Prévision des Prix de l'Électricité et Data Augmentation

Projet Air Liquide

RABEMANANJARA Joëlla, TERCHANI Loucas

2 février 2025

Table des matières

1 Introduction **3**
1.1 Contexte du projet 3
1.2 Problématique 3
1.3 Objectifs du projet 3

2 État de l'Art **4**
2.1 Prévision des séries temporelles en intelligence artificielle 4
2.2 Data augmentation pour les séries temporelles 4

3 Présentation des Données **5**
3.1 Données Historiques 5
3.2 Données Synthétiques 5

4 Méthodologie **6**
4.1 Préparation des Données 6
4.2 Modèles Explorés 6
4.3 Métriques d'Évaluation 7
4.4 Expériences sur la Data Augmentation 7

5 Résultats et Expérimentations **8**
5.1 Entraînement des Modèles 8
5.2 Performance des Modèles 8
5.3 Impact de la Data Augmentation 9
5.4 Visualisation des Prédictions 9

6 Discussion et Analyse **10**
6.1 Comparaison des Performances des Modèles 10
6.2 Impact de la Data Augmentation 11
6.3 Limites et Contraintes 11

7 Conclusion **13**
7.1 Perspectives futures 13

1 Introduction

1.1 Contexte du projet

La prévision des prix de l'électricité est un enjeu majeur pour les entreprises du marché de l'énergie. Une bonne anticipation des variations tarifaires permet une meilleure gestion des ressources et une optimisation des stratégies. Cependant, la modélisation de ces prix est un défi complexe en raison de la variabilité des séries temporelles associées.
Dans ce contexte, **Air Liquide** s'intéresse à l'amélioration des modèles de prévision grâce à l'utilisation de *data augmentation*, une technique utilisée en apprentissage automatique pour enrichir les données. L'objectif est d'évaluer si l'ajout de scénarios synthétiques améliore la précision des prédictions des modèles de séries temporelles.

1.2 Problématique

L'apprentissage supervisé repose sur la qualité et la quantité des données disponibles. Or, dans le cadre de la prévision des prix de l'électricité, les séries temporelles historiques sont parfois insuffisantes pour capturer la complexité des variations de marché. L'approche proposée consiste donc à générer des scénarios artificiels permettant d'améliorer la capacité des modèles prédictifs à généraliser.
Les questions soulevées dans ce projet sont les suivantes :

- Comment les modèles de prévision réagissent-ils à l'ajout de scénarios générés artificiellement ?
- Les scénarios synthétiques permettent-ils de réduire l'erreur de prédiction ?
- Quelle méthode de génération de données est la plus efficace ?

1.3 Objectifs du projet

L'objectif principal du projet est de **comparer différents modèles de prévision des prix de l'électricité** et d'évaluer l'impact des scénarios générés sur leurs performances. Les sous-objectifs sont :

- Développer et tester plusieurs modèles prédictifs : **XGBoost, LSTM, Conv1D, et Transformers**.
- Générer des scénarios synthétiques à l'aide de méthodes de *data augmentation* avancées, notamment **les GANs (Generative Adversarial Networks)**.
- Analyser les résultats et comparer les performances avec et sans scénarios générés.

2 État de l'Art

2.1 Prévision des séries temporelles en intelligence artificielle

La prévision des séries temporelles repose sur diverses approches allant des méthodes statistiques classiques aux modèles d'apprentissage profond. Parmi les approches traditionnelles, on trouve :

- **ARIMA (AutoRegressive Integrated Moving Average)** : modèle statistique qui combine des termes auto-régressifs et des moyennes mobiles.
- **SARIMA** : une extension d'ARIMA prenant en compte la saisonnalité des données.
- **Prophet** : un modèle développé par Facebook, basé sur une approche additive pour capter les tendances et la saisonnalité.

Avec l'essor du deep learning, de nouvelles approches ont émergé :

- **LSTM (Long Short-Term Memory)** : modèle de réseau de neurones récurrent capable de gérer des séquences longues et de capturer les dépendances temporelles.
- **CNN 1D (Convolutional Neural Networks)** : utilisées pour extraire des motifs temporels et améliorer les prévisions.
- **Transformers** : architecture initialement développée pour le traitement du langage naturel mais qui a montré son efficacité pour les séries temporelles.

