Université de Lille

SIAD
Systèmes d'information & aide à la décision

MLOPS

La consommation d'électricité en France Métropolitaine

Enseignant : Corentin Vasseur

Etudiants :
Metuengo Fotso Ange-Cécile
Mamoudou Kaka Abdoul-Kader
Abdou Dangaladima Nana
M2 SIAD DS

Introduction

Éclairage, chauffage, cuisson, téléphone, transports, informatique, santé…, l'électricité est aujourd'hui présente dans toutes les activités de la vie quotidienne. Sa consommation va de pair avec l'amélioration de la qualité de vie des habitants, la création de richesse et le développement des loisirs. Il paraît très difficile de vivre aujourd'hui sans électricité. Celle-ci est indispensable au développement économique, social et industriel dans tous les pays du monde. Elle fait partie des indicateurs permettant de mesurer les écarts de développement entre les différentes régions et représente l'un des enjeux majeurs actuels du développement durable.

Responsable du réseau public de transport d'électricité haute tension en France métropolitaine, RTE (Réseau de transport d'électricité) a pour mission fondamentale d'assurer à tous ses clients (distributeurs d'électricité ou entreprises grosses consommatrices) l'accès à une alimentation électrique économique, sûre et propre.

À cet effet, il exploite, maintient et développe le réseau à haute et très haute tension et est le garant du bon fonctionnement et de la continuité de l'alimentation du réseau en fonction des besoins. Pour ce faire, il lui est nécessaire de construire des modèles de prévision de la demande d'électricité en temps continu.

La mission nous a été confiée d'établir un modèle de prévision, tenant compte des différents processus temporels, mouvements répétitifs et aléatoires afin de prévoir avec le plus d'exactitude possible la consommation quotidienne d'électricité en France dans le but de garantir un meilleur service à sa clientèle.

Afin d'atteindre notre objectif, nous procéderons dans un premier temps à l'étude statistique de notre base de données, afin d'identifier des données manquantes ou d'éventuelles valeurs aberrantes dans notre fichier de données et ainsi juger de la qualité des données. Ensuite, nous nous attèlerons à effectuer une analyse statistique des différents processus temporels (trend et mouvements répétitifs) et aléatoires pouvant avoir une influence sur la consommation électrique en France.

Dans la seconde partie, en nous basant sur les résultats de la première partie nous pourrons entraîner et valider différents modèles, dans le but d'arriver à la prédiction la plus précise et robuste possible.

Etude préalable des données

La base de données qui nous a été fournie est constituée de 4261 observations correspondant aux données de consommation électrique journalière dans la France métropolitaine sur une période s'étalant du 01/01/2008 au 31/08/2019. Ces données correspondent aux données de consommation réelle, issus des outils de suivi de la consommation en temps continu, de la RTE .

