

Consommation électrique en france : analyse et prédiction

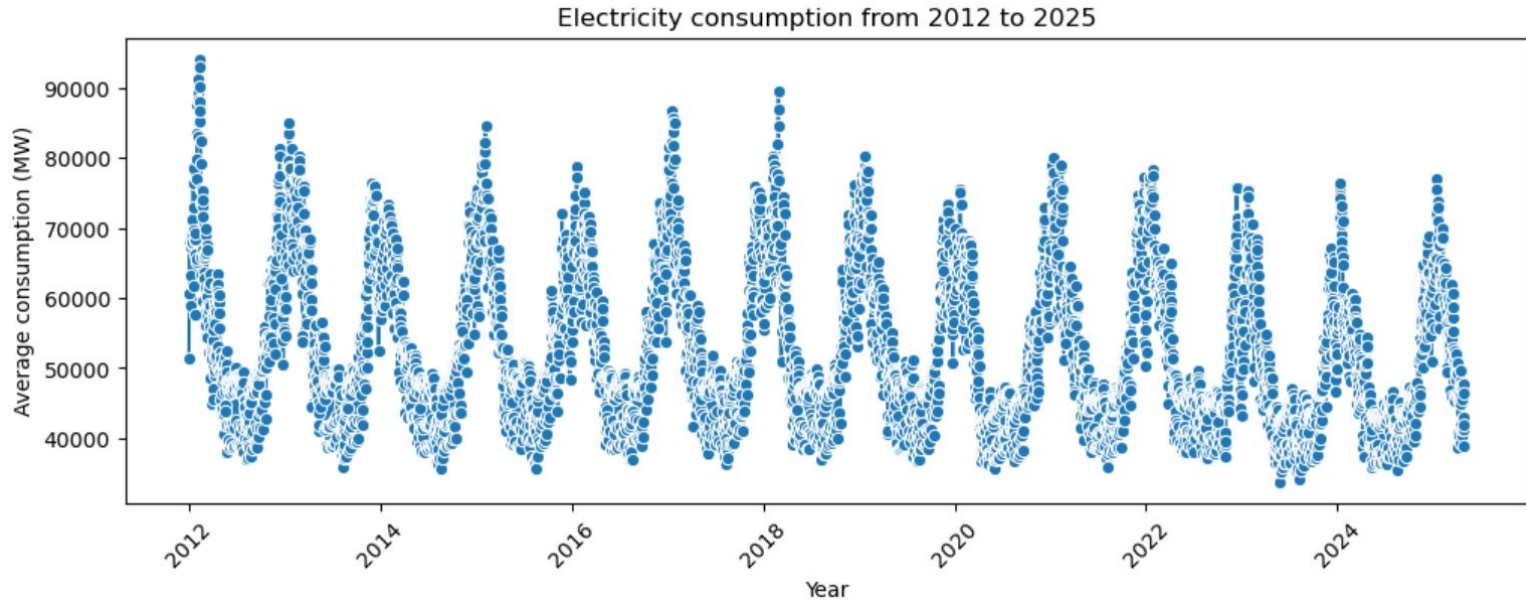
Plan

1. Contexte et objectif
2. Analyse de données
3. Prétraitement et choix du modèle
4. Résultats

1. Contexte et objectif

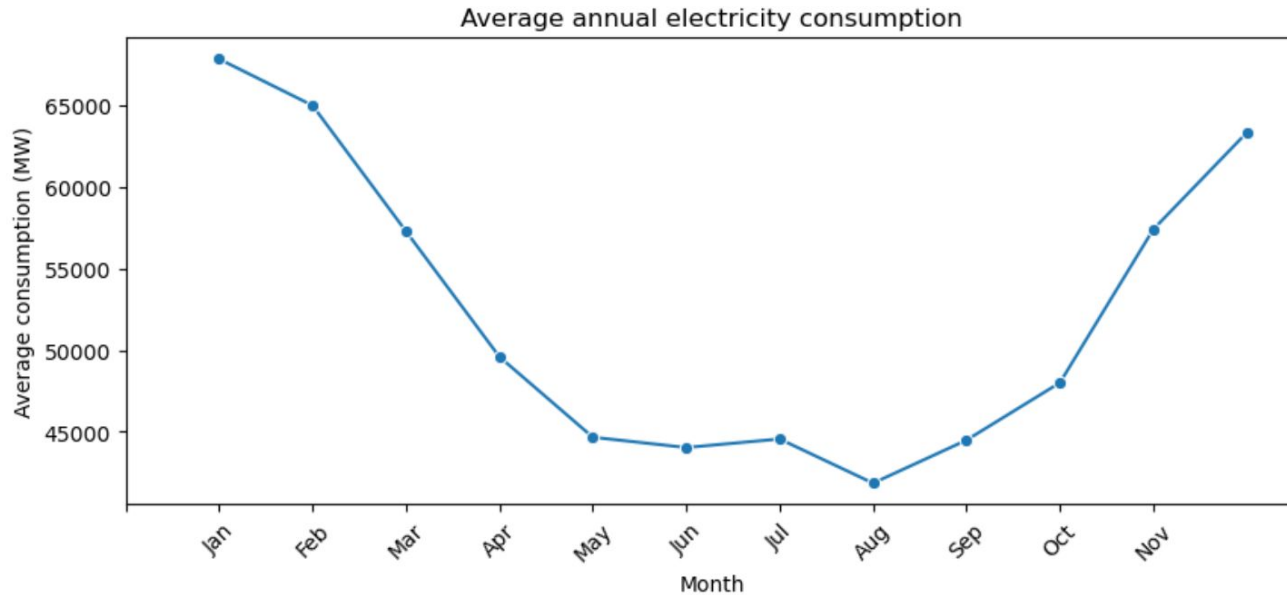
- **Objectif** : Au jour N à midi, prévoir la consommation totale du jour N+1 en France.
- **Données** :
 - Consommation RTE par tranche de 30 minutes depuis
 - Températures quotidiennes nationales
 - Calendrier TEMPO, jours fériés, vacances scolaires