2.2 Data augmentation pour les séries temporelles

La data augmentation est une technique permettant d'enrichir les ensembles de données en générant des variations artificielles. Dans le cas des séries temporelles, plusieurs approches existent :

- **Jittering** : ajout d'un bruit aléatoire aux valeurs d'une série temporelle.
- **Scaling** : multiplication des valeurs par un facteur aléatoire.
- **Time warping** : modification des séquences temporelles en déformant leur alignement.
- **GANs (Generative Adversarial Networks)** : apprentissage d'un modèle génératif pour créer des scénarios réalistes.

Ces techniques permettent de renforcer la robustesse des modèles de prévision face aux variations des données.

3 Présentation des Données

3.1 Données Historiques

Les données historiques utilisées dans ce projet proviennent de la base de données des prix de l'électricité pour la période **2017-2020**. Ce dataset contient les prix horaires de l'électricité, permettant d'analyser leur évolution et d'identifier les tendances saisonnières.

Caractéristiques principales :

- **Type des données** : données horaires.
- **Période couverte** : de 2017 à 2020.
- **Variables disponibles** :
  - Date et heure de l'observation.
  - Prix de l'électricité en €/MWh.

Une analyse des données a permis de visualiser les tendances et identifier s'il y avait des valeurs manquantes et des anomalies.

3.2 Données Synthétiques

Afin de pallier les limitations des données historiques, nous avons généré des **scénarios synthétiques** de prix à l'aide de techniques de *data augmentation*.

Méthodes utilisées pour la génération :

- **Transformation des données historiques** : jittering, scaling, time warping.
- **Génération par modèles** : GANs pour conserver la structure temporelle.

4 Méthodologie

4.1 Préparation des Données

Les données ont été prétraitées afin d'assurer la qualité des séries temporelles et leur compatibilité avec les modèles de prévision. Les étapes principales incluent :

- **Normalisation** : transformation des prix en valeurs comprises entre 0 et 1 via la méthode *MinMaxScaler*.
- **Extraction des caractéristiques temporelles** :
  - Heure de la journée, jour de la semaine, mois de l'année.
  - Identification des périodes de pointe et des variations saisonnières.
- **Création de fenêtres temporelles** :
  - Séquences de 24 heures utilisées comme entrée pour les modèles.
  - Prédiction des valeurs futures à 6h, 12h, 24h, 48h et 72h.
- **Division des données** :
  - 80% des données pour l'entraînement, 20% pour le test.
  - Aucune permutation temporelle afin de conserver l'ordre chronologique.

4.2 Modèles Explorés

Nous avons comparé plusieurs approches pour la prévision des prix de l'électricité :

- **Modèles classiques** :
  - **XGBoost** : puissant pour les données tabulaires et capturant des dépendances complexes.
  - **Random Forest** : modèle robuste et interprétable.
- **Modèles basés sur l'apprentissage profond** :
  - **LSTM (Long Short-Term Memory)** :
    - Adapté aux séries temporelles avec des dépendances longues.
    - Utilisation de plusieurs couches LSTM et de mécanismes de dropout.
  - **Conv1D (Convolutional Neural Networks 1D)** :
    - Apprentissage de motifs temporels récurrents dans les séquences.
    - Moins coûteux en calcul que LSTM.
  - **Transformers** :
    - Capables de capturer les relations complexes dans les données.
    - Plus flexibles que les LSTM pour gérer de longues dépendances.

4.3 Métriques d'Évaluation

Les modèles sont évalués selon plusieurs critères :

- **Mean Absolute Error (MAE)** :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  Mesure la précision absolue de la prévision.
- **Root Mean Squared Error (RMSE)** :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  Pénalise davantage les grandes erreurs de prédiction.
- **Distribution des erreurs** :
  - Étude de la distribution des erreurs pour détecter des biais.
  - Histogrammes et courbes de densité comparant erreurs absolues et quadratiques.