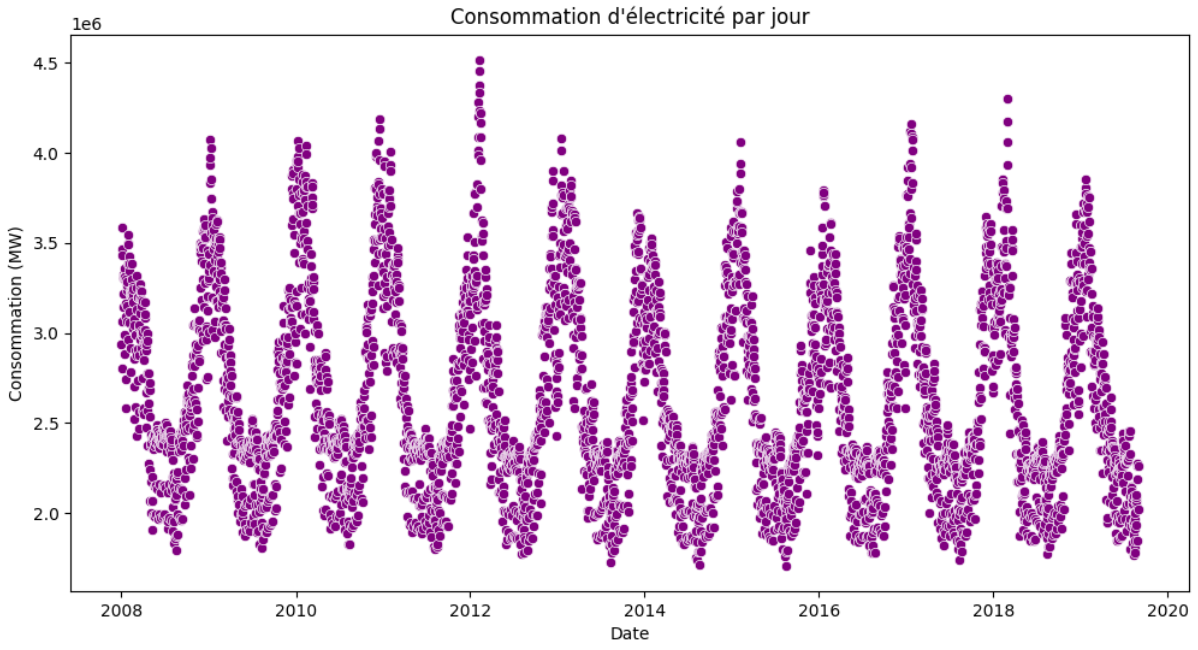
Analyse des données manquantes

Communément, on appelle donnée manquante toute donnée non renseignée dans un fichier de données ; elles sont très souvent le fruit d'un oubli de saisie ou d'une valeur inconnue.
L'analyse préalable de nos données ne nous a pas permis de détecter d'éventuels trous dans notre base de données ce qui revient à dire que pour tous les jours de notre période d'intérêt, nous disposons des données quotidiennes de consommation électrique.

Analyse des valeurs aberrantes

On considère une valeur comme aberrante lorsqu'elle a mal été enregistrée dans le fichier de données, soit par une erreur de frappe, soit une saisie erronée.

Afin de relever les données anormalement élevées, ou anormalement basses dans notre fichier, nous avons choisi d'établir un nuage de points afin d'avoir une vue globale de nos données brutes.



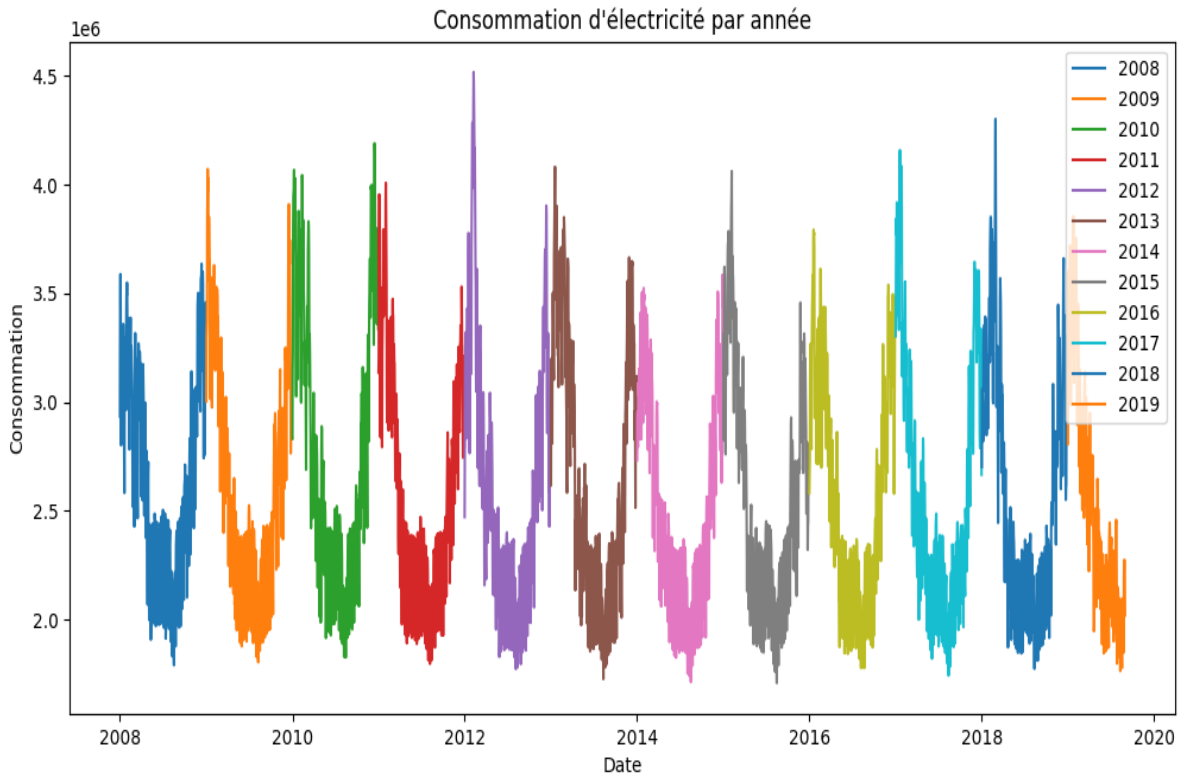
Les données de ce graphique montrent une tendance cyclique marquée, avec des pics réguliers correspondant aux périodes hivernales, où la demande en chauffage augmente. Par exemple, des pics notables sont observés entre janvier et février 2012 attribuables à des températures exceptionnellement basses (-10°C à -20°C). Ces fluctuations s'expliquent par des conditions climatiques rudes augmentant la consommation électrique.