2. Analyse de données



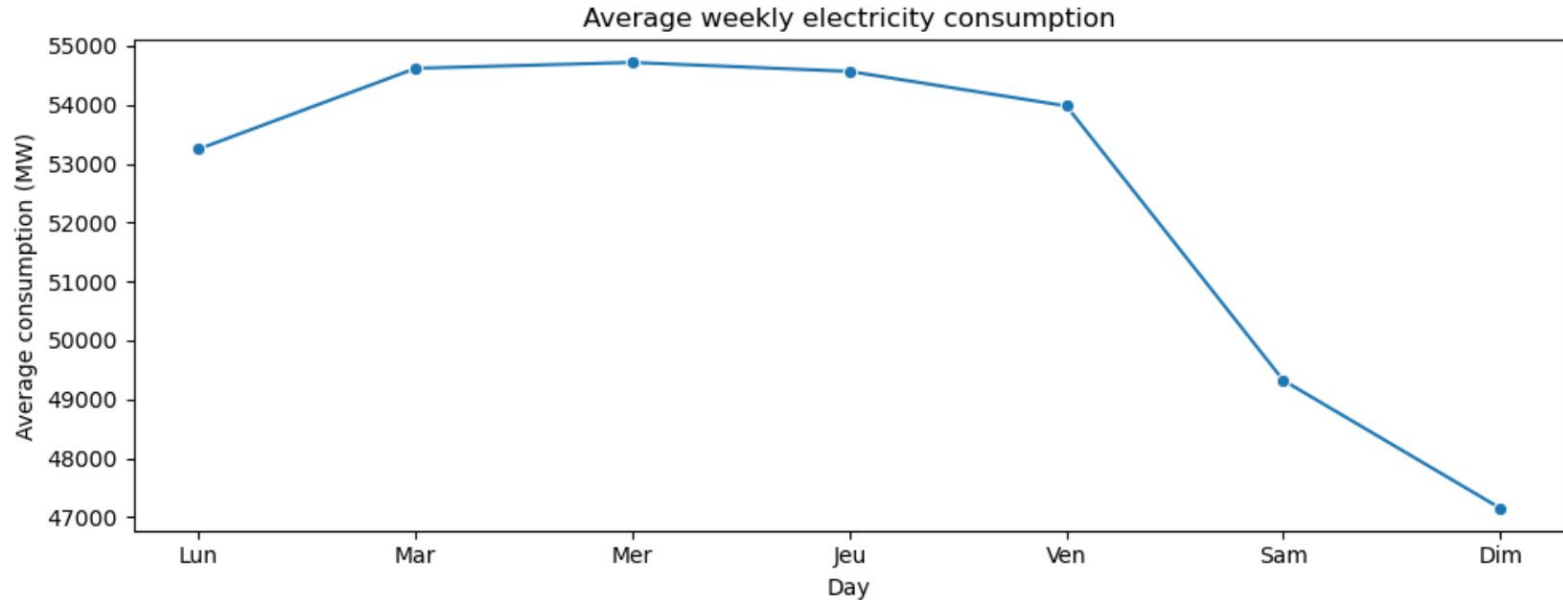
- Effet important de la saisonnalité : la consommation augmente en hiver et chute en été.
- Le profil de consommation se répète année après année sans changement notable.

2. Analyse de données



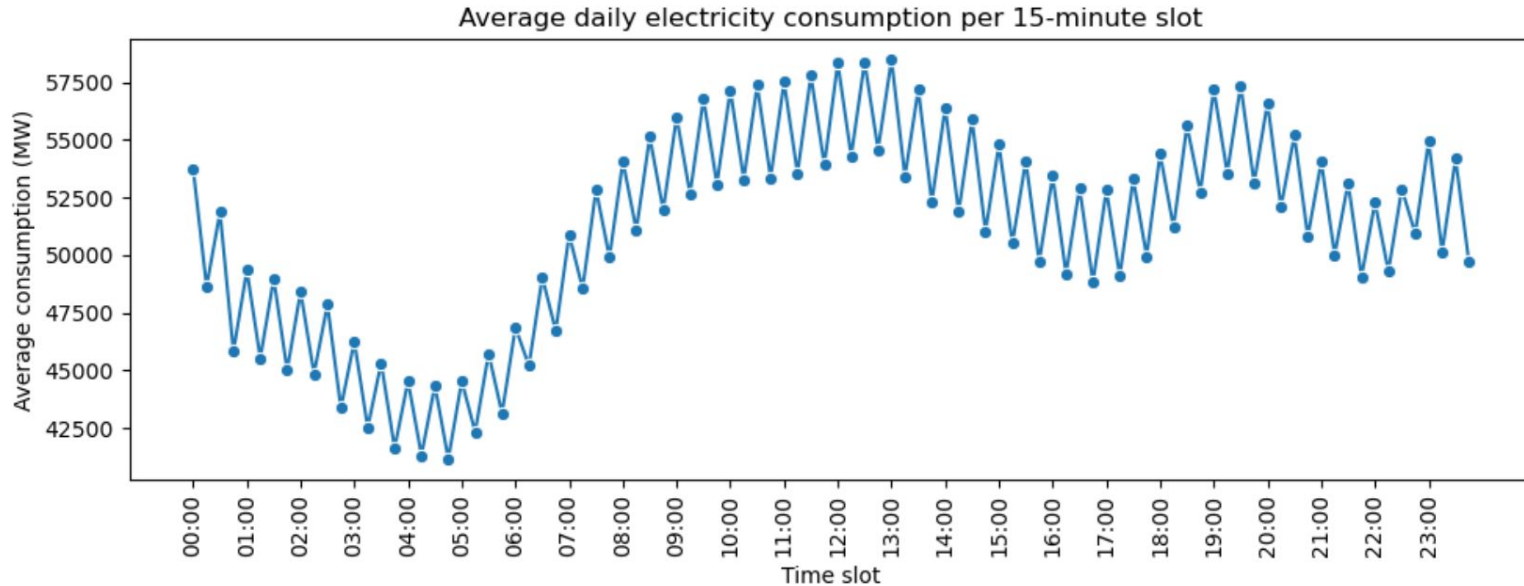
Effet important de la saisonnalité : la consommation augmente en hivers et chute en été.

