École nationale de la statistique et de l'analyse de l'information

Mastère Data Science - Connaissance Client

Projet Séries Temporelles

Thème

Prévision de la Consommation Énergétique en France

Réalisé par
DIAKITE GAOUSSOU

Destiné au professeur
ESSTAFA YOUSSEF

2022/2023

Table des matières

Chapitre 1

Résumé

Ce projet explore la prévision de la consommation énergétique en France, un enjeu crucial dans le contexte de l'efficacité énergétique et de la durabilité. À travers une analyse détaillée des données historiques, nous examinons comment la consommation énergétique est influencée par divers facteurs, notamment la température. Nous utilisons des techniques avancées comme l'analyse exploratoire, le prétraitement des données, et des modèles prédictifs tels que ARIMA et des Modèles d'Apprentissage Automatique. L'objectif est de développer un modèle fiable pour anticiper les besoins énergétiques, optimiser la distribution d'énergie, et comprendre l'impact des variables environnementales.

Chapitre 2

Introduction

Dans un monde où la gestion de l'énergie devient de plus en plus importante, notre projet se concentre sur la prévision de la consommation énergétique en France. Nous analysons des séries temporelles pour identifier des tendances et des modèles, en tenant compte des variations saisonnières et des événements imprévus. En combinant des méthodes de modélisation avancées, notre étude vise à fournir des prévisions précises, contribuant ainsi à une meilleure planification et optimisation des ressources énergétiques. Cette approche innovante offre une perspective approfondie sur la manière dont les facteurs externes influencent la consommation d'énergie.

Chapitre 3

Présentation de la Base de Données

Notre étude se concentre sur une base de données énergétique couvrant la période de 2020 à 2023, détaillant la consommation énergétique en France. Les principales variables étudiées sont la consommation brute d'énergie et la température. Nous intégrons des données complémentaires pour enrichir notre analyse :

- **Date :** La date des enregistrements, utilisée pour observer les tendances sur différentes périodes.
- **Pic Journalier Consommation (MW) :** Mesure le pic quotidien de consommation énergétique en mégawatts.
- **Température Moyenne (°C) :** Température moyenne, un facteur important influençant la consommation d'énergie.
- **Température Référence (°C) :** Pour des comparaisons standardisées avec la température moyenne.
- **Consommation à Température Normale (MW) :** Ajuste la consommation d'énergie à une température de référence.
- **Consommation Brute (MW) :** Mesure principale de notre étude, reflétant la consommation totale d'énergie.

Des fonctionnalités supplémentaires telles que les années, les mois, et les jours fériés seront créées pour aider dans la visualisation et la modélisation des données. Ces variables supplémentaires nous permettront d'examiner la consommation énergétique sous différents angles.

Les données proviennent de sources fiables, accessibles aux liens suivants :

- Pic Journalier de Consommation Brute
- Consommation Quotidienne Corrigée Brute
- Synthèse de la Consommation Brute

Chapitre 4

Préparation et Traitement de la Base de Données

4.1 Préparation de la Base de Données

La première phase consiste à préparer et standardiser la base de données sur la consommation énergétique brute. Cette étape inclut la mise à jour du format des colonnes pour une meilleure lisibilité, le renommage des variables pour une compréhension facilitée, et la conversion des valeurs en mégawatts (MW). Une sélection minutieuse des données pertinentes est également effectuée pour assurer la cohérence et la précision de l'analyse.

4.2 Conversion et Jointure des Bases de Données

Pour une intégration harmonieuse, les colonnes de dates ont été converties au format datetime. L'objectif était de fusionner divers jeux de données en un seul ensemble cohérent et complet. En joignant ces données sur les dates correspondantes, nous avons concentré nos efforts sur la conservation des informations pertinentes communes à tous les jeux. Cette étape essentielle a permis de créer une base de données consolidée, riche de 1430 lignes et 15 colonnes, qui servira de fondement à nos analyses futures.

4.3 Visualisation et Traitement des Valeurs Manquantes

Ici, nous abordons la visualisation des données pour identifier les tendances et les valeurs manquantes. Des techniques spécifiques sont mises en œuvre pour traiter ces valeurs manquantes, assurant ainsi l'intégrité et la fiabilité de nos analyses. Cette étape est cruciale pour comprendre pleinement la dynamique de la consommation énergétique et pour garantir que les modèles prédictifs sont basés sur des données complètes et précises.

Identification des Valeurs Manquantes

À travers une carte de chaleur, nous avons identifié que la colonne 'Consommation Brut (MW)' comportait 37 valeurs manquantes. Cette visualisation est cruciale car elle met en

évidence la répartition et l'importance des données manquantes dans notre analyse.

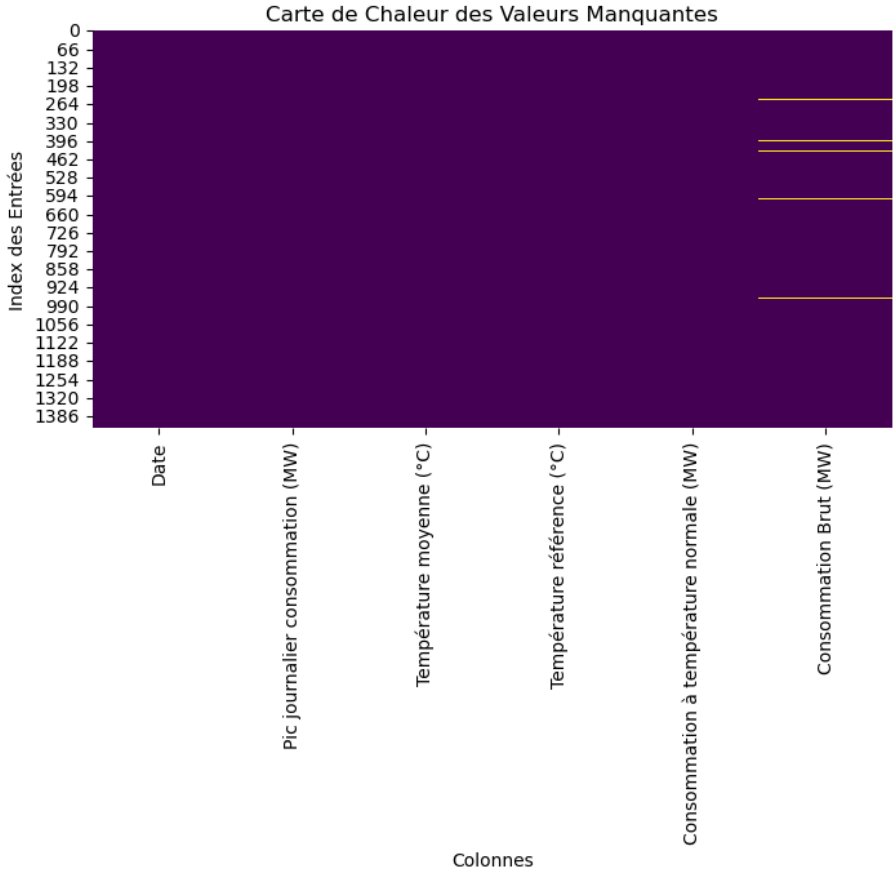

FIGURE 4.1 – Carte de Chaleur des Valeurs Manquantes

Analyse Statistique Descriptive

L'analyse descriptive de 'Consommation Brut (MW)' montre les tendances centrales et la dispersion des données, comme suit :

Statistique	**Valeur**
Nombre total d'observations	1393
Moyenne	51667.23 MW
Écart-type	10386.46 MW
Valeur minimale	35288 MW
1er quartile	43868 MW
Médiane	47554.5 MW
3ème quartile	60611.5 MW
Valeur maximale	77591 MW

TABLE 4.1 – Statistiques Descriptives de 'Consommation Brut (MW)'

Imputation des Valeurs Manquantes

Nous avons choisi d'imputer les valeurs manquantes en utilisant la médiane, qui est de 47554.5 MW. Cette méthode est préférée à la moyenne en raison de sa résilience aux

valeurs extrêmes, comme en témoigne l'écart-type significatif. L'imputation médiane maintient l'intégrité de la distribution des données tout en comblant les lacunes, permettant des analyses ultérieures plus robustes et fiables.

