Étude de Cas : Prédiction des Prix Quotidiens des Cryptomonnaies BTC

Introduction

Cette étude se concentre sur le développement d'une méthode rigoureuse et méthodique pour prédire les fluctuations quotidiennes des prix de la cryptomonnaie Bitcoin (BTC). Ma démarche s'articule autour de plusieurs étapes clés : la collecte de données historiques, une analyse exploratoire, la mise en place de modèles prédictifs, l'évaluation des performances des modèles et la synthèse finale des résultats.

Objectifs :

L'objectif principal est de prévoir le prix de clôture de BTC pour le lendemain. Les objectifs secondaires incluent :

- Tester différents modèles statistiques et d'apprentissage profond.
- Évaluer les avantages et inconvénients des approches pour des données journalières et horaires.
- Synthétiser les résultats et proposer des améliorations possibles.

Étape 1: Extraction des Données.

J'ai utilisé **yfinance** pour extraire les données historiques de BTC.

yfinance est une bibliothèque Python open-source qui offre une méthode simple pour télécharger des données de marché historiques depuis Yahoo Finance. Elle a été créée par Ran Aroussi pour pallier la dépréciation de l'API Yahoo Finance. yfinance se distingue par sa facilité d'utilisation et ses capacités complètes de récupération de données, ce qui en fait un choix privilégié pour de nombreux membres des communautés financières et de science des données.

Fonctionnalités clés de yfinance :

- **Données de marché historiques :** Les utilisateurs peuvent télécharger les prix historiques des actions, incluant les prix d'ouverture, de clôture, les plus hauts, les plus bas, les prix ajustés, ainsi que les volumes de trading.

- **Actions corporatives :** La bibliothèque fournit des informations sur les actions corporatives, telles que les dividendes et les fractionnements d'actions, essentielles pour une analyse financière précise.
- **États financiers :** yfinance permet d'obtenir des états financiers, incluant les bilans, les comptes de résultat et les flux de trésorerie.
- **Données sur les résultats financiers :** Les utilisateurs ont accès à des informations sur les résultats, y compris les dates de publication, les revenus et le bénéfice par action (BPA).
- **Gestion de plusieurs tickers :** La bibliothèque prend en charge la récupération de données pour plusieurs tickers simultanément, permettant une analyse efficace d'un portefeuille.

J'ai récupéré des données journalières sur une période de trois mois, allant du 13 septembre 2024 à 00:00:00 UTC au 11 décembre 2024 à 23:00:00 UTC, soit un total de 2160 lignes. Pour contourner les limitations de l'API, j'ai mis en place un processus de traitement par "chunks" de temps, chacun couvrant une période de huit jours maximums. Cela permet de répartir les requêtes et d'éviter de dépasser les limites imposées par l'API.

Étape 2 : Nettoyage et Préparation des Données

Etapes principales

- Suppression des valeurs manquantes et des doublons.
- Transformation des dates en format datetime pour permettre une analyse temporelle.
- Agrégation des données journalières avec calcul des prix minimum, maximum, médian et moyen pour chaque cryptomonnaie.
- Stockage des données localement dans un fichier CSV pour une réutilisation ultérieure.

Exemple de structure des données :

```
df = pd.read_csv("daily_btc_price.csv")
df.head(10)
```

✓ 0.0s Open 'df' in Data Wrangler

	Date	Min Close	Max Close	Median Close	Average Close
0	2024-09-12	57506.429688	58391.359375	58046.193359	57994.997884
1	2024-09-13	57774.347656	60638.910156	58131.279297	58753.855794
2	2024-09-14	59736.859375	60457.496094	59996.277344	60032.770182
3	2024-09-15	59188.644531	60352.726562	60058.562500	60009.556478
4	2024-09-16	57609.585938	58909.468750	58384.205078	58324.196126
5	2024-09-17	57647.492188	61137.273438	59167.304688	59381.903646
6	2024-09-18	59439.792969	61516.207031	60214.464844	60172.569499
7	2024-09-19	61940.191406	63768.570312	62804.966797	62676.471517
8	2024-09-20	62695.214844	63837.421875	63177.144531	63236.537272
9	2024-09-21	62864.562500	63389.632812	63141.125000	63113.887533