2. Analyse de données



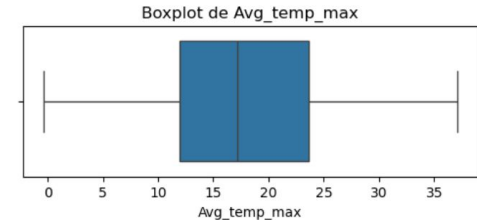
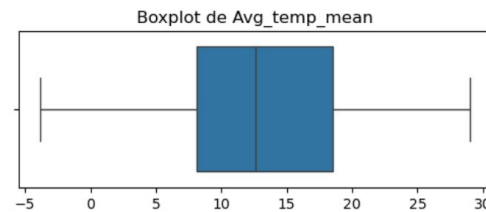
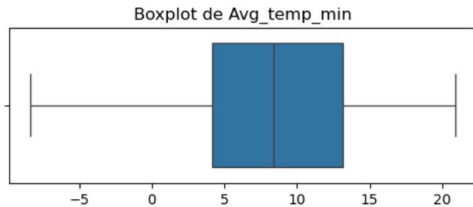
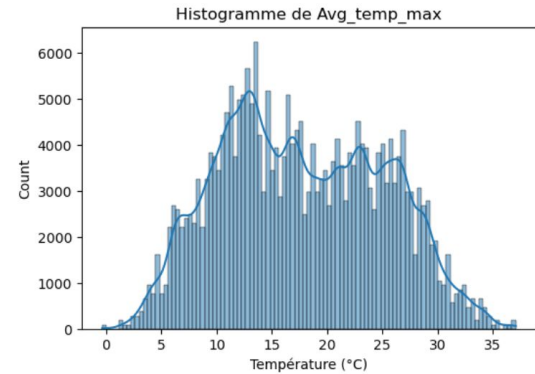
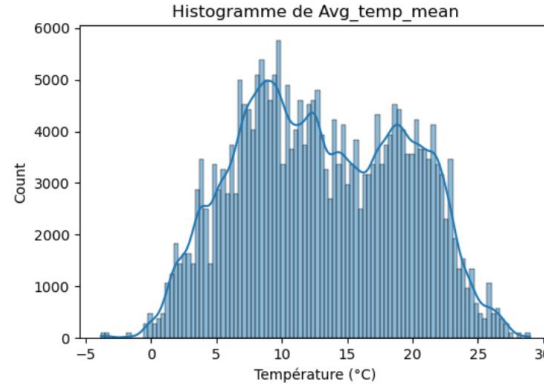
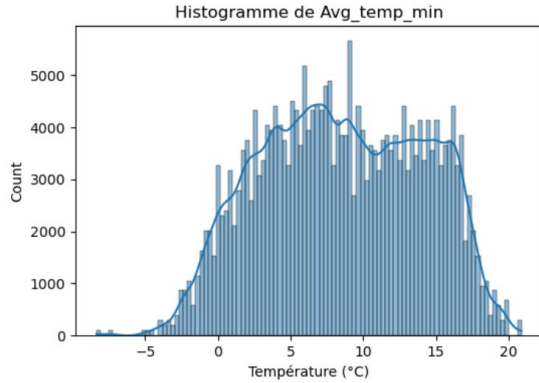
Effet important du jour de la semaine : la consommation augmente la semaine, lorsque l'activité humaine bat son plein, et retombe le week-end.

2. Analyse de données



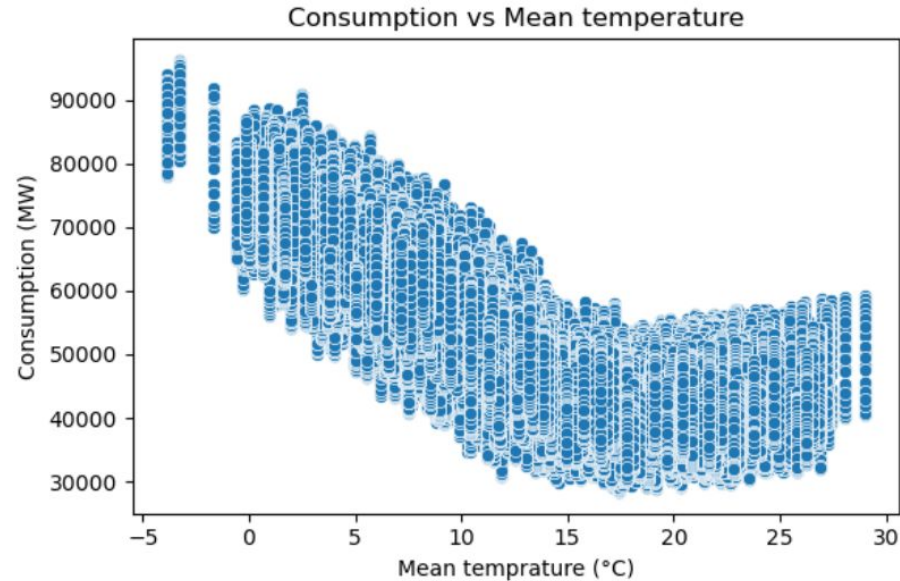
- Effet important de l'heure : la consommation augmente en journée lorsque l'activité humaine bat son plein, et retombe la nuit.
- Le "cisaillement" est dû au fait qu'à ce stade l'analyse, j'avais gardé les lignes HH:15 et HH:45.

2. Analyse de données



- On observe plusieurs bosses pour chaque histo, signe que la température suit une distribution multimodale qui reflète les saisons.
- On a un pic pour les températures basses (hiver), un autre pour les températures élevées (été), et parfois un autre au printemps/automne ?

2. Analyse de données



- Forte corrélation négative avec la température ($-0,76$).
- Quand la température moyenne nationale diminue, la consommation électrique augmente de façon assez linéaire, car en hiver, les besoins de chauffage explosent. Inversement pour l'été.

2. Analyse de données

Impact de la période de la semaine

- Moyenne en semaine $\approx 54\,228$ MW
- Moyenne le week-end $\approx 48\,237$ MW

-> En moyenne, la consommation le week-end est ~ 11 % plus basse qu'en semaine (t-test p-value < 0.05 , Cohen's d = 0.527).

-> Pourquoi ? Moins d'activité industrielle et tertiaire le week-end, plus de foyers inoccupés (?) ou plus d'activités domestiques (qui consomme moins qu'une usine).

Impact des vacances scolaires

- Jours hors vacances $\approx 53\,937$ MW
- Jours de vacances $\approx 50\,550$ MW

-> En moyenne, la consommation durant les vacances est ~ 6 % plus basses que le reste du temps (t-test p-value < 0.05 , Cohen's d = 0.293).

-> Pourquoi ? Fermeture partielle des écoles et des bureaux, moindre activité économique, etc.

