Prédiction de la variation du prix de l'électricité en France et en Allemagne

Statistical Learning

Tania Admane et Lola Chardigny

7 janvier 2026

Table des matières

1 Introduction

Nous avons récupéré la base de données mise à disposition par QRT dans le cadre du Data-Challenge 2023. L'objectif est de proposer un modèle pour expliquer les mouvements de variations des prix de l'électricité en Allemagne et en France. Dans un premier temps, nous allons traiter nos données, ensuite nous proposerons 4 types de modèles différents : une régression linéaire, une régression linéaire avec régularisation, un Random Forest, un LightGBM et un réseau de neurones. Enfin, nous apporterons des améliorations pour ne conserver qu'un unique modèle.

2 Traitement des données

2.1 Données initiales

Le dataset initial contient 35 colonnes :

— `ID` : Identifiant unique associé à un jour (`DAY_ID`) et à un pays (`COUNTRY`).
— `DAY_ID` : Identifiant du jour. Les dates ont été anonymisées tout en préservant la structure temporelle des données.
— `COUNTRY` : Identifiant du pays (`DE` pour l'Allemagne, `FR` pour la France).
— `GAS_RET` : Variation du prix du gaz en Europe.
— `COAL_RET` : Variation du prix du charbon en Europe.
— `CARBON_RET` : Variation des Futures sur les émissions de carbone.
— `x_TEMP` : Température du pays x.
— `x_RAIN` : Précipitations du pays x.
— `x_WIND` : Vent du pays x.
— `x_GAS` : Production d'électricité à partir de gaz naturel.
— `x_COAL` : Production d'électricité à partir de charbon.
— `x_HYDRO` : Production hydroélectrique.
— `x_NUCLEAR` : Production nucléaire.
— `x_SOLAR` : Production photovoltaïque.
— `x_WINDPOW` : Production éolienne.
— `x_LIGNITE` : Production à partir de lignite.
— `x_CONSUMPTION` : Consommation totale d'électricité.
— `x_RESIDUAL_LOAD` : Consommation d'électricité après prise en compte des énergies renouvelables.
— `x_NET_IMPORT` : Quantité d'électricité importée depuis l'Europe.
— `x_NET_EXPORT` : Quantité d'électricité exportée vers l'Europe.
— `DE_FR_EXCHANGE` : Électricité échangée de l'Allemagne vers la France.
— `FR_DE_EXCHANGE` : Électricité échangée de la France vers l'Allemagne.

2.2 Valeurs manquantes

Nous choisissons de compléter les valeurs manquantes à l'aide d'interpolation polynomiale. Au début, nous avions trié les données par ordre chronologique, puis nous avions utilisé la méthode `ffill`, c'est-à-dire propager la dernière valeur disponible, mais cette première méthode ne nous paraissait pas pertinente au vu des données.
Nous transformons la variable catégorielle `COUNTRY` en variable binaire (FR = 1, DE = 0).
Après ce traitement, aucune donnée manquante ne subsiste dans le dataset.

2.3 Valeurs extrêmes

Plusieurs variables présentent des valeurs extrêmes. Cependant, ces valeurs n'ont pas été supprimées car elles correspondent à des événements économiques et météorologiques réels (pics

de prix de l'énergie, stress du réseau, conditions exceptionnelles) et font partie intégrante du phénomène étudié. Au début, nous avions utilisé une méthode de suppression d'outliers basée sur un calcul de Z-score, néanmoins nous n'avons pas retenu cette méthode car elle constitue une perte d'information économiquement et météorologiquement pertinente.