Bien que extrêmes, ces observations sont cohérentes avec les conditions climatiques, ce qui n'en fait pas forcément des données aberrantes.

Evolution de la consommation d'électricité

Ce graphique montre l'évolution de la consommation d'électricité au fil des années, de 2008 à 2019. On observe des pics de consommation chaque hiver, vraisemblablement dus aux besoins de chauffage en période froide.

Les creux de consommation correspondent aux périodes estivales, où la demande en énergie est généralement plus faible.



La forme des courbes est quasiment identique d'une année à l'autre, ce qui met en évidence une forte régularité saisonnière dans les habitudes de consommation d'électricité.

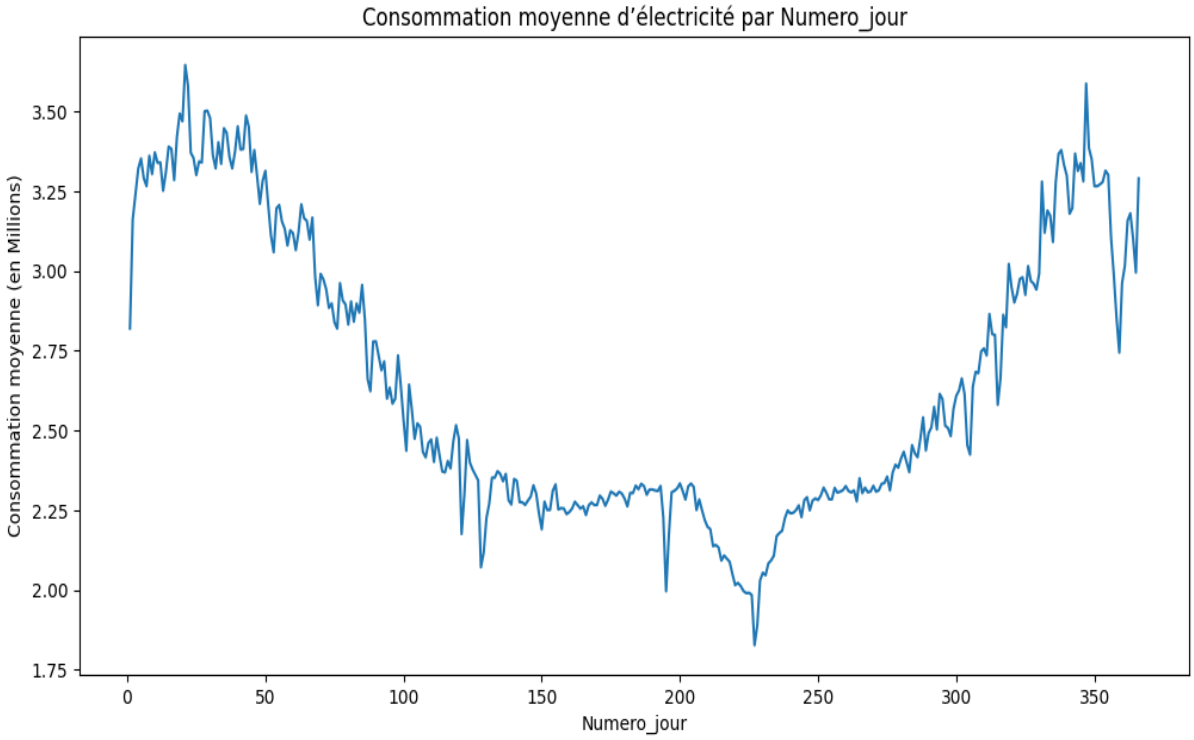
Cette saisonnalité annuelle justifierait l'utilisation de variables temporelles, comme le numéro du jour ou les composantes trigonométriques (sinus et cosinus), pour capturer cette structure cyclique.

On ne note pas de tendance forte à la hausse ou à la baisse sur la période étudiée, ce qui suggère que la consommation reste relativement stable à long terme, bien que des variations interannuelles mineures puissent exister.