Chapitre 5

Ingénierie des Caractéristiques

5.1 Extraction des Caractéristiques Temporelles

L'extraction des composants temporels est une première étape déterminante. En utilisant la date comme index, nous avons extrait l'année, le mois, le jour, le jour de la semaine, et le numéro de la semaine. Ces informations sont fondamentales pour identifier les tendances et les cycles saisonniers influençant la consommation d'énergie.

5.2 Identification des Jours Fériés

L'ajout d'une variable 'IsHoliday' pour signaler les jours fériés est une innovation clé de notre modèle. Cela nous aide à intégrer l'impact des jours fériés sur la consommation d'énergie dans notre analyse prédictive.

5.3 Ajout de Caractéristiques Complémentaires

Nous avons enrichi le modèle avec des noms explicites pour les jours de la semaine et des indicateurs pour les week-ends, permettant une analyse plus nuancée des habitudes de consommation.

Ces étapes améliorent la robustesse du modèle en fournissant une vue holistique des facteurs influençant la consommation énergétique.

Chapitre 6

Analyse Exploratoire des Données (AED)

L'AED est notre première fenêtre sur le monde complexe de la consommation énergétique. Nous plongeons dans les tendances, les variations saisonnières et les anomalies. Comprendre ces éléments nous éclaire sur l'impact du climat et d'autres facteurs sur la demande en énergie. Nous analysons les données pour détecter des motifs répétitifs et établir des corrélations entre les variables. Cette étape est vitale pour anticiper les besoins futurs et optimiser la gestion de l'énergie.

6.1 Visualisation des Tendances de Consommation

Les graphiques suivants illustrent les tendances de la consommation énergétique brute, ajustée à la température normale, les pics journaliers, et la température moyenne.

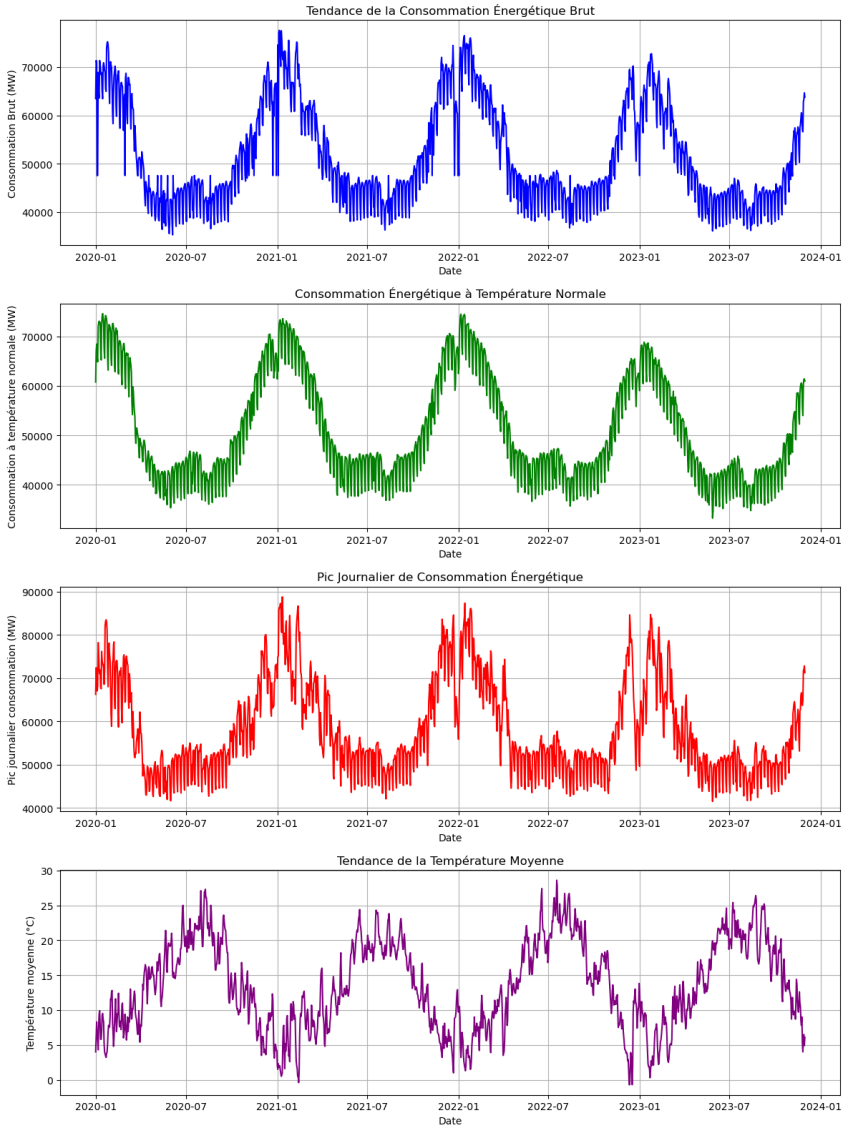

FIGURE 6.1 – Tendances de la consommation énergétique brute et ajustée.

Le graphique en bleu (Figure 6.1, haut) montre les fluctuations de la consommation brute, avec des pics qui peuvent refléter la demande accrue pendant les saisons froides ou chaudes. La courbe verte (Figure 6.1, milieu) nous indique comment la consommation s'adapte aux températures normales, démontrant l'influence directe du climat.

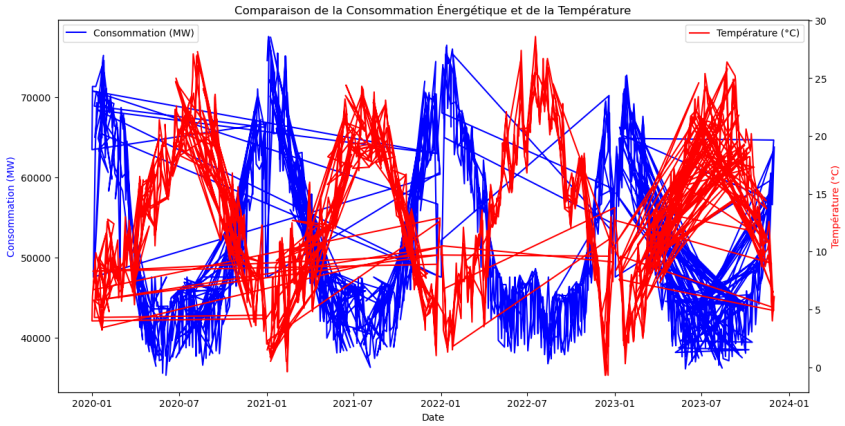

FIGURE 6.2 – Comparaison de la consommation énergétique et de la température.