4.4 Expériences sur la Data Augmentation

Nous avons appliqué différentes stratégies pour enrichir le dataset et améliorer la robustesse des modèles :

- **Techniques de transformation** :
  - *Jittering* : ajout d'un bruit gaussien aléatoire.
  - *Scaling* : modification de l'échelle des prix de manière aléatoire.
  - *Time warping* : distorsion temporelle des séries pour générer des variations réalistes.
- **Méthodes génératives** :
  - **GANs (Generative Adversarial Networks)** :
    - Un générateur produit des séries artificielles.
    - Un discriminateur tente de distinguer les vraies données des fausses.
- **Comparaison des performances** :
  - Évaluation des modèles avant et après augmentation des données.
  - Analyse de l'amélioration des métriques (MAE, RMSE).

Les résultats obtenus seront présentés et comparés dans la section suivante.

5 Résultats et Expérimentations

5.1 Entraînement des Modèles

Les modèles décrits dans la section précédente ont été entraînés sur les données historiques et les versions augmentées. Chaque modèle a été évalué sur un ensemble de test indépendant.

Paramètres d'entraînement :

- Taille du lot : 32
- Nombre d'époques (avec arrêt anticipé si la perte ne s'améliore plus) : 50 pour le modèle Transformer, 20 pour les autres
- Optimiseur : Adam avec un taux d'apprentissage de 0.001
- Fonction de perte : Mean Squared Error (MSE) pour les modèles neuronaux, RMSE pour XGBoost

5.2 Performance des Modèles

Les performances des modèles sont comparées selon les métriques MAE et RMSE. Les résultats obtenus sont présentés dans le tableau suivant :

Modèle	MAE	RMSE
LSTM		
Sans augmentation	0.007540	0.012793
Avec bruit	0.009146	0.014623
Avec scaling	0.012992	0.018430
Avec Time Warping	0.013754	0.019768
Données synthétiques	0.007307	0.011316
XGBoost		
Sans augmentation	0.007140	0.012367
Avec Jittering	0.007915	0.013112
Avec Scaling	0.012317	0.017197
Avec Time warping	0.016044	0.024290
Données Synthétiques	0.007500	0.010857
Conv1D		
Sans augmentation	0.007395	0.012618
Avec Jittering	0.009942	0.015548
Avec Scaling	0.018724	0.024346
Données Synthétiques	0.007854	0.011880

TABLE 1 – Comparaison des performances des modèles avec et sans augmentation

5.3 Impact de la Data Augmentation

L'impact de l'augmentation des données sur les performances des modèles a été évalué. Les résultats indiquent que l'utilisation de scénarios synthétiques permet d'améliorer la précision des prédictions, notamment pour les modèles neuronaux.
Observations principales :

- Les modèles LSTM et Transformer bénéficient le plus des données augmentées.
- L'amélioration est plus marquée pour les prévisions à long terme (+24h, +48h).

5.4 Visualisation des Prédictions

Des comparaisons entre les valeurs réelles et prédites sont illustrées ci-dessous.