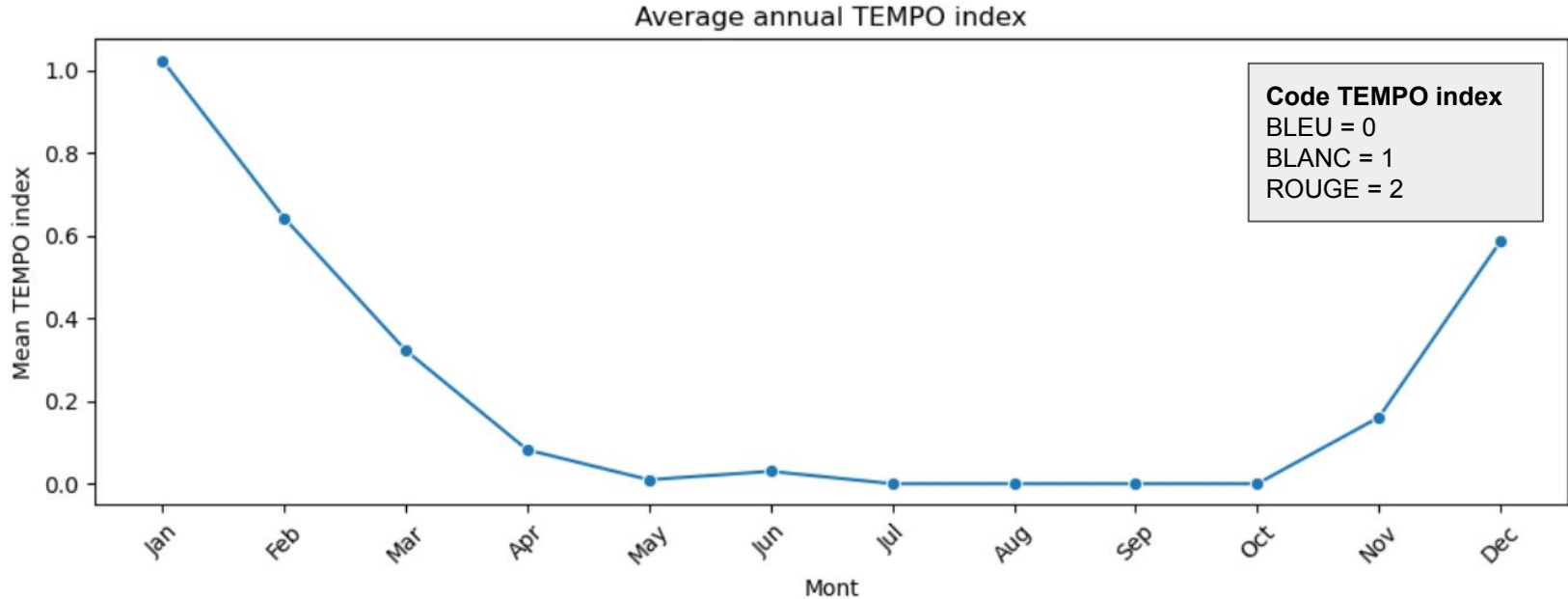
Impact des jours fériés

- Jours classiques $\approx 52\,741$ MW
- Jours fériés $\approx 45\,114$ MW

-> Chute spectaculaire (~ 15 %) de la demande lors des jours fériés (t-test p-value < 0.05 , Cohen's d = 0.657).

-> Pourquoi ? Presque toute l'industrie et les services sont à l'arrêt, seuls les usages domestiques persistent.

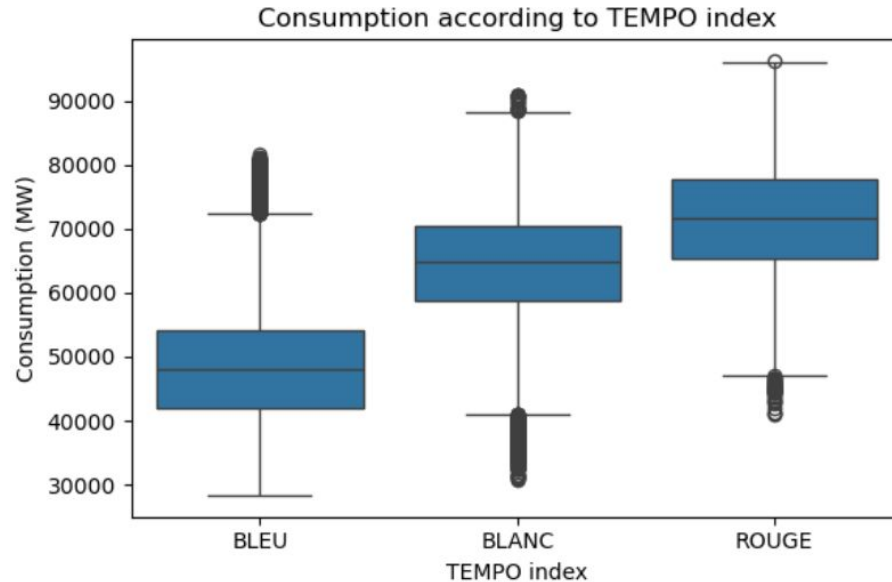
2. Analyse de données



- Ordonnée proche de 0 → quasi-exclusivement des jours BLEU (tarif le plus bas).
- Ordonnée ≈ 1 → mix de jours BLEU et BLANC (tarif moyen).
- Ordonnée > 1 → présence non négligeable de jours ROUGE (tarif le plus élevé). La proportion des index montre qu'il y a beaucoup moins de jours ROUGES (6%) que BLEU (81%) ou BLANC (12%), ce qui explique pourquoi 1 est à peine dépassé.

-> Le prix est haut en hiver et bas en été.