La consommation d'électricité durant l'année

Le graphique suivant représente la consommation moyenne d'électricité en fonction du numéro du jour dans l'année. On observe clairement une composante saisonnière marquée :

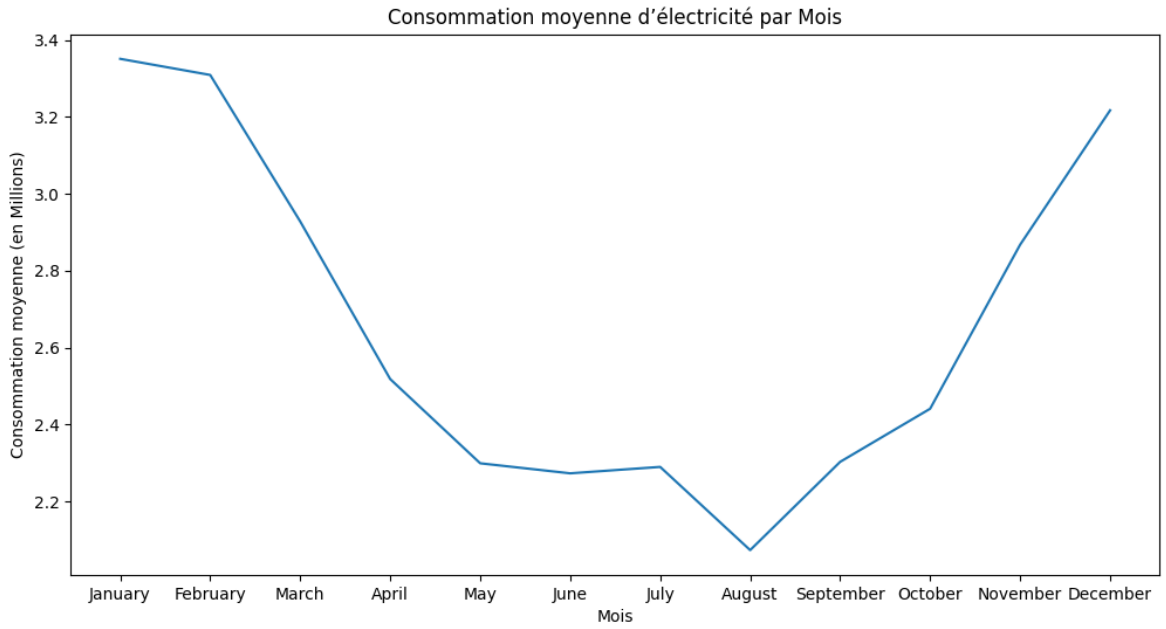
- Début d'année (jours 1 à ~90) : La consommation est élevée, probablement en raison des besoins énergétiques hivernaux, notamment pour le chauffage.
- Printemps et eté (jours ~90 à ~240) : La consommation diminue progressivement, atteignant un minimum pendant l'été, lorsque les besoins de chauffage sont réduits.
- Fin d'année (jours ~240 à 365) : La consommation augmente de nouveau, culminant à l'approche de l'hiver, avec une légère baisse autour des jours correspondant aux fêtes de fin d'année.



On en déduit que le numéro du jour dans l'année encapsule cette saisonnalité naturelle dans la consommation d'électricité. Il pourra permettre au modèle d'identifier les tendances cycliques liées aux variations climatiques et aux comportements humains au fil de l'année.