Dans le cadre de cette étude, le choix entre utiliser des données de clôture quotidienne ou des données horaires dépend des objectifs et des compromis à considérer. Voici une analyse détaillée justifiant pourquoi j'ai choisi de travailler avec des données de clôture quotidienne :

- **Les données de clôture quotidienne :**

  - Pertinence pour la prise de décision :

    L'objectif est d'analyser les tendances à long terme ou de prévoir les prix quotidiens pour cela les données de clôture quotidienne sont suffisent. Cela convient aux tâches comme l'élaboration de stratégies de trading quotidiennes ou l'analyse du comportement général du marché.

  - Lissage du bruit :

    L'agrégation des données horaires en données quotidiennes permet de réduire le bruit et les valeurs aberrantes, qui dominent souvent les données de haute fréquence (horaires). Cela rend les modèles plus clairs et facilite l'identification de tendances et de saisons. Les modèles traditionnels, comme ARIMA ou le lissage exponentiel, fonctionnent très bien avec des données quotidiennes où les fluctuations à haute fréquence sont moins pertinentes.

  - Simplicité computationnelle :

    Les données quotidiennes nécessitent moins de points de données à traiter, ce qui réduit le coût et la complexité computationnelle, notamment pour des périodes longues. Cela permet de simplifier l'analyse tout en obtenant des résultats robustes.

- **Les données horaires :**

  - Trading haute fréquence ou modélisation de la volatilité :

Si l'objectif est de prévoir les mouvements de prix intra journaliers ou la volatilité, les données horaires sont plus pertinentes. Elles sont essentielles pour les stratégies de trading à court terme ou les algorithmes visant à exploiter de petits changements de prix.

  - Capturer les motifs à court terme :

Les motifs intra journaliers, comme les pics matinaux ou les baisses en soirée, ne sont visibles qu'avec des données horaires. Elles permettent d'identifier et de modéliser ces détails fins.

  - Plus de données pour la modélisation :

Les données horaires offrent un ensemble de données plus riche avec plus d'observations, ce qui peut être bénéfique pour les modèles d'apprentissage automatique ou d'apprentissage profond. Cependant, cela peut aussi augmenter le bruit et les exigences computationnelles.

  - Modèles complexes et caractéristiques supplémentaires :

Avec des données horaires, il est possible d'ajouter des caractéristiques supplémentaires comme le volume de transactions, la volatilité intra journalière ou les spreads acheteur-vendeur, ce qui renforce la capacité prédictive des modèles.

- **Compromis**

Aspect	***Données de clôture quotidienne***	***Données horaires***
Bruit	Réduit, tendances plus lisses	Bruit élevé et fluctuations à court terme
Coût computationnel	Plus faible	Plus élevé
Complexité des modèles	Modèles simples (ARIMA, Lissage exponentiel)	Modèles avancés (LSTM, Prophet)
Résolution	Capture les tendances quotidiennes	Capture les motifs intrajournaliers
Cas d'utilisation	Prévisions à long terme, gestion de portefeuille	Trading à court terme, modélisation de la volatilité