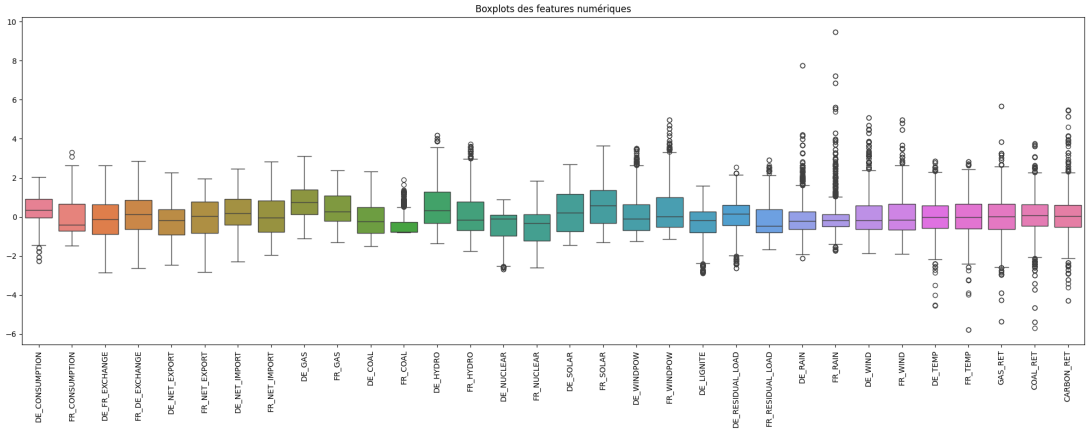

FIGURE 1 – Valeurs extrêmes – Boxplots des features numériques

2.4 Ingénierie des features

Nous créons de nouvelles variables qui nous semblent intéressantes. Pour la plupart des variables suivantes, il faut travailler indépendamment sur le dataset allemand, puis français, avant de les reconcaténer.

- `time_block` : Les jours sont séparés en quatre quartiles afin de représenter approximativement quatre saisons annuelles.
- `GAS_RET_i` : Moyenne mobile sur i jours des variations du prix du gaz en Europe.
- `COAL_RET_i` : Moyenne mobile sur i jours des variations du prix du charbon en Europe.
- `CARBON_RET_i` : Moyenne mobile sur i jours des variations du prix des contrats à terme sur les émissions de carbone.
- `PREVIOUS_Y` : Dernière variation du prix de l'électricité disponible.
- `VARIATION_TEMP` : Variation de la température par rapport à la dernière valeur disponible.
- `x_WIND_WINDPOW` : Interaction Vent × Production éolienne dans le pays x.
- `x_TEMP_RESLOAD` : Interaction Température × Residual Load dans le pays x.
- `x_RAIN_HYDRO` : Interaction Pluie × Production hydraulique dans le pays x.
- `LOAD_IMPORT_x` : Interactions entre le Residual Load et les importations nettes pour le pays x.
- `x_NON_RENEWABLE` : Production d'énergie non renouvelable (gaz + charbon + lignite + nucléaire).
- `x_RENEWABLE` : Production d'énergie renouvelable (hydraulique + solaire + éolien).
- `x_PRODUCTION` : Production totale + importations nettes dans le pays x.
- `x_EQUILIBRIUM` : Production − consommation dans le pays x.

2.5 La multicolinéarité

Nous calculons la matrice de corrélation sur les features pour repérer les variables fortement corrélées, pour éviter la multicolinéarité.