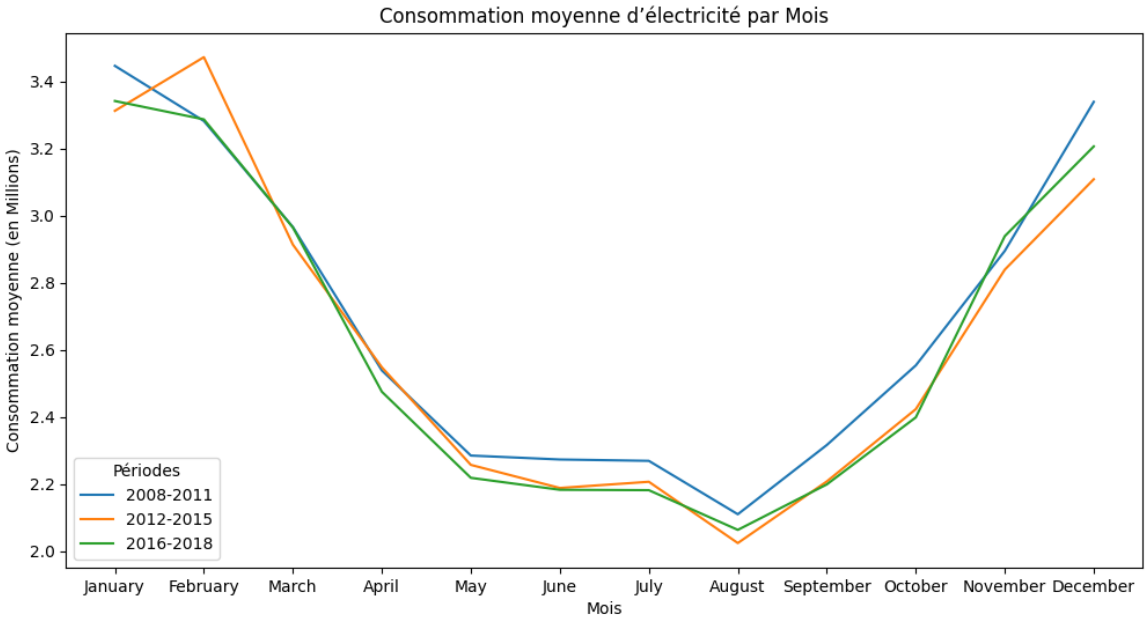
Les mouvements saisonniers sur les mois

Nous avons supposé que la consommation d'électricité en France suit une saisonnalité qui se répète sur les 12 mois de l'année. Afin de vérifier cette hypothèse, nous avons calculé la consommation moyenne d'électricité par mois.



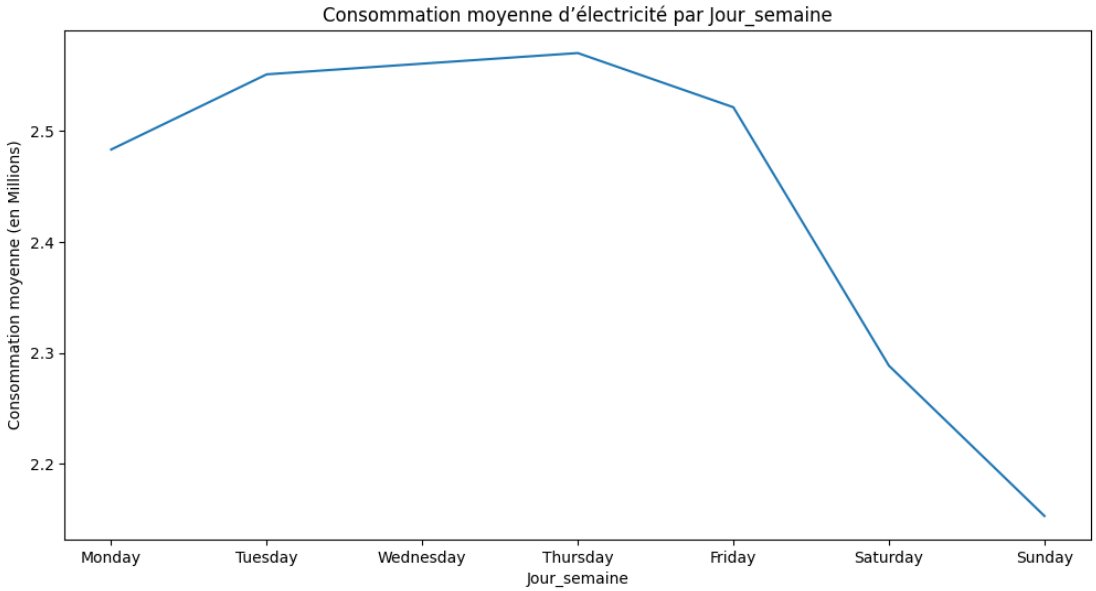
Le graphique ci-dessus illustre la courbe de consommation électrique marquée par une variation saisonnière significative. La consommation est élevée au début et à la fin de l'année, atteignant un pic en hiver, puis diminue progressivement jusqu'à atteindre son point le plus bas en milieu d'année, correspondant aux mois d'été. Cette tendance révèle une corrélation directe entre la consommation d'énergie et les conditions saisonnières, avec une demande accrue en hiver et plus faible en été.

De manière générale, nous observons une régularité et une stabilité dans ce mouvement saisonnier, permettant de d'affirmer que la distribution de la consommation électrique au courant de l'année reste la même quelque soit l'année d'observation.