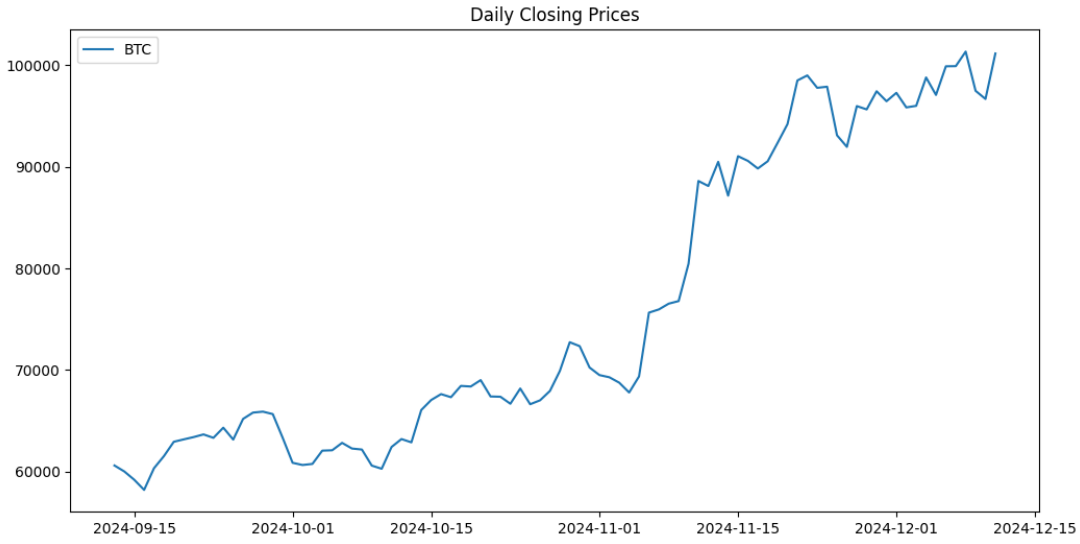
- **Décision :**

J'ai choisi de travailler avec des données journalières dans la mesure où l'objectif est de prédire la valeur de la cryptomonnaie pour le lendemain.

Analyse exploratoire :

- Objectif : Comprendre les tendances, la saisonnalité et les anomalies dans les données.
Étapes :

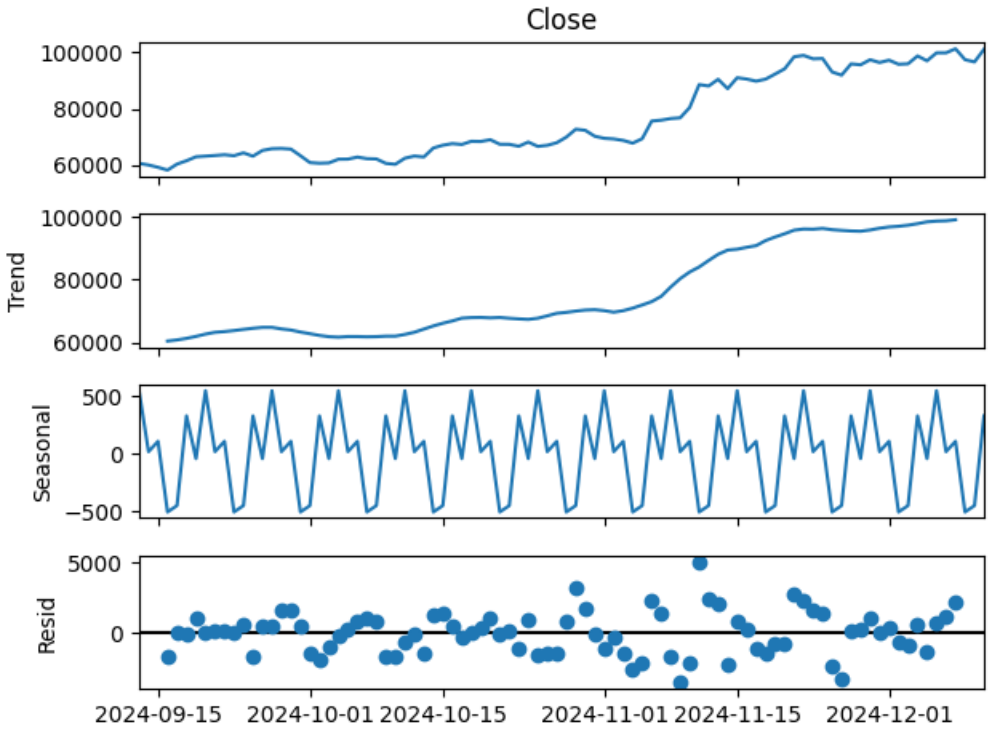
1. Tracer les prix de clôture quotidiens.