La Figure 6.2 dépeint la consommation (en bleu) contre la température (en rouge), mettant en lumière leur interdépendance. Des pics de consommation correspondent souvent à des baisses de température.

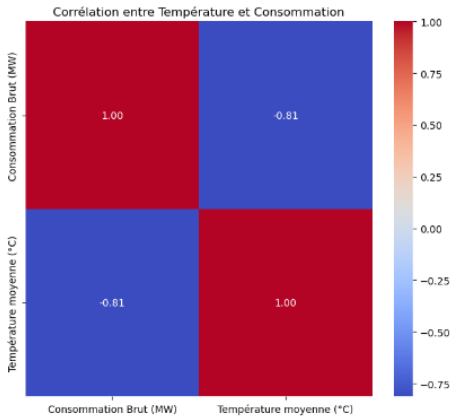

FIGURE 6.3 – Corrélation entre la température moyenne et la consommation énergétique.

La matrice de corrélation (Figure 6.3) confirme une forte liaison entre la température et la consommation, essentielle pour prévoir les besoins énergétiques et ajuster les stratégies de production.

Ces analyses graphiques enrichissent notre compréhension et nous guident vers un modèle prédictif affiné et pertinent.

6.2 Variation Saisonnière et Hebdomadaire de la Consommation

Le comportement de la consommation énergétique varie à la fois avec les saisons et les jours de la semaine, comme le démontrent les visualisations suivantes.

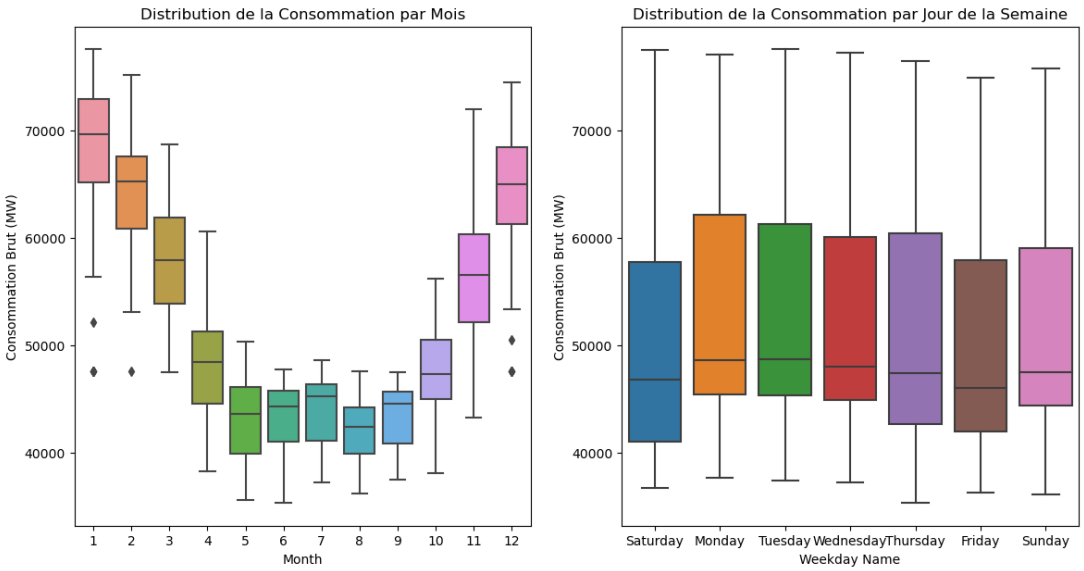

FIGURE 6.4 – Distribution de la Consommation Énergétique par Mois et par Jour de la Semaine

La Figure 6.4 (gauche) illustre une tendance marquée de consommation accrue pendant les mois d'hiver, ce qui est probablement dû à l'usage plus intensif de chauffage. Inversement, les mois d'été affichent une consommation réduite, en partie grâce à des jours plus longs et potentiellement moins de besoin en chauffage ou refroidissement.

Quant à la distribution hebdomadaire (Figure 6.4 (droite)), la consommation reste relativement constante, indiquant que la consommation quotidienne reste stable tout au long de la semaine, sans variations majeures entre les jours ouvrables et le week-end.

Ces observations sont essentielles pour anticiper les périodes de forte demande et pour une gestion équilibrée de l'approvisionnement en énergie.

Chapitre 7

Modélisation

Dans ce chapitre, nous décrirons en détail les étapes méthodologiques et les analyses effectuées pour développer des modèles de prévision de la consommation énergétique. Voici les étapes clés qui seront explorées :

Évaluation de la Stationnarité : Nous débuterons par examiner la stationnarité des séries temporelles, en utilisant des tests statistiques pour déterminer si une différenciation ou une transformation est nécessaire.

Analyse des Fonctions d'Auto-Corrélation : Les fonctions d'Auto-Corrélation (ACF) et de Corrélation Partielle (PACF) seront analysées pour identifier les ordres appropriés des composantes Auto-Régressives (AR) et Moyennes Mobiles (MA) pour les modèles ARIMA.

Sélection des Modèles : Différents modèles prédictifs seront testés, des modèles linéaires simples aux modèles plus complexes comme les réseaux de neurones LSTM, pour évaluer leur performance.

Validation des Modèles : Les modèles sélectionnés seront validés à l'aide de techniques telles que la validation croisée pour assurer leur robustesse et leur fiabilité.

Chaque étape sera accompagnée d'analyses et de visualisations pour illustrer notre démarche et les découvertes qui en découlent.

7.1 Analyse Visuelle de la Stationnarité

Nous commençons par examiner visuellement la série temporelle de la consommation énergétique brute.

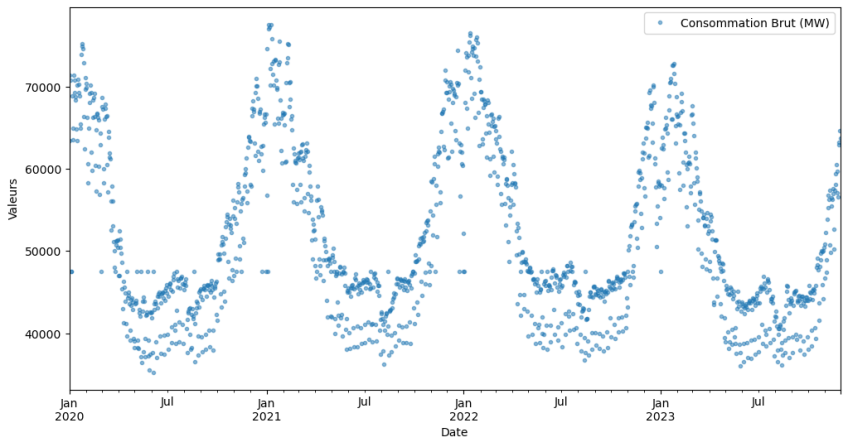

FIGURE 7.1 – Dispersion de la consommation énergétique brute à travers le temps.

Comme le montre la Figure 7.1, la consommation affiche une saisonnalité claire, avec des pics en hiver et des creux en été, indiquant une possible non-stationnarité saisonnière.

7.2 Test de Stationnarité

Après l'observation visuelle, nous validons ces conclusions à l'aide du test de Dickey-Fuller augmenté.