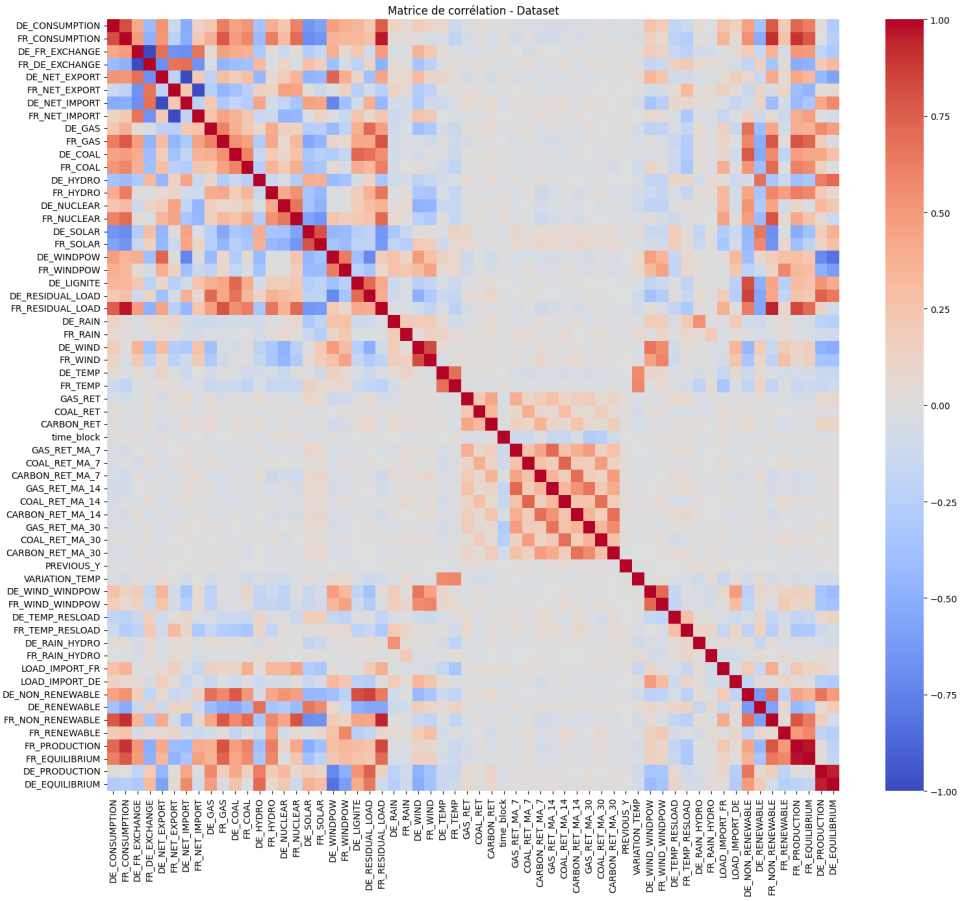

FIGURE 2 – Matrice de corrélation des features numériques

Nous décidons de supprimer les variables ayant des corrélations supérieures à 0.80.

2.6 Corrélation avec la target Y

Nous réalisons un calcul de corrélation de Spearman entre les features et la variable target Y.

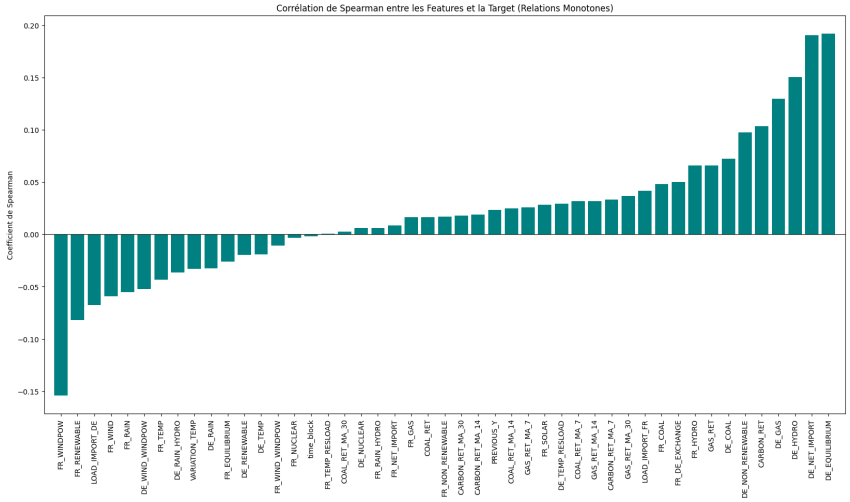

FIGURE 3 – Corrélation de Spearman entre les features et la variable cible

Plusieurs variables ont une corrélation très faible ($|\text{corrélation}| < 0.03$) avec la target. Nous

décidons donc de les supprimer.

2.7 Séparation du dataset

La temporalité n'est pas respectée dans le dataset initial ; les données ne sont pas ordonnées chronologiquement. Il est donc nécessaire d'ordonner le dataset avant de le séparer en ensembles d'entraînement et de validation pour éviter toute fuite de données. La séparation a été effectuée en respectant la chronologie au sein des données allemandes et françaises séparément, avec un ratio 80/20 % pour le jeu d'entraînement et le jeu de validation.

2.8 Normalisation

Les variables ont des échelles très différentes et certaines contiennent des valeurs extrêmes. Il est donc nécessaire de standardiser les données. Le scaler a été ajusté uniquement sur l'ensemble d'entraînement, puis appliqué sur l'ensemble d'entraînement et de validation.

3 Modélisation

L'objectif du DataChallenge est de capter correctement l'intensité relative et le sens des mouvements des futures de l'électricité. QRT recommande donc d'utiliser une corrélation de Spearman pour l'évaluation des modèles, car elle mesure la qualité du classement plutôt que l'erreur en niveau.