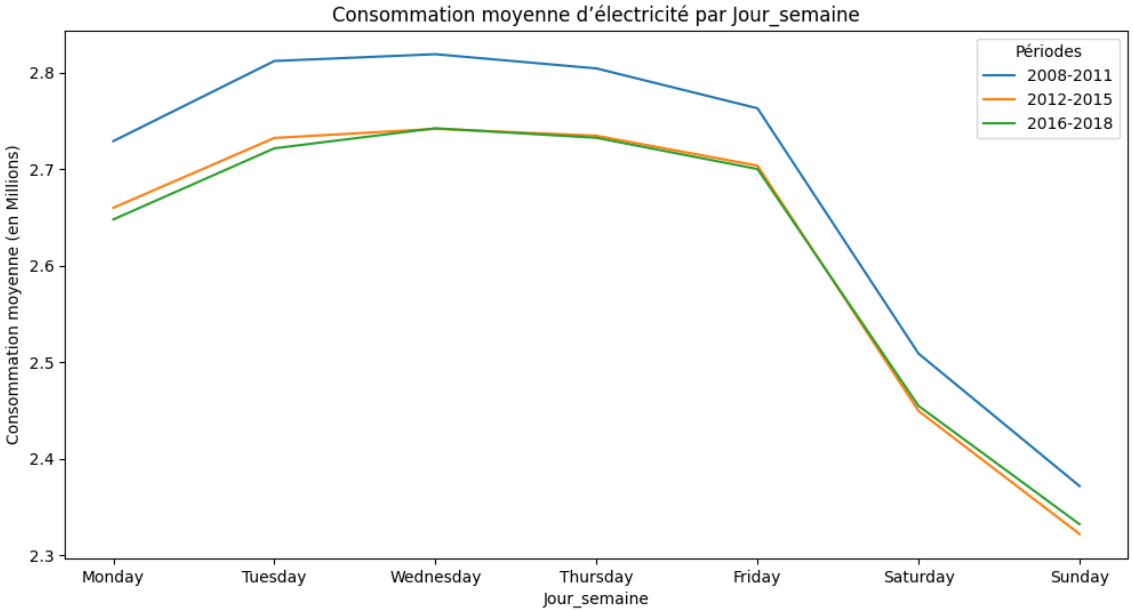


Les mouvements saisonniers sur la semaine

Le graphique illustre une saisonnalité hebdomadaire marquée dans la consommation d'électricité, avec des pics en milieu de semaine (mardi et mercredi) et une décroissance progressive jusqu'à un minimum atteint durant le week-end. Cette dynamique reflète une variabilité intra-semaine fortement corrélée à l'intensité des activités économiques et industrielles.



Une analyse sur des périodes distinctes (2008-2011, 2012-2015, 2016-2018) révèle la persistance de ce schéma, indiquant une structure stable dans la répartition hebdomadaire de la demande.



Les données confirment une asymétrie nette, où les jours ouvrables concentrent les valeurs les plus élevées, et les jours non ouvrables présentent une chute significative, traduisant l'arrêt ou la réduction des activités professionnelles et industrielles.

Modélisation

Les variables

Pour entraîner nos modèles, nous avons choisi d'utiliser les fonctions sinus et cosinus pour représenter la saisonnalité des jours et des mois, plutôt que d'utiliser les jours et mois sous forme directe. Cette décision repose sur les raisons suivantes :

- Cyclicité naturelle : Les jours de la semaine et les mois de l'année suivent un cycle (par exemple, lundi précède toujours mardi, décembre précède janvier). En utilisant les indices numériques (1-7 pour les jours et 1-12 pour les mois), ces transitions sont perçues de manière linéaire (par exemple, la différence entre décembre et janvier est de 11, ce qui n'est pas représentatif d'un cycle). Les fonctions sinus et cosinus, en revanche, capturent cette continuité cyclique, rendant la représentation plus fidèle aux comportements réels.
- Simplicité et efficacité : Si l'on utilisait les jours et mois sous forme directe, cela nécessiterait la création de variables catégorielles ou numériques supplémentaires, comme des indicateurs binaires (dummies). Avec les sinus et cosinus, seulement deux variables suffisent pour modéliser la saisonnalité, réduisant ainsi la complexité du modèle.
- Compatibilité avec les modèles linéaires : Les sinus et cosinus permettent de représenter des relations non linéaires de manière fluide, contrairement aux jours et mois, où une simple variable numérique ne permettrait pas de modéliser correctement les comportements saisonniers.

Ainsi, l'utilisation des sinus et cosinus pour représenter les jours et mois nous permettra de mieux capturer les cycles naturels de temporalité, tout en simplifiant et en rendant le modèle plus efficace.

Partitionnement du jeu de données

Dans le cadre de la prédiction de la consommation d'électricité, la division du jeu de données en ensembles d'entraînement et de test est une étape cruciale pour garantir la robustesse et la fiabilité des modèles.

- Ensemble d'entraînement : Constitué de 80% des données disponibles, cet ensemble sera utilisé pour ajuster les paramètres du modèle. Il permettra au modèle d'apprendre les relations entre les variables explicatives (comme la saisonnalité, les variations quotidiennes, etc.) et la consommation électrique cible.