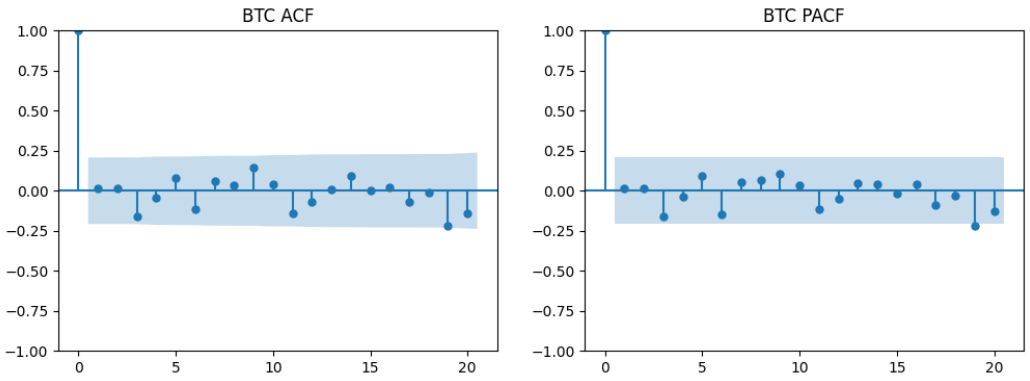
2. Vérifier la stationnarité.

Les résultats du test ADF pour les prix de clôture de Bitcoin montrent une statistique ADF de 0.0955 et une valeur p de 0.9658. La série temporelle des prix de clôture de Bitcoin semble présenter une tendance et/ou une saisonnalité, et n'est pas stationnaire.

3. Décomposer la série temporelle.



- ✓ On observe une tendance croissante
- ✓ La composante saisonnière, représente les fluctuations périodiques autour de la tendance. Les pics et creux dans cette composante suggèrent une répétition régulière de certains patterns sur une base régulière hebdomadaire.
- ✓ Quant aux résidus, les points montrent une dispersion autour de zéro, ce qui indique que le modèle de décomposition explique bien la majorité des variations de la série originale.



- ✓ **ACF (Fonction d'autocorrélation) :**
  - Une décroissance rapide vers 0 après le décalage (lag) 1 suggère que la composante MA (Moyenne Mobile) est minimale.
  - L'absence de pics saisonniers marqués dans le graphique de l'ACF indique que la saisonnalité ne domine probablement pas.
- ✓ **PACF (Fonction d'autocorrélation partielle) :**

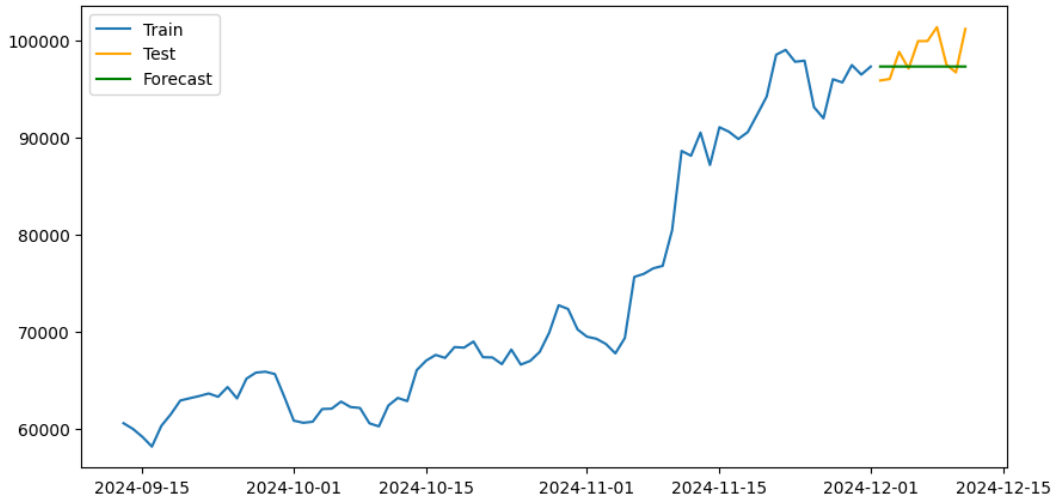
- Un pic important au décalage (lag) 1 dans le graphique du PACF soutient l'inclusion d'un terme AR(1) (Auto-Régressif d'ordre 1).