Modèle	**Loss func**	**Spearman Val**	**Spearman Test**	**MAE**
Linear Regression	MAE	0.22	0.29	0.55
Lasso	MAE	0.26	0.26	0.55
Random Forest	MAE	0.27	0.47	0.55
LightGBM	MAE	0.28	0.63	0.55
Réseau de neurones	MAE	0.26	0.29	0.53

TABLE 1 – Comparaison des performances des modèles

3.1 Régression linéaire

— Nous minimisons une MAE plutôt que la RMSE, pour être plus robustes aux valeurs extrêmes.
— Le score de Spearman est plus élevé sur l'ensemble d'entraînement que sur l'ensemble de validation, indiquant un surapprentissage.

3.2 Lasso

— On introduit une régularisation $L1$ pour tenter de réduire l'overfitting et faire de la sélection de variables.
— Les résultats obtenus sont meilleurs que ceux obtenus avec la régression linéaire. D'une part, notre score de Spearman obtenu sur le jeu de validation (26 %) est meilleur. D'autre part, l'écart entre les scores de Spearman sur les jeux d'entraînement et de validation est nettement plus faible.
— Le modèle est donc plus fiable que le premier.

3.3 Random Forest

— Les modèles linéaires ne sont pas suffisants pour prédire les variations des prix de l'électricité. Nous allons donc utiliser un modèle non linéaire. Le Random Forest avec MAE est un bon choix de modèle ensembliste car il est robuste au bruit et à la multicolinéarité.
— Le meilleur modèle sélectionné par le GridSearch arrive à légèrement accroître notre score de Spearman sur le jeu de validation (27 %).
— On note tout de même que l'écart entre le score de Spearman sur les jeux d'entraînement et de validation s'est creusé.

3.4 LightGBM

— À la différence du Random Forest, les arbres sont construits séquentiellement. Chaque arbre essaie de corriger les erreurs du précédent.
— Le modèle améliore légèrement le score de Spearman sur le jeu de validation (28 %).

3.5 Réseau de neurones

— Nous choisirons de tester un réseau de neurones. Après plusieurs tests, c'est un réseau de neurones très simple à 1 couche cachée et 4 neurones qui est le plus performant. Cela peut s'expliquer par la taille limitée du dataset et la nature majoritairement linéaire des relations entre les variables. Un modèle plus complexe conduirait à du surapprentissage.
— Pour éviter le surapprentissage, on utilise un bruit gaussien léger pendant l'entraînement pour forcer la robustesse, une régularisation L2 pour pénaliser les poids trop importants et un dropout de 20 %
— On entraîne un ensemble de 7 modèles avec des initialisations différentes pour réduire la variance des prédictions, puis on moyenne leurs sorties.
— On obtient un des meilleurs resultats jusque la avec un écart raisonnable entre le set d'entraînement et de validation

Nous ne sommes pas totalement convaincus par nos modèles, nous proposons donc de construire un modèle unique pour les variations des prix de l'électricité allemande.

4 Amélioration : un modèle allemand

Nous construisons un dataset contenant uniquement les variations allemandes. Nous pouvons nous permettre de prendre cette décision puisque le dataset allemand contient plus de 500 observations.

Notre modèle le plus robuste jusqu'à présent était le LASSO, en considérant plusieurs critères : le score de Spearman sur la validation, l'écart entre les scores d'entraînement et de validation, et la MAE.

4.1 Modèle LASSO Allemagne

TABLE 2 – Comparaison des performances : modèle global vs modèle allemand

Modèle	**Spearman Train**	**Spearman Val**	**MAE**
LASSO Allemagne	0.38	0.38	0.57
LASSO	0.26	0.26	0.55

— Les résultats montrent qu'une approche pour l'Allemagne, utilisant un modèle LASSO avec régularisation, permet d'obtenir de meilleures performances.
— Il n'y a pas de problème d'overfitting, la différence entre le Spearman sur le jeu de test et sur le jeu de validation est presque nulle.
— La fonction de perte MAE s'est révélée judicieuse dans ce contexte où l'objectif est le classement plutôt que la précision absolue.

5 Conclusion

Cette étude démontre l'importance d'une préparation soigneuse des données pour la prédiction des prix de l'électricité. L'ingénierie des features, notamment la création d'interactions physiquement significatives et de moyennes mobiles, est primordiale pour capturer les dynamiques du marché. Pour améliorer nos prédictions, il faudrait continuer à rechercher des prédictions avec des modèles de deep learning plus complexes.

Références

— Weron, R. (2014). Electricity price forecasting : A review of the state-of-the-art with a look into the future.
— Ziel, F., & Weron, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures.
— Gianfreda, A., Parisio, L., & Pelagatti, M. (2016). Revisiting long-run relations in power markets.