- Ensemble de test : Les 20% restants serviront à évaluer la capacité du modèle à prédire la consommation électrique sur des données qu'il n'a jamais vues. Cela est essentiel pour s'assurer que le modèle sera capable de généraliser ses prédictions à des périodes futures et ne se limitera pas à mémoriser les données d'entraînement.

Cette approche permettra de détecter et d'éviter le sur-ajustement (overfitting), garantissant que le modèle soit non seulement précis sur les données passées, mais également fiable pour des prédictions sur des données nouvelles.

Les métriques d'évaluation

Pour évaluer la performance des modèles de prédiction de la consommation d'électricité, nous avons retenu deux métriques principales : l'erreur quadratique moyenne et le coefficient de détermination. Ces métriques sont adaptées à notre tâche de prédiction et permettent de comparer objectivement les performances des différents modèles utilisés : régression linéaire, les arbres de décision, et XG Boost.

Le MSE (Mean Squared Error) mesure l'écart moyen entre les valeurs réelles de la consommation d'électricité et celles prédites par le modèle. Elle pénalise fortement les grandes erreurs grâce à l'élévation au carré des écarts, ce qui en fait une métrique particulièrement utile pour détecter des modèles susceptibles de produire des prédictions erronées ou incohérentes. Cette caractéristique est essentielle pour la prédiction de la consommation d'électricité, où des erreurs importantes pourraient entraîner des décisions opérationnelles inadaptées. De plus, en étant exprimée dans la même unité que la variable cible (au carré), le MSE facilitera la comparaison directe entre les différents modèles entraînés.

Le coefficient de détermination évaluera la proportion de la variabilité totale de la consommation d'électricité qui sera expliquée par le modèle.
Une valeur élevée indiquera que le modèle est capable de capturer efficacement les tendances et les variations des données. Le R2 est particulièrement utile pour comparer les modèles entre eux, notamment lorsque leurs structures sont très différentes, comme c'est le cas avec un modèle simple de régression linéaire et un modèle plus complexe comme le XGBoost.
En complément de la MSE, il apportera une perspective relative, permettant de juger si un modèle explique bien les variations globales de la consommation, même si ses prédictions absolues peuvent être légèrement décalées.

Les modèles

La régression linéaire

La régression linéaire est un modèle simple qui peut être utilisé pour établir une relation linéaire entre la consommation d'électricité et des variables explicatives. Elle permet de quantifier l'impact de chaque variable grâce à des coefficients facilement interprétables.
Dans le cadre de notre étude, nous l'utiliserons comme point de départ pour modéliser les relations entre les facteurs tels que les jours et les mois, tout en offrant une base simple pour la comparer avec des modèles plus complexes comme les arbres de décisions ou le XGBoost.

Mean Squared Error	62196382742.90496
Coefficient de détermination	0.67

Le R2 de 0,67 indique que le modèle explique 67 % de la variabilité de la consommation d'électricité, laissant 33 % non expliquée.
Ce résultat suggère que, bien que le modèle capture une partie significative des tendances, il pourrait être limité pour représenter la complexité de la tâche. La consommation d'électricité dépend souvent de relations non linéaires, d'interactions complexes entre les variables, et de phénomènes saisonniers, que la régression linéaire ne peut pas bien modéliser.
Pour améliorer la performance, il serait pertinent d'explorer des modèles plus complexes, comme les arbres de décision ou les modèles basés sur le gradient boosting (par exemple, XG Boost), qui sont mieux adaptés pour capturer ces relations complexes et non linéaires.

Le Random Forest Estimator

Face aux limitations de la régression linéaire dans la modélisation de la consommation d'électricité, notamment son incapacité à capturer les relations non linéaires et les interactions complexes entre variables, il est pertinent d'explorer des modèles plus robustes. L'estimateur Random Forest est une option pertinente pour cette tâche car le il peut capturer les non-linéarités et les interactions entre les variables explicatives. En combinant différentes caractéristiques comme les sinus et cosinus, le modèle peut exploiter les schémas saisonniers plus efficacement.

Grâce à l’agrégation des prédictions de multiples arbres de décision, le modèle peut être moins sensible aux fluctuations aléatoires dans les données.

Mean Squared Error	54250756288.300156
Coefficient de détermination	0.80

Le modèle Random Forest (R^2 de 0,80) surpasse le modèle de régression linéaire (R^2 de 0,67), ce qui signifie que Random Forest est plus efficace pour expliquer la variance des données et qu'il est mieux adapté à la tâche, capturant potentiellement les relations non linéaires complexes entre les variables.

En termes de MSE, Random Forest obtient également un meilleur score (54,25 milliards contre 62,19 milliards pour la régression linéaire). Cela signifie que les erreurs de prédiction sont plus faibles avec Random Forest, et les prédictions sont plus précises.