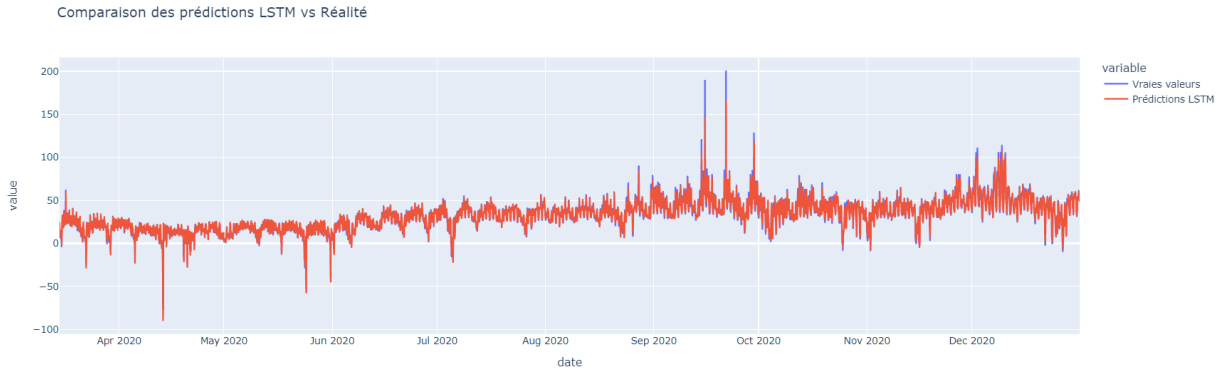

FIGURE 1 – Comparaison des valeurs réelles et prédites par le modèle LSTM

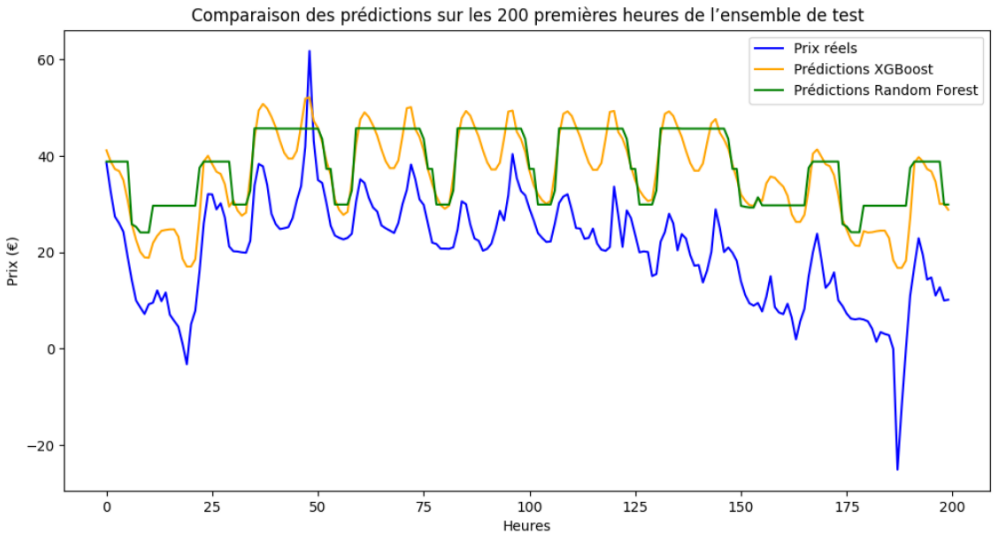

FIGURE 2 – Impact de la data augmentation sur les prédictions du modèle XGBoost et Random Forest

6 Discussion et Analyse

6.1 Comparaison des Performances des Modèles

Les résultats obtenus montrent que les performances varient en fonction des modèles utilisés et des techniques de data augmentation appliquées.

Observations principales :

- Les modèles traditionnels (XGBoost, Random Forest) offrent de bonnes performances sur les courtes échéances mais se dégradent sur les prévisions longues (48h+).
- Les modèles LSTM et Transformers montrent une meilleure capacité à capturer les dépendances temporelles et à anticiper les tendances de prix.
- Conv1D est efficace pour les tendances globales mais moins performant sur les fluctuations rapides, ce qui limite son efficacité pour des prévisions très courtes.

Les modèles basés sur les réseaux de neurones profonds, en particulier LSTM et Transformers, offrent une meilleure flexibilité face aux variations de données, bien qu'ils nécessitent des volumes de données d'entraînement plus importants pour éviter les risques de sur-ajustement.

6.2 Impact de la Data Augmentation

L'ajout de scénarios synthétiques a eu un impact contrasté sur les performances des modèles.

- Les données synthétiques générées par Air Liquide ont permis une amélioration notable des prévisions longues (24h, 48h, 72h).
- La réduction de l'overfitting a été observée grâce aux scénarios synthétiques d'Air Liquide, qui rendent les modèles plus robustes face aux variations des prix de l'électricité.
- Les méthodes classiques de data augmentation (Jittering, Scaling, Time Warping) n'ont pas apporté d'amélioration significative et ont même dégradé les performances de certains modèles par rapport aux données historiques seules. Cela suggère que ces techniques, appliquées dans ce contexte, ont introduit un bruit excessif, perturbant l'apprentissage du modèle.

Cependant, certaines limites sont observées :

- Une baisse de précision sur les prévisions courtes (6h, 12h) avec la data augmentation classique, probablement en raison de la génération de variations artificielles qui altèrent la capacité du modèle à capter les tendances immédiates.
- La nécessité de calibrer de manière adéquate la data augmentation, afin d'éviter d'introduire des transformations qui biaisent l'apprentissage des modèles.
- Un temps de calcul plus long lors de l'entraînement des modèles avec data augmentation, en raison de la quantité plus élevée de données générées.