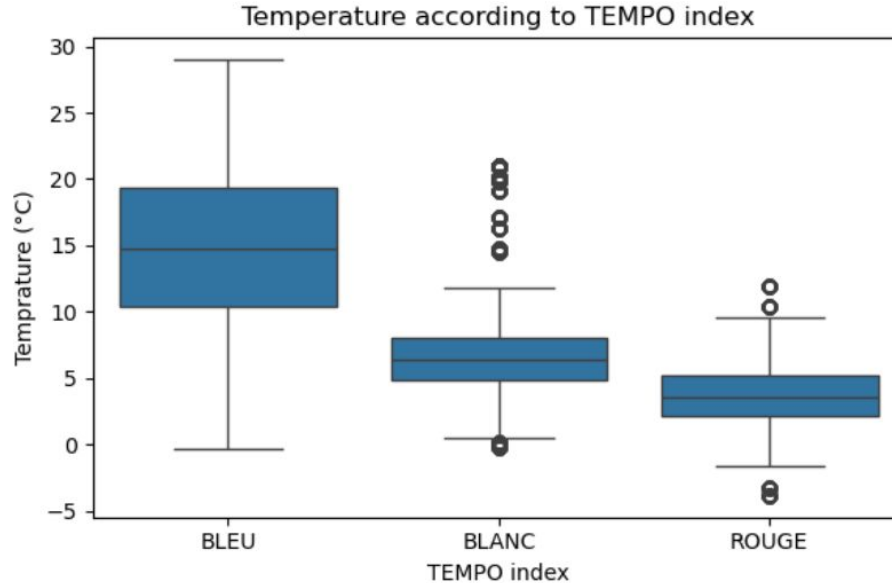
2. Analyse de données



- BLEU < BLANC < ROUGE : le tarif TEMPO est justement calibré pour être plus élevé les jours où la demande (consommation) est forte.
- L'IQR s'élargit de BLEU à ROUGE : les jours "Rouge" voient non seulement une plus haute consommation moyenne, mais aussi des fluctuations plus marquées.

-> Le découpage TEMPO capte bien les régimes de basse, moyenne et haute demande.

2. Analyse de données



- Plus l'indice passe de BLEU → BLANC → ROUGE, plus la température médiane décroît.
- Les jours bleus couvrent un spectre très large (0–29 °C).
- Quelques journées atypiques : Des journées assez chaudes classées « Blanc » (pic solaire, vent, incidents réseau ?), des journées très froides à 0°C où l'indice reste « Blanc » (mauvais prévisions ?), etc ?

-> TEMPO synthétise bien l'effet thermique : les jours rouges sont typiquement les plus froids, les jours bleus les plus doux ou chauds.

2. Analyse de données

Conclusion

- **Saisonnalité forte** : la consommation augmente en hiver et diminue l'été.
- **Effet de l'heure de la journée** : la consommation augmente en journée et diminue en soirée.
- **Effet jour de semaine vs. week-end** : la consommation augmente la semaine et diminue le week-end.
- **Vacances & jours fériés** : la consommation diminue les vacances et les jours fériés.
- **Jours TEMPO** : signal additionnel de l'effet de la température sur la consommation.

3. Prétraitement et modèle

Features sélectionnées

- **HDD et CDD (num)** : Heating/Cooling Degree Days sont des indicateurs de la rigueur du climat pour le chauffage et la climatisation qui traduisent l'écart de température par rapport à une base (18 °C).
- **hour_sin et hour_cos (num)** : Représentation cyclique de l'heure (sinus & cosinus) pour modéliser le cycle journalier de la consommation.
- **day_of_week (cat)** : Jour de la semaine (0 = Lundi, ..., 6 = Dimanche). L'EDA montre que la consommation est ~11 % plus basse le week-end.
- **lag_24h (num)** : Consommation 24 h avant à la même heure. Capturer l'autocorrélation jour-à-jour est utile parce que la demande électrique à une heure donnée ressemble souvent à celle de la veille à la même heure.
- **roll_mean_7d (num)** : Moyenne glissante sur 7 jours de la consommation à la même heure. Permet de donner la tendance hebdomadaire de la consommation.
- **resid_j-1 (num)** : *Consommation (J) - Prédiction (faite à J-1 pour J), avec Prédiction (faite à J-1 pour J) = Consommation (J-1)*. Autrement dit, hier à heure H, la prédiction pour aujourd'hui à H s'est trompée de $\pm X$ MW.
- **type_de_jour_TEMPO (cat)** : Catégorie réglementaire RTE qui permet de capturer la contrainte réseau et incitations prix selon le plan TEMPO.
- **bank_holidays (binaire)** : Permet de capter la baisse de consommation les jours fériés.
- **school_holidays (binaire)** : Permet de capter la baisse de consommation durant les vacances.

3. Prétraitement et modèle

Scaling, encoding et imputing

- **Imputing**
 - **HDD et CDD** : Remplacement des NaN par la moyenne et création d'un flag *_missing pour informer le modèle de l'absence de ces données.
 - **type_de_jour_TEMPO** : Remplacement des jours manquants par la modalité par défaut (BLEU).
- **Encoding**
 - **day_of_week** : Transformation en 6 indicateurs binaires (Lundi, ..., Samedi) pour casser la colinéarité linéaire et réduire le nombre de features.
 - **Type de jour TEMPO (après imputation)** : Transformation en 2 indicateurs binaires (BLANC, ROUGE) pour casser la colinéarité linéaire et réduire le nombre de features.
 - **bank_holidays, school_holidays** : Utilisé tel quel.
- **Scaling**
 - **resid_j-1, hour_sin, hour_cos, lag_24h, roll_mean_7d** : Centré à 0 et réduit à écart-type 1 pour un entraînement stable.

3. Prétraitement et modèle

Variable cible

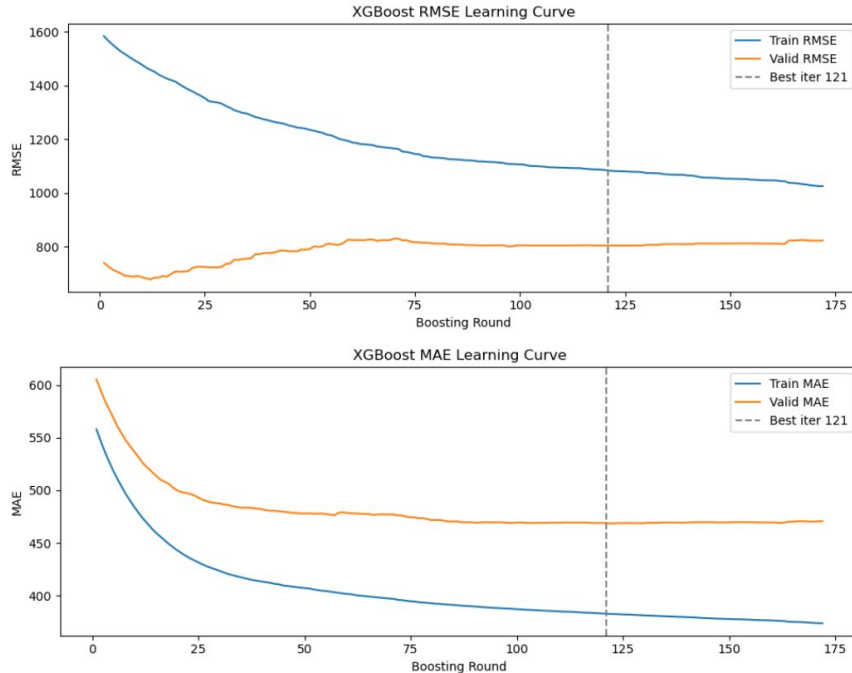
$Y = \text{Consommation}(J + 1) - \text{Prévision}(\text{faite à } J \text{ pour } J+1)$