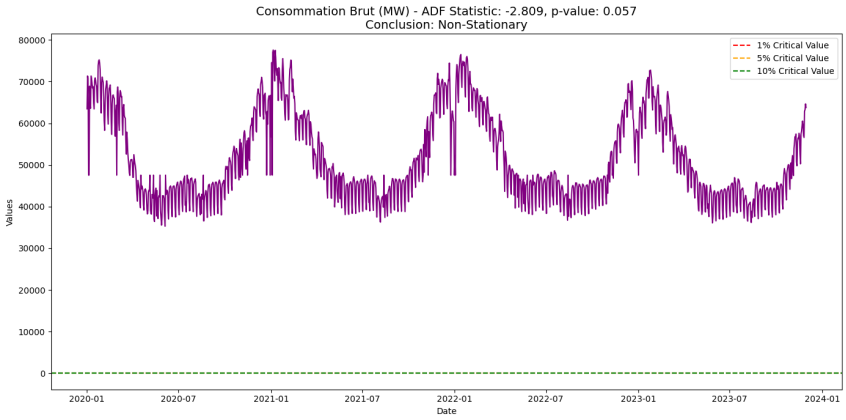

FIGURE 7.2 – Résultats du test de Dickey-Fuller augmenté pour la consommation énergétique brute.

La Figure 7.2 illustre que la série n'est pas stationnaire, avec une valeur-p de 0.057, ce qui nous oblige à envisager des différenciations saisonnières ou d'autres transformations pour stabiliser la série.

7.3 Différenciation pour atteindre la Stationnarité

Après avoir identifié une non-stationnarité initiale, nous avons appliqué une différenciation à la série temporelle pour stabiliser la moyenne et réduire la tendance saisonnière.

Cette technique de différenciation est couramment utilisée pour transformer une série non-stationnaire en une série stationnaire, condition préalable pour de nombreux modèles prédictifs.

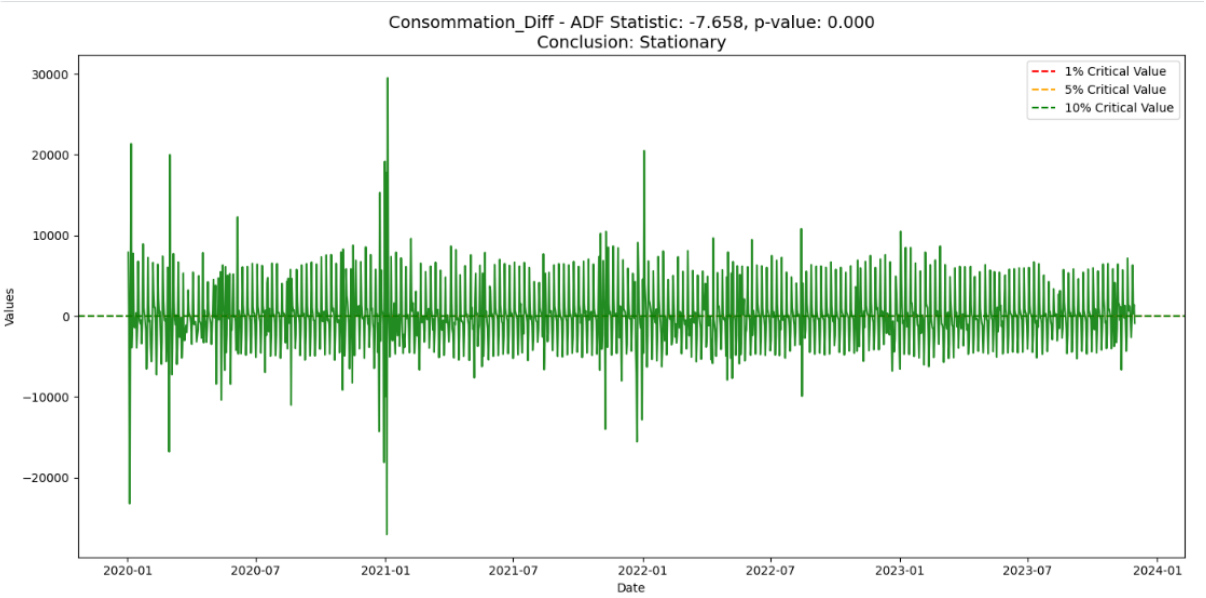

FIGURE 7.3 – Série temporelle après différenciation.

La Figure 7.3 montre la série après application de la différenciation. Visuellement, la série semble maintenant fluctuer autour d'une moyenne constante, sans tendance ou saisonnalité apparente. De plus, le test de Dickey-Fuller augmenté indique une statistique ADF de -7.658 et une valeur-p de 0.000, confirmant la stationnarité de la série.

Ces résultats valident l'efficacité de la différenciation effectuée et permettent de poursuivre avec la modélisation de la série temporelle désormais stationnaire.

7.4 Modèle ARIMA

Dans cette section, nous développerons un modèle ARIMA (AutoRegressive Integrated Moving Average), une approche classique pour la modélisation de séries temporelles stationnaires. Le modèle ARIMA est choisi pour sa capacité à capturer à la fois les dépendances dans les données ainsi que les aspects de tendance et de saisonnalité après application de différenciation.

7.4.1 Sélection de l'ordre du modèle ARIMA

Pour sélectionner l'ordre approprié du modèle ARIMA, nous nous basons sur l'analyse des fonctions d'auto-corrélation (ACF) et de corrélation partielle (PACF).

article graphicx float

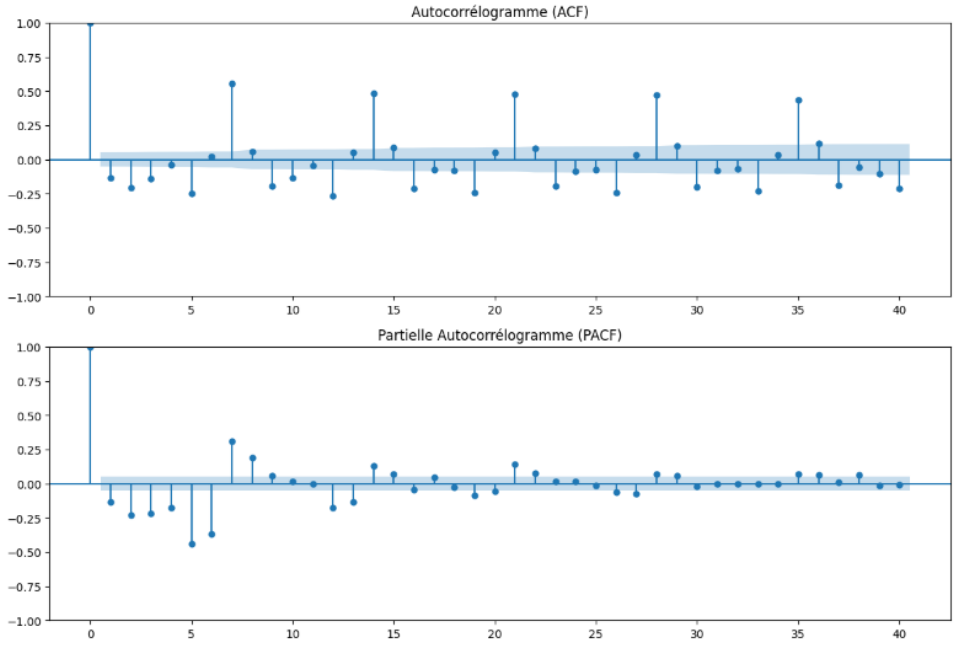

FIGURE 7.4 – Fonctions d'Auto-Corrélation (ACF) et de Corrélation Partielle (PACF)

La Figure 7.4 illustre les fonctions ACF et PACF pour notre série temporelle après différenciation. L'ACF révèle des pics significatifs qui s'étendent au-delà de l'intervalle de confiance, indiquant des autocorrélations significatives à plusieurs retards. Cela suggère que la série temporelle ne peut être modélisée par un modèle de moyenne mobile (MA) simple. De manière similaire, la PACF présente également des pics significatifs à divers intervalles, ce qui écarte un modèle autorégressif (AR) simple.