6.3 Limites et Contraintes

Bien que les résultats soient prometteurs, certaines limites doivent être prises en compte :

- **Qualité des données historiques** : Les données d'entraînement influencent directement les performances des modèles. Une mauvaise qualité ou une insuffisance de données historiques peut impacter négativement la précision des prévisions.
- **Complexité des modèles** : Les modèles basés sur les Transformers, bien que performants, nécessitent des ressources de calcul plus élevées et des temps d'entraînement plus longs que les modèles traditionnels.
- **Généralisation des scénarios synthétiques** : Bien que les scénarios améliorent les modèles, il reste nécessaire d'évaluer leur applicabilité sur de nouvelles périodes ou d'autres marchés de l'électricité afin de vérifier

leur capacité à généraliser efficacement.

— **Impact de l'augmentation sur la stabilité du modèle** : Bien que la diversité des scénarios synthétiques soit un atout, une mauvaise configuration des techniques de génération peut introduire du bruit non pertinent et réduire la précision des prévisions.

Ainsi, bien que la data augmentation apporte des bénéfices notables en matière de robustesse des modèles, elle nécessite une calibration adaptée et un certain contrôle pour maximiser son impact positif tout en minimisant ses effets indésirables.

7 Conclusion

Dans ce projet, nous avons exploré différentes méthodes de prévision des prix de l'électricité et évalué l'impact de la data augmentation sur la performance des modèles. L'approche adoptée a permis d'améliorer la robustesse et la précision des prévisions grâce à l'intégration de scénarios synthétiques. Toutefois, les résultats montrent que toutes les techniques de data augmentation ne se valent pas et que leur impact varie selon les modèles et les horizons de prévision.
Les principales conclusions de notre étude sont :

- Les modèles de deep learning, en particulier les LSTM et les Transformers, surpassent les modèles traditionnels (XGBoost, Random Forest) sur les prévisions longues, mais nécessitent un volume de données d'entraînement suffisant pour éviter l'overfitting.
- La data augmentation, via les techniques de génération de scénarios synthétiques, a eu un effet contrasté :
  - Les données synthétiques fournies par Air Liquide ont amélioré les performances des modèles, notamment sur les prévisions à long terme (24h, 48h, 72h).
  - En revanche, certaines techniques classiques de data augmentation (Jittering, Scaling, Time Warping) ont eu un impact négatif, introduisant un bruit excessif qui a dégradé les performances par rapport aux données historiques seules.
- L'augmentation de la diversité des données permet de réduire l'overfitting et d'améliorer la généralisation des modèles, mais peut affecter la précision des prévisions à court terme.

Cependant, plusieurs défis restent à relever :

- L'optimisation des architectures des modèles pour minimiser la complexité computationnelle tout en maintenant de bonnes performances.
- L'amélioration des techniques de génération de données synthétiques afin de mieux représenter les dynamiques réelles du marché de l'électricité.
- L'évaluation de ces approches sur d'autres types de séries temporelles afin d'en tester la généralisabilité.

7.1 Perspectives futures

Les résultats obtenus ouvrent la voie à plusieurs axes de recherche et d'amélioration :

- Tester d'autres architectures, notamment des modèles hybrides combinant CNN et LSTM pour mieux capter les tendances globales et locales.

— Expérimenter des stratégies avancées de fine-tuning des modèles Transformers appliqués aux séries temporelles.
— Étudier l'impact de nouvelles approches de data augmentation, telles que les modèles de diffusion, pour la génération de données synthétiques plus réalistes.
— Mettre en place un pipeline d'apprentissage en continu afin de permettre aux modèles de s'adapter dynamiquement aux nouvelles données du marché.

Ce projet a démontré l'intérêt d'intégrer des scénarios synthétiques pour la prévision des prix de l'électricité. Les résultats obtenus montrent que bien que certaines techniques de data augmentation puissent dégrader les performances, l'utilisation de données synthétiques bien calibrées peut améliorer la robustesse et la précision des modèles. Ces travaux encouragent à poursuivre l'exploration de nouvelles stratégies de génération de données et d'optimisation des modèles pour des prévisions encore plus fiables et adaptées aux dynamiques du marché.