- On connaît *Prévision (faite à J pour J+1)*, qui est la valeur de consommation d'aujourd'hui pour le même créneau (prévision naïve).
- On prédit Y .
- On peut ensuite reconstruire $\text{Consommation}(J + 1) = Y + \text{Prévision}(\text{faite à } J \text{ pour } J+1)$

Modèle

```
xgb = XGBRegressor(  
    n_estimators = 500,  
    learning_rate = 0.05,  
    max_depth = 7,  
    objective = 'reg:squarederror',  
    random_state = 42,  
    verbosity = 1,  
    eval_metric=['rmse','mae'],  
    early_stopping_rounds = 50  
)
```

4. Résultats



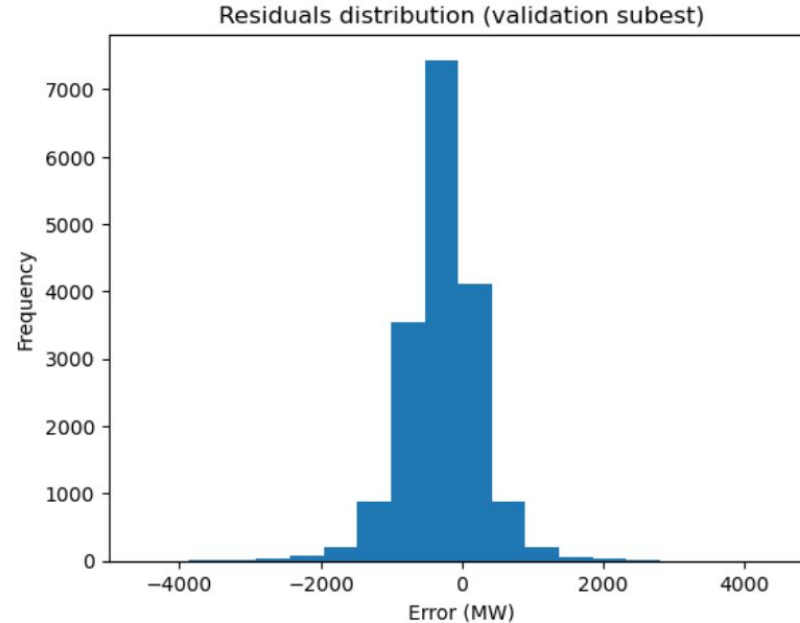
- Phase initiale (rounds 0–10) : forte baisse de l'erreur sur train et validation (underfitting réduit).
- Rounds 10–60 : la validation RMSE/MAE remonte alors que le train continue de baisser (sur-apprentissage modéré).
- Au-delà, la courbe de validation se stabilise autour de l'itération optimale (~121).

Échantillon d'évaluation : RMSE = 804.39, MAE = 468.53

Échantillon de test : RMSE = 572.47, MAE = 442.47

- Les erreurs sont dans les normes d'une prévision de charge électrique à J+1 définies par la littérature scientifique (MAE de l'ordre de 1 % et 3 %, et RMSE de l'ordre de 2 % à 4 %),
- Le gain sur le sous-ensemble de test montre peut-être un problème légèrement « trop facile » sur 2023, un léger sur-apprentissage sur 2023 (échantillon d'évaluation), ou un problème de fuite de données de l'échantillon de test.

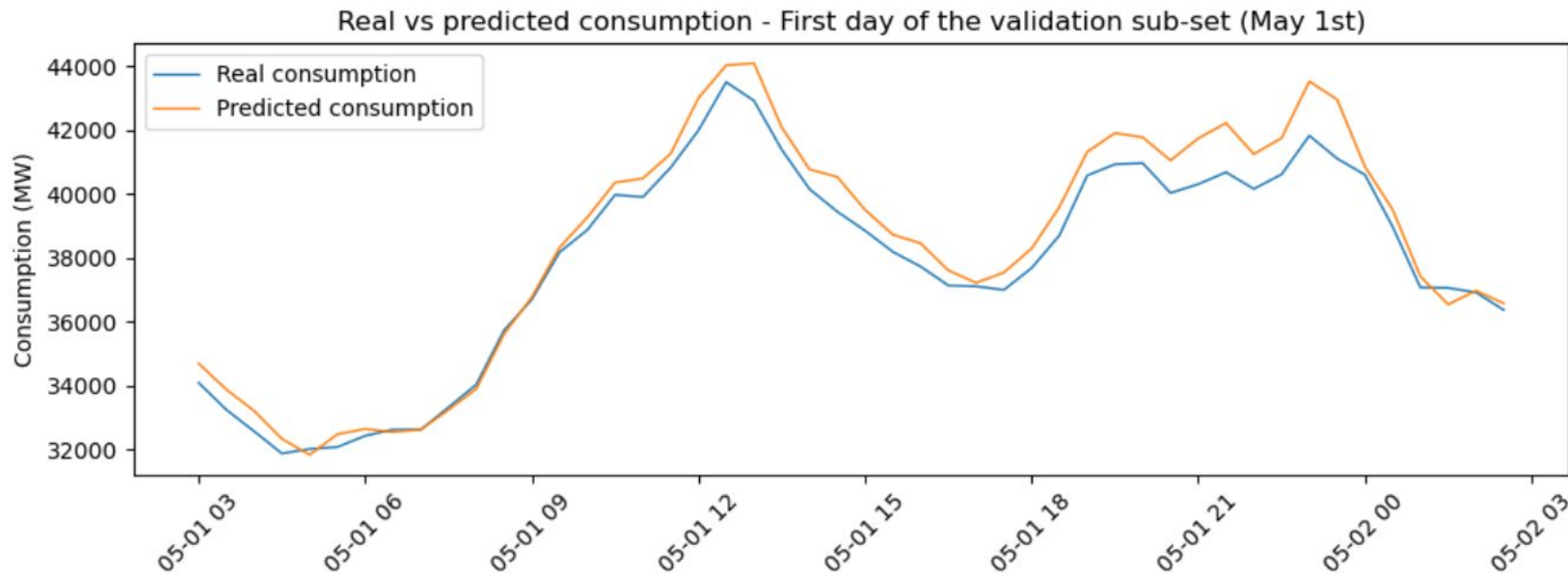
4. Résultats



Residuals mean : -308.48149767559795 – Median : -264.74608612060547
t = -52.29, p-value = 0.000e+00
Cohen's d = -0.395 (effect size)