Ce deuxième modèle est donc plus robuste car il est capable de mieux capturer des relations complexes entre les variables, ce qui le rend adapté à la tâche de prédiction.

Le XG Boost

Malgré l’amélioration de la performance de prédictions avec le Random Forest, le modèle conserve toujours une marge d’erreur significative, indiquant qu’il existe des relations entre les variables qui ne peuvent pas être comprises par les modèles entraînés jusqu’ici.
Il est judicieux de tester un modèle plus complexe, capable de mieux capturer les relations non linéaires et complexes dans les données de consommation d'électricité.
Le XG Boost semble se prêter à la tâche car ce type de modèle utilise un algorithme de boosting, une technique où les modèles sont ajoutés séquentiellement, chaque modèle suivant corrigeant les erreurs du précédent. Cela permet à XG Boost de mieux exploiter les erreurs résiduelles des prédictions passées, ce qui est une approche plus robuste pour obtenir une meilleure généralisation sur de nouveaux jeux de données.
De plus, il est possible d’optimiser les performances du XG Boost, non seulement en ajustant les hyperparamètres, mais aussi grâce au mécanisme interne de régularisation qui permet de contrôler l'overfitting.

Mean Squared Error	39334028295.118454
Coefficient de détermination	0.86

Le XG Boost a un très bon R2, expliquant 86% de la variance des données, ce qui indique qu'il est le modèle le plus performant pour capturer la variabilité de la consommation d'électricité.
On constate en effet que le XG Boost surpasse le Random Forest et la régression linéaire avec les meilleures performances sur les deux métriques : un MSE plus bas et un R^2 plus élevé. Il offre une meilleure précision et une meilleure capacité à expliquer la variance de la consommation d'électricité quotidienne.

Choix du meilleur modèle

En se basant sur les performances des trois modèles, le modèle XG Boost est celui qui offre objectivement des meilleures performances sur de nouvelles données.

Cette meilleure performance peut être expliquée par sa capacité à capturer les relations non linéaires complexes et les interactions entre les variables et sa robustesse compte tenu de l'algorithme sous-jacent.
Ce gain de performance est toutefois associé à une majeure complexité du modèle au vu des nombreux hyper paramètres à ajuster et du temps d'entraînement supérieur à celui des deux autres modèles.

Toutefois, dans le cadre de notre étude, la précision et la robustesse maximales du modèle sont fondamentales car une estimation exacte de la demande est cruciale pour mieux planifier l'approvisionnement et réduire les risques de pénurie ou de surproduction. En effet, si les prévisions sont trop éloignées de la réalité, cela peut entraîner des déséquilibres sur le réseau, augmentant les coûts ou même causant des pannes et une erreur de prédiction, même modérée, pourrait avoir un impact financier important sur la gestion des réseaux électriques, le stockage d'énergie et l'allocation des ressources.

Conclusions

L'étude de la prédiction de la consommation quotidienne d'électricité en France a permis de comparer plusieurs modèles d'apprentissage automatique afin d'identifier celui offrant la meilleure performance en termes de précision et de robustesse. Après avoir testé trois modèles (régression linéaire, Random Forest, et XG Boost), il ressort que XG Boost surpasse les autres avec un R^2 de 0,86 et un MSE significativement plus bas (39,33 milliards), offrant ainsi la meilleure capacité à expliquer la variance des données et à prédire la consommation d'électricité avec une grande précision.

Cette supériorité en termes de performance s'explique par la capacité de XG Boost à capturer les relations non linéaires complexes et les interactions entre les variables, tout en étant robuste face aux variations subtiles et aux anomalies potentielles dans les données. Bien que la complexité de XG Boost demande un ajustement minutieux des hyperparamètres, les résultats montrent clairement que l'amélioration de la précision justifie cet investissement en temps et en ressources.

Dans un contexte où la gestion optimale de la consommation d'électricité est cruciale pour le bon fonctionnement du réseau, la précision et la robustesse des prévisions est essentielle. Par conséquent, choisir XG Boost pour ce type de tâche est justifié par son aptitude à réduire les erreurs de prévision, contribuant ainsi à une gestion plus efficace et durable du réseau énergétique.

En conclusion, cette étude démontre que pour des problématiques complexes comme la prédiction de la consommation d’électricité, où la précision et la fiabilité sont essentielles, l'usage de modèles avancés comme XG Boost est fortement recommandé. Les gains en performance offerts par ce modèle surpassent largement la complexité qu'il introduit, et son adoption pourrait avoir un impact significatif sur l'efficacité de la gestion énergétique à long terme.