L'observation des pics s'étendant à des retards supérieurs dans les deux fonctions suggère que la structure de la série temporelle est plus complexe et qu'elle pourrait bénéficier d'un modèle combinant à la fois les composantes AR et MA. Par conséquent, une sélection automatique des paramètres à l'aide de la méthode auto_arima est justifiée pour identifier un modèle ARIMA approprié qui capture à la fois les autorégressions et les moyennes mobiles inhérentes à la série.

7.4.2 Sélection du Modèle Prédictif : ARIMA ET SARIMAX

Dans cette phase essentielle, nous cherchons à identifier le modèle ARIMA ou SARIMA le mieux adapté pour nos données sur la consommation énergétique. Nous employons la fonction `auto_arima` de `pmdarima`, qui teste automatiquement diverses combinaisons de paramètres pour minimiser l'AIC ou le BIC.

- `start_p` et `start_q` fixent les points de départ pour les ordres AR et MA.
- `max_p` et `max_q` délimitent les ordres maximaux pour les composantes AR et MA.
- `m` dénote la périodicité saisonnière des données.
- `start_P`, `D`, et les paramètres saisonniers permettent d'explorer la saisonnalité.
- `trace` active l'affichage du processus de recherche.
- `stepwise` implémente une recherche optimisée pour accélérer le processus.

Le modèle retenu est ensuite résumé pour évaluer son adéquation avec nos besoins prévisionnels.

Entraînement et Validation des Modèles ARIMA, SARIMA et SARIMAX

Pour la prévision de la consommation énergétique, nous abordons un processus méthodique d'entraînement et de validation des modèles ARIMA, SARIMA et SARIMAX. Le choix de ces modèles est dicté par leur capacité à saisir les tendances et les saisonnalités inhérentes aux données de consommation d'énergie. Le processus est le suivant :

1. **Séparation des données** : Nous commençons par diviser notre série temporelle en deux ensembles distincts. Les données allant jusqu'au 31 décembre 2022 constituent l'ensemble d'entraînement, tandis que les données à partir du 1er janvier 2023 servent d'ensemble de test.
2. **Modélisation ARIMA** : Nous ajustons d'abord un modèle ARIMA aux données d'entraînement. Le modèle est sélectionné sur la base de critères tels que l'AIC ou le BIC, et l'ordre des composantes ARIMA est déterminé par les fonctions d'auto-corrélation et de corrélation partielle.
3. **Modélisation SARIMA** : Pour capturer la saisonnalité, nous étendons l'approche ARIMA en intégrant des termes saisonniers, formant ainsi un modèle SARIMA. Les périodes saisonnières sont identifiées grâce à une analyse exploratoire approfondie des données.
4. **Modélisation SARIMAX** : Nous adoptons également un modèle SARIMAX qui incorpore des variables exogènes potentiellement influentes, telles que les températures moyennes ou les jours fériés, pour améliorer la précision des prévisions.
5. **Validation et Comparaison** : Chaque modèle est évalué sur l'ensemble de test pour vérifier sa capacité de généralisation. Les métriques telles que le RMSE et le R^2 nous permettent de comparer les performances des modèles et de sélectionner le plus adéquat.
6. **Amélioration continue** : Nous nous engageons dans un cycle itératif d'évaluation et d'amélioration, affinant les modèles en fonction des retours de validation pour atteindre l'optimum de précision.

L'objectif est de développer un modèle robuste et précis qui puisse être utilisé pour des prévisions énergétiques fiables et éclairées.

Meilleur modèle ARIMA

Le modèle ARIMA(5,0,5) a été identifié comme le plus adapté pour notre série temporelle 'Consommation_Diff'. Ce modèle illustre la complexité des dynamiques de consommation énergétique, avec une interdépendance notable entre les valeurs passées et la capacité à saisir les cycles et tendances. Cependant, un RMSE de 2659.52 et un R^2 de 0.2791 indiquent que malgré certaines tendances capturées, des écarts notables persistent entre les prévisions et les valeurs réelles.

La nécessité d'explorer d'autres ordres et modèles comme SARIMA et SARIMAX est mise en évidence, afin de raffiner l'adéquation aux données observées. La matrice de covariance, proche de la singularité, peut indiquer des erreurs standard potentiellement instables, suggérant une possible simplification du modèle.

SARIMAX Results

Dep. Variable:	Consommation_Diff	No. Observations:	1088
Model:	SARIMAX(5, 0, 5)	Log Likelihood	-10132.348
Date:	Thu, 11 Jan 2024	AIC	20286.696
Time:	22:25:17	BIC	20341.548
Sample:	01-09-2020	HQIC	20307.463
	- 12-31-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4238	0.032	13.067	0.000	0.360	0.487
ar.L2	-1.1062	0.029	-38.122	0.000	-1.163	-1.049
ar.L3	0.2350	0.047	4.993	0.000	0.143	0.327
ar.L4	-0.6703	0.029	-22.942	0.000	-0.728	-0.613
ar.L5	-0.3694	0.033	-11.343	0.000	-0.433	-0.306
ma.L1	-0.9731	0.043	-22.703	0.000	-1.057	-0.889
ma.L2	1.3790	0.052	26.759	0.000	1.278	1.480
ma.L3	-0.9488	0.069	-13.764	0.000	-1.084	-0.814
ma.L4	0.9596	0.048	19.952	0.000	0.865	1.054
ma.L5	-0.2559	0.042	-6.105	0.000	-0.338	-0.174
sigma2	1.003e+07	3.95e-09	2.54e+15	0.000	1e+07	1e+07

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	3587.28
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	0.45	Skew:	-0.33
Prob(H) (two-sided):	0.00	Kurtosis:	11.90

FIGURE 7.5 – Estimation des paramètres du modèle ARIMA(5,0,5)