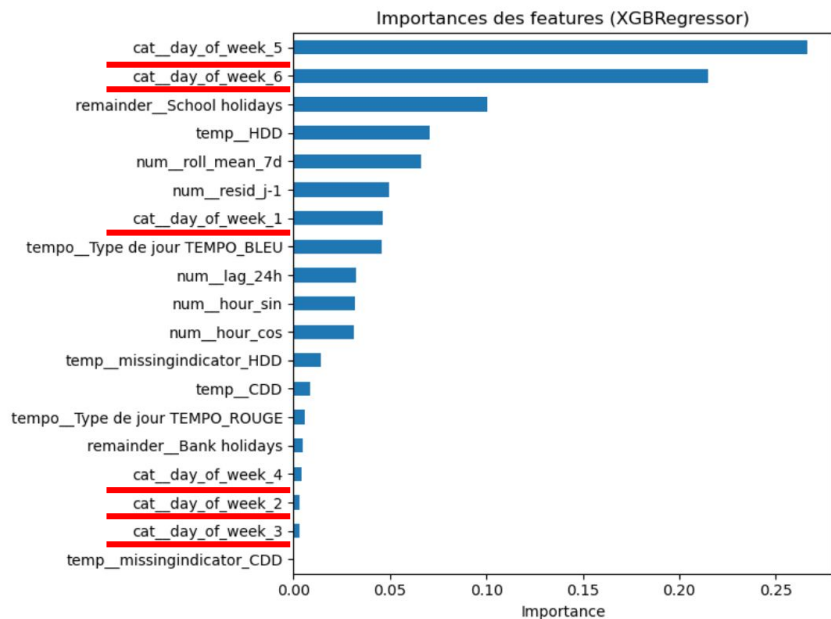
- **Négatif** -> on pourrait s'attendre à une distribution centrée sur 0, mais il y a un biais de sur-prévision significatif (p-value < 0) et modéré (Cohen's d = -0.395).
- **Positif** -> les résidus sont néanmoins densément centrés sur 0.

4. Résultats



- La courbe du modèle colle bien au cycle de la journée.
- On observe la sur-prévision identifiée précédemment.
- Le modèle semble lisser légèrement certaines variations.

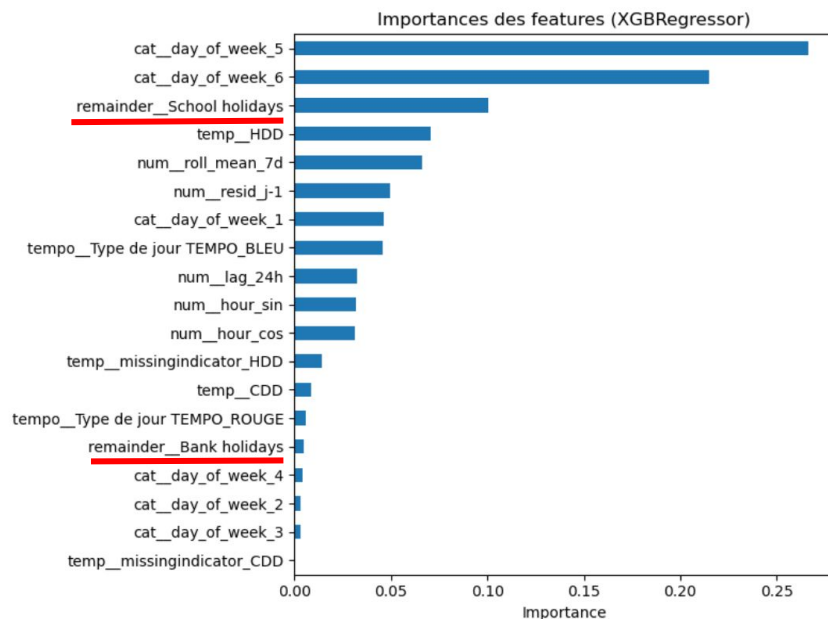
4. Résultats



Les variables de jour de semaine dominant :

- Les deux plus grosses importances tombent sur *cat_day_of_week_5* et *cat_day_of_week_6*, qui correspondent respectivement aux samedis et dimanches (avec un encodage où lundi = 0). Le modèle capte avant tout la baisse de conso le week-end.
- *cat_day_of_week_1* (mardi) est plus bas que samedi et dimanche, mais aussi beaucoup plus haut que mercredi, jeudi et vendredi. Pourquoi ? Vu que *cat_day_of_week_1* traduit la différence entre mardi et lundi (qui est la catégorie de référence) il est probable qu'on observe ici le rebond d'activité du au retour de congés du week-end, au retour du télétravail le lundi, ou plus largement au démarrage de la semaine.

