.com/time-series-models-d9266f8ac7b0, **Time Series Models, AR, MA, ARMA, ARIMA** - Charanraj Shetty, *Septembre 2020*
3. https://medium.com/@ooemma83/clear-explanations-of-ar-ma-arma-and-arima-in-times-series-analysis-9a72ff569dee, **Clear Explanations of AR, MA, ARMA, and ARIMA in Times Series Analysis** - TrainDataHub, *Novembre 2021*
4. https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/, **Quick way to find p, d and q values for ARIMA** - Yugesh Verma, *Mai 2022*
5. https://avram.perso.univ-pau.fr/sertemp/ser.pdf, **Séries temporelles : régression, et modélisation ARIMA(p,d,q)** - Florin Avram, *décembre 2012*
6. https://blog.devgenius.io/ma-arma-and-arima-models-in-time-series-forecasting-40ad5152a6b9, **MA, ARMA, and ARIMA Models in Time Series Forecasting** - Okan Yenigün, *Août 2020*
7. https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles/5001226-les-processus-non-stationnaires-arima-et-sarima, **Les processus non stationnaires : ARIMA et SARIMA** - OpenClassRoom avec école ENSAE-ENSAI, *Juillet 2020*
8. http://www.jeeng.net/pdf-70200-8365?filename=VARIMAX%20MODEL%20TO%20FORECAST.pdf, **Varimax model to forecast the emission of carbon dioxide from energy consumption in rubber and petroleum industries sectors in Thailand** - Pruethsan Sutthichaimethee - Faculty of Economics, Chulalongkorn University, Thailand, *Mai 2017*
9. https://ledatascientist.com/facebook-prophet-la-prevision-a-grande-echelle/, **Facebook prophet : La prévision à grande échelle** - Henri Michel, *Janvier 2021*
10. https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050921X00026/1-s2.0-S1877050921000417/main.pdf, **Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET** - Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia, *2021*
11. https://bmjopen.bmj.com/content/10/12/e039676, **Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study - Department of Epidemiology, Department of Mathematics, China Medical University, Shenyang, China** - Mirxat Alim, Guo-Hua Ye, Peng Guan , De-Sheng Huang, Bao-Sen Zhou, *Novembre 2020*