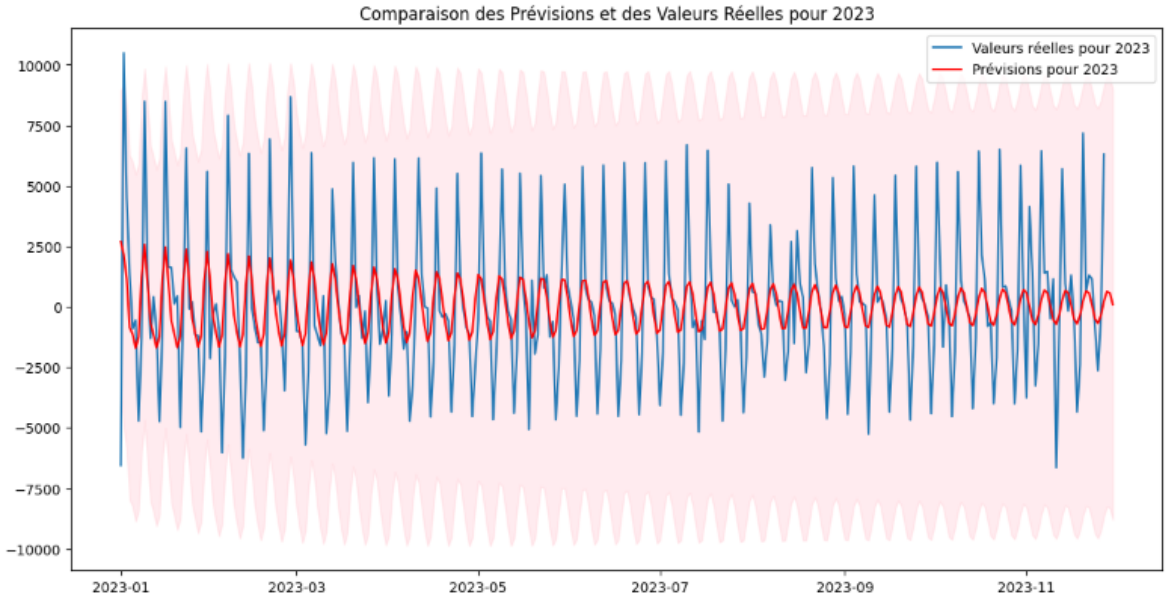

FIGURE 7.6 – Prédictions du modèle ARIMA(5,0,5) comparées aux valeurs réelles pour 2023

L'équation du modèle ARIMA(5,0,5) peut être exprimée comme suit :

$$y_t = c + \phi_1 y_{t-1} + ... + \phi_5 y_{t-5} + \theta_1 \varepsilon_{t-1} + ... + \theta_5 \varepsilon_{t-5} + \varepsilon_t \tag{7.1}$$

où y_t est la consommation énergétique à l'instant t, c est une constante, ϕ_i sont les paramètres autorégressifs, θ_i sont les paramètres des moyennes mobiles, et ε_t est le terme d'erreur.

Présentation et Interprétation du Meilleur Modèle SARIMAX

Le modèle SARIMAX sélectionné avec un ordre de (6,0,5) et un ordre saisonnier de (2,1,[1],12) a montré une précision améliorée en incluant la température moyenne et les jours fériés comme variables exogènes. Le RMSE de 2533.82 et le R^2 de 0.3456 indiquent une bonne adéquation du modèle par rapport aux données observées, bien que des améliorations soient possibles.

SARIMAX Results

Dep. Variable:	Consommation_Diff	No. Observations:	1088
Model:	SARIMAX(6, 0, 5)x(2, 1, [1], 12)	Log Likelihood	-9875.211
Date:	Thu, 11 Jan 2024	AIC	19784.421
Time:	23:14:19	BIC	19868.618
Sample:	01-09-2020 - 12-31-2022	HQIC	19816.352
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
Température moyenne (°C)	-35.5111	13.840	-2.566	0.010	-62.638	-8.384
IsHoliday	-2731.6427	599.627	-4.556	0.000	-3906.890	-1556.395
ar.L1	-0.7063	0.056	-12.512	0.000	-0.817	-0.596
ar.L2	-0.7693	0.028	-27.951	0.000	-0.823	-0.715
ar.L3	-0.9157	0.055	-16.771	0.000	-1.023	-0.809
ar.L4	-0.6069	0.030	-20.270	0.000	-0.666	-0.548
ar.L5	-1.0257	0.037	-28.101	0.000	-1.097	-0.954
ar.L6	-0.5595	0.031	-18.113	0.000	-0.620	-0.499
ma.L1	0.2364	0.071	3.349	0.001	0.098	0.375
ma.L2	0.5345	0.057	9.394	0.000	0.423	0.646
ma.L3	0.5474	0.081	6.769	0.000	0.389	0.706
ma.L4	0.2783	0.057	4.900	0.000	0.167	0.390
ma.L5	0.7268	0.055	13.292	0.000	0.620	0.834
ar.S.L12	0.0277	0.056	0.493	0.622	-0.082	0.138
ar.S.L24	0.0105	0.070	0.151	0.880	-0.126	0.147
ma.S.L12	-0.9582	0.030	-31.467	0.000	-1.018	-0.898
sigma2	1.397e+07	0.102	1.37e+08	0.000	1.4e+07	1.4e+07

Ljung-Box (L1) (Q):	0.17	Jarque-Bera (JB):	3263.23
Prob(Q):	0.68	Prob(JB):	0.00
Heteroskedasticity (H):	0.35	Skew:	-0.23
Prob(H) (two-sided):	0.00	Kurtosis:	11.64

FIGURE 7.7 – Estimation des paramètres du modèle SARIMAX

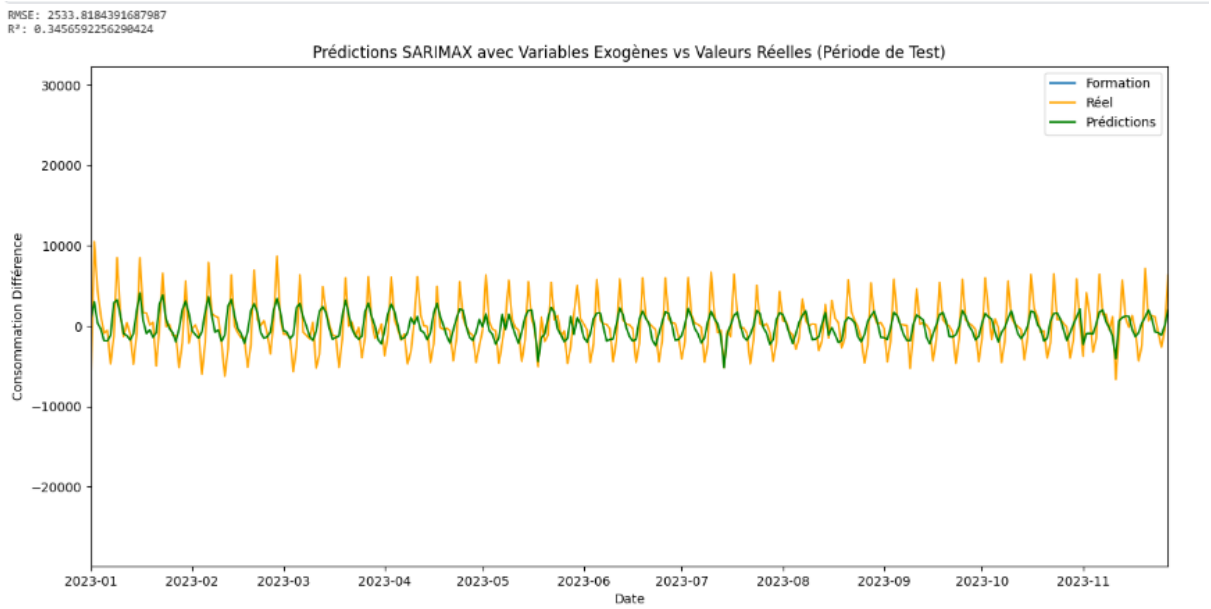

FIGURE 7.8 – Comparaison des prévisions du modèle SARIMAX avec les valeurs réelles pour 2023

L'équation du modèle SARIMAX peut être formulée comme suit :

$$Consommation\_Diff_t = c + \sum_{i=1}^{6} \phi_i Consommation\_Diff_{t-i} + \sum_{i=1}^{5} \theta_i \varepsilon_{t-i} + \varepsilon_t + saisonnalit + variablesexog \tag{7.2}$$

où $Consommation\_Diff_t$ est la différence de consommation à l'instant t, c est une constante, ϕ_i sont les coefficients AR, θ_i sont les coefficients MA, ε_t est le terme d'erreur et les composants de saisonnalité et variables exogènes sont inclus pour capter les effets saisonniers et l'impact des jours fériés et de la température.