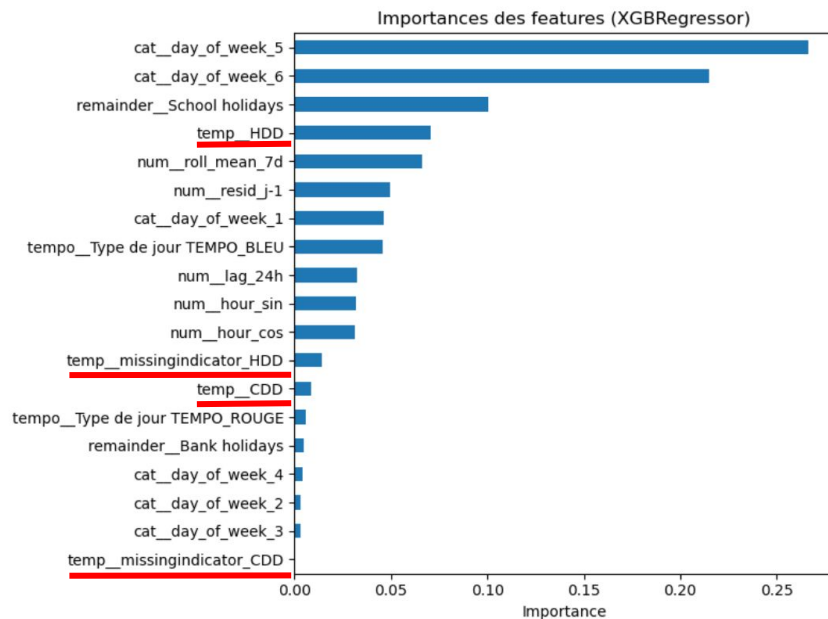
4. Résultats



Vacances scolaires vs jours fériés :

- L'indicateur *remainder__School holidays* se classe troisième, montrant que les vacances scolaires ont un impact significatif (baisse de la demande).
- À l'inverse, *remainder__Bank holidays* (jours fériés) pèse très peu : même si on a vu dans l'EDA que la consommation baissait beaucoup durant les jours fériés, l'effet ponctuel semble beaucoup moins structurant qu'une semaine de vacances entière.

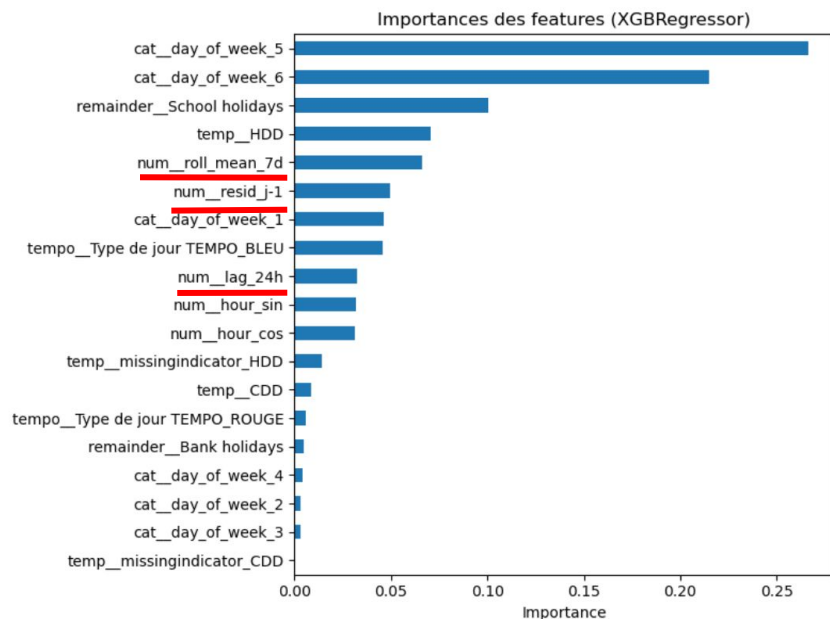
4. Résultats



Météo :

- *temp__HDD* (Heating Degree Days) pèse beaucoup plus lourd que *temp__CDD* (Cooling Degree Days), ce qui est cohérent pour le climat tempéré de la France, où le chauffage est un moteur de demande plus stable que la climatisation.
- *temp__missingindicator_HDD* et *temp__missingindicator_CDD* apportent vraiment peu d'information additionnelle.

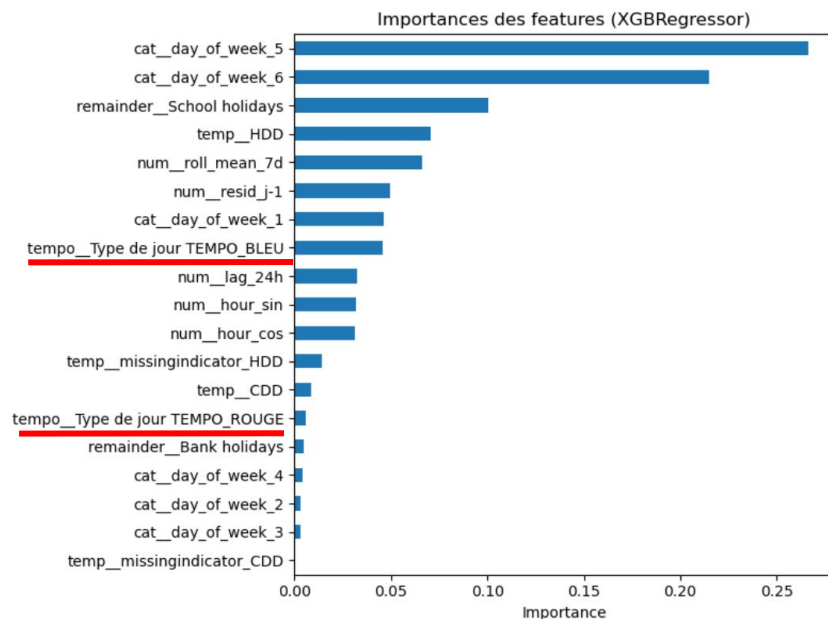
4. Résultats



Les lags :

- *num__roll_mean_7d* pèse assez lourd et arrive juste après *temp__HDD*, ce qui montre que la tendance de consommation hebdo est structurante.
- *num__resid_j-1* (erreur de la prévision naïve d'hier) est aussi importante, pas beaucoup moins que *num__roll_mean_7d*, ce qui montre que le modèle s'appuie autant sur la tendance hebdo que sur la performance de la prévision naïve précédente.
- *num__lag_24h* et les composantes cycliques *hour_sin* et *hour_cos* ont un rôle plus modéré mais non négligeable. Leur information est peut-être en partie redondante avec le rolling ?
- C'était une plutôt bonne idée de créer ces features !

4. Résultats



TEMPO :

- *TEMPO BLEU* (jours les moins chers) a une importance modérée mais non négligeable, du au fait qu'il représente les jours majoritaires et homogènes d'un point de vue conso.
- *TEMPO ROUGE* (jours les plus chers) a une importance très faible, probablement du au fait qu'il n'y a qu'une poignée de jours ROUGE dans le dataset, mais aussi probablement qu'il représente des jours plus chaotiques, qui coïncident souvent avec des pointes de froid extrême et d'autres facteurs moins facilement prévisible.