Le modèle SARIMAX retenu avec les paramètres (6, 0, 5) pour la partie ARIMA et (2, 1, 1, 12) pour la composante saisonnière indique une capacité à capturer les dynamiques complexes et saisonnières de la consommation énergétique. Bien que le RMSE de 2533.81 et le R^2 de 0.3456 montrent une amélioration par rapport aux modèles précédents, nous sommes à la recherche d'une performance encore meilleure.L'ajustement des composantes AR et MA démontre une corrélation temporelle significative, tandis que les coefficients des variables exogènes soulignent leur influence directe. Le fait que la matrice de covariance soit presque singulière suggère que le modèle pourrait être amélioré pour renforcer la stabilité des estimations des paramètres.

En poursuivant notre analyse, nous allons maintenant explorer des modèles d'apprentissage automatique comme XGBoost et LSTM pour leur potentiel à modéliser des relations non linéaires et à améliorer la précision des prédictions..

7.5 Modèles d'Apprentissage Automatique

Dans cette section du rapport, nous aborderons les différentes étapes de l'utilisation des modèles d'apprentissage automatique pour la prévision énergétique. Nous créerons d'abord des variables rétardées pour intégrer l'histoire dans nos prévisions. Ensuite, nous séparerons nos données en ensembles d'entraînement et de test pour valider la performance des modèles. Nous entraînerons et évaluerons un modèle XGBoost, et nous utiliserons également un modèle LSTM pour sa capacité à traiter les séries temporelles. La métrique RMSE nous aidera à mesurer la précision des prévisions du modèle.

7.5.1 Amélioration de l'Ingénierie des Caractéristiques

L'ingénierie des caractéristiques est un aspect crucial dans la modélisation énergétique. Elle implique la création de variables rétardées (lags) pour comprendre l'influence des valeurs passées sur la consommation future. De plus, l'intégration de fenêtres glissantes aide à lisser les données et à révéler des tendances sur une période plus longue. La décomposition de la série temporelle en tendance, saisonnalité et composantes résiduelles offre une compréhension plus profonde des dynamiques de consommation. Enfin, en prenant en compte les résidus, nous pouvons détecter des anomalies et des changements soudains.

Toutefois, cette approche peut augmenter la complexité du modèle et présenter des risques de surinterprétation ou de multicollinéarité. Il est donc essentiel de procéder avec prudence et de valider les modèles de manière rigoureuse.

7.5.2 Séparation des Données et Préparation pour la Modélisation

Nous avons divisé nos données en deux ensembles : entraînement et test. La séparation est basée sur la date du 31 décembre 2022. Les ensembles ont les tailles suivantes :

— Entraînement : 1088 observations pour les caractéristiques (24 variables) et 1088 pour la variable cible.

— Test : 331 observations pour les caractéristiques (24 variables) et 331 pour la variable cible.

Cette division nous permet d'entraîner le modèle sur un large éventail de données historiques, tout en réservant un ensemble de test significatif pour évaluer sa performance.

7.5.3 Entraînement et Évaluation du Modèle XGBoost

L'entraînement de notre modèle XGBoost a commencé par une approche simple, sans paramètres personnalisés, afin d'établir une base de référence. Pour affiner notre modèle, nous avons ensuite effectué une recherche croisée (cross-validation) pour déterminer les paramètres optimaux. Cette méthode systématique nous a permis d'améliorer les performances du modèle.

Résultats des Modèles XGBoost

Deux modèles ont été testés : un modèle de base et un modèle avec des paramètres optimisés via la cross-validation. Le modèle optimisé a démontré une amélioration significative par rapport au modèle de base, comme en témoignent les métriques de performance.

Modèle Basique Le premier modèle, sans paramètres personnalisés, a produit un RMSE de 457.046, indiquant une certaine distance entre les prévisions et les valeurs réelles.

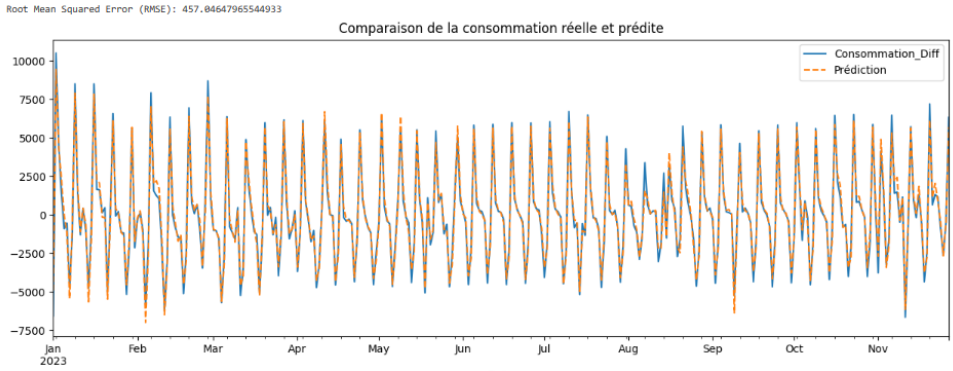

FIGURE 7.9 – Comparaison des prévisions du modèle basique XGBoost et des valeurs réelles

Modèle Optimisé Le second modèle, après optimisation, a affiché un RMSE réduit de 292.292, soulignant une précision accrue dans les prévisions.

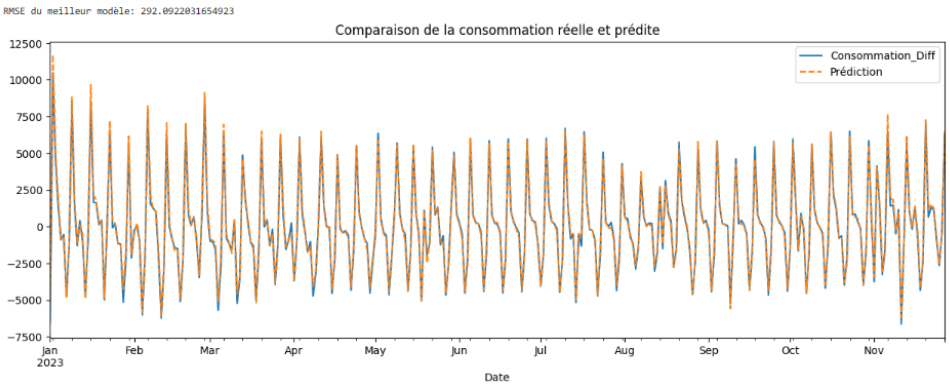

FIGURE 7.10 – Comparaison des prévisions du modèle optimisé XGBoost et des valeurs réelles

Les caractéristiques les plus influentes pour le modèle optimisé sont illustrées dans le graphique suivant, qui montre l'importance relative de chaque attribut.

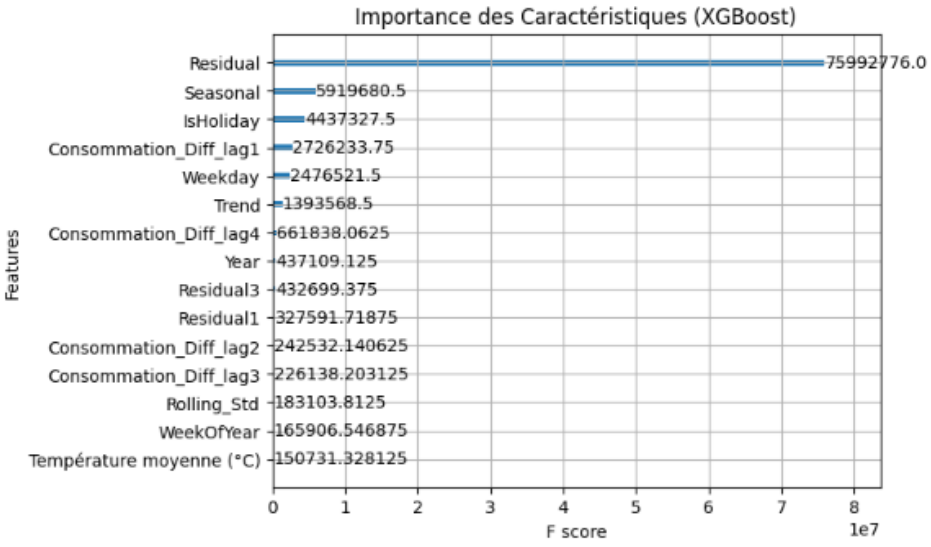

FIGURE 7.11 – Importance des caractéristiques dans le modèle optimisé XGBoost

Interprétation Le modèle optimisé, avec son RMSE plus faible, indique une amélioration de l'adéquation du modèle aux données. Cela implique que les prévisions sont plus proches des valeurs réelles, renforçant la confiance dans les prédictions générées par le modèle.

Conclusion Les résultats montrent que l'application de la cross-validation pour l'optimisation des hyperparamètres de XGBoost est bénéfique. Cela permet non seulement d'améliorer la précision des prédictions mais aussi de comprendre l'importance de chaque variable dans le modèle, ce qui peut guider des ajustements supplémentaires et l'interprétation des résultats de la modélisation.

7.5.4 Entraînement et Évaluation du Modèle LSTM

L'approche adoptée pour l'entraînement du modèle LSTM a consisté à expérimenter avec différentes architectures de réseau. Après divers essais, un modèle avec une configuration de couches spécifique a montré des résultats prometteurs.

Résultats du Modèle LSTM

Le modèle LSTM a généré une erreur quadratique moyenne (MSE) de 290, indiquant une adéquation raisonnable avec les données de consommation énergétique réelles. Ce résultat suggère que le modèle a appris avec succès les motifs sous-jacents de la série temporelle.

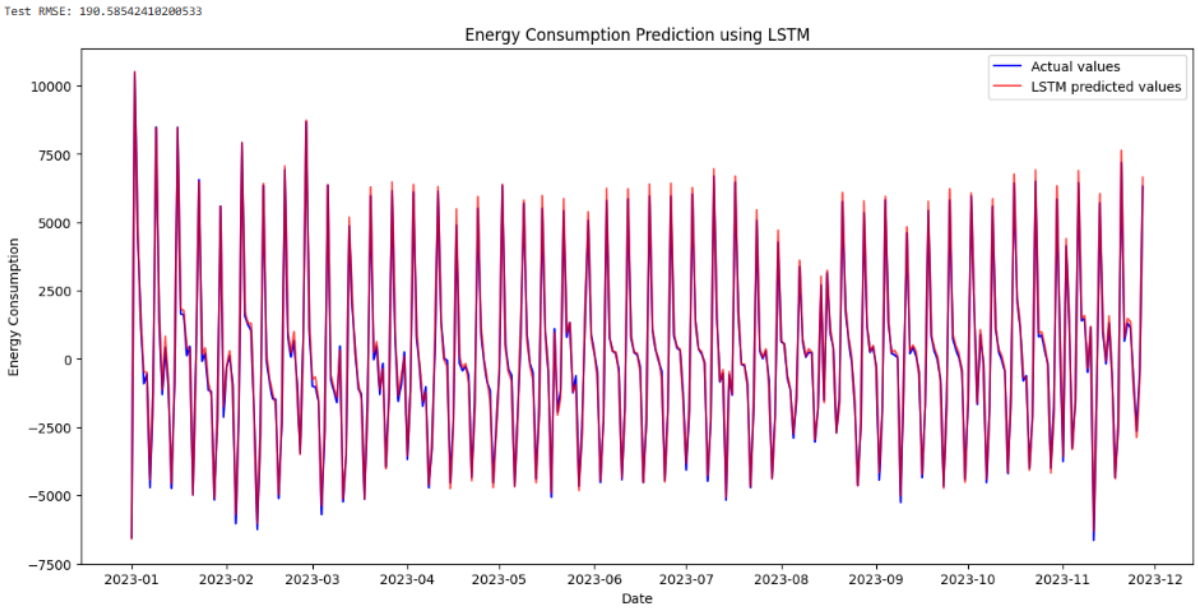

FIGURE 7.12 – Prédictions de consommation d'énergie à l'aide du modèle LSTM

Interprétation et Discussion Le modèle sélectionné offre un compromis entre précision des prédictions et complexité du modèle. La valeur de MSE obtenue suggère que le modèle peut être utilisé pour des prévisions à court terme. Cependant, une validation supplémentaire pourrait être nécessaire pour confirmer sa performance dans différents scénarios.

Conclusion Le modèle LSTM choisi, avec un MSE de 290, sera utilisé pour les prévisions futures. L'analyse des prédictions par rapport aux valeurs réelles met en évidence l'efficacité du modèle dans la capture des tendances de consommation d'énergie.

Chapitre 8

Conclusion

À travers ce projet, nous avons exploré et comparé plusieurs modèles pour prédire la consommation énergétique, avec un accent particulier sur les méthodes de séries temporelles et d'apprentissage automatique. Le modèle XGBoost optimisé a démontré une performance significative avec un RMSE de 292.292, indiquant une forte adéquation aux données observées. Parallèlement, le modèle LSTM a produit un MSE de 290, révélant son potentiel pour des prévisions précises à court terme. Ces modèles, caractérisés par une meilleure compréhension des facteurs influençant la consommation, s'avèrent prometteurs pour des prédictions fiables et précises.

En ce qui concerne la surveillance continue, l'intégration d'un système de surveillance en temps réel est suggérée pour permettre des mises à jour dynamiques des prévisions énergétiques. Cela contribuera à une gestion plus efficace de la demande énergétique, s'adaptant rapidement aux changements des habitudes de consommation et aux conditions extérieures.

Finalement, ce projet souligne l'importance de l'évaluation continue et de l'amélioration des modèles prédictifs dans le domaine de la consommation énergétique. Il ouvre également la voie à l'intégration de techniques avancées d'apprentissage automatique pour améliorer davantage la précision et la robustesse des prédictions.

8.1 Réponses aux Questions Spécifiques

8.1.1 Surveillance Continue de la Consommation Énergétique

Pour la question 10 sur la surveillance continue, l'implémentation d'un système de monitoring en temps réel est recommandée. Ce système devrait être capable de collecter des données de consommation énergétique, les analyser avec les modèles prédictifs mis à jour, et ajuster les prévisions en fonction des nouvelles informations. Une telle approche garantirait que la gestion énergétique reste réactive et efficace, s'adaptant aux tendances émergentes et aux anomalies dans les modèles de consommation.