Etape 3 : proche prédictive

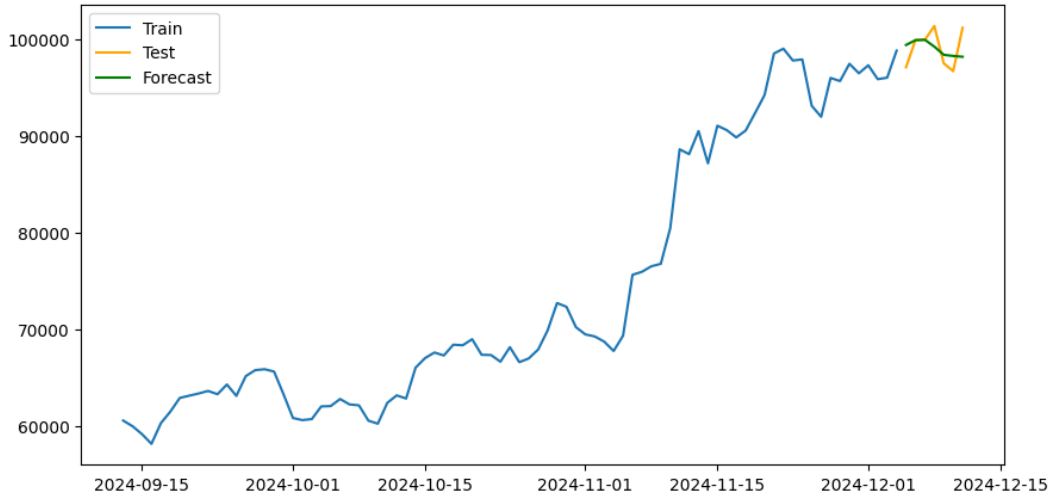
ARIMA:

- ✓ Modèle statistique efficace pour les séries temporelles stationnaires.
- ✓ Nécessite une différenciation pour rendre les données stationnaires.
- ✓ Donne de bons résultats pour les séries à court terme avec une faible saisonnalité.

- ARIMA (1,1,1)



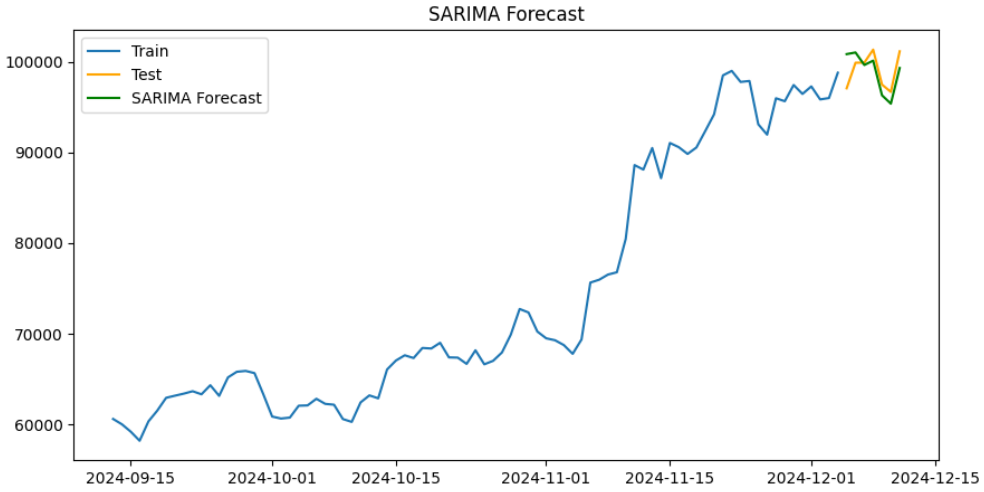
- ARIMA (20,1,7)



Commentaire : la composante saisonnière rend les prédictions du modèle loin d'être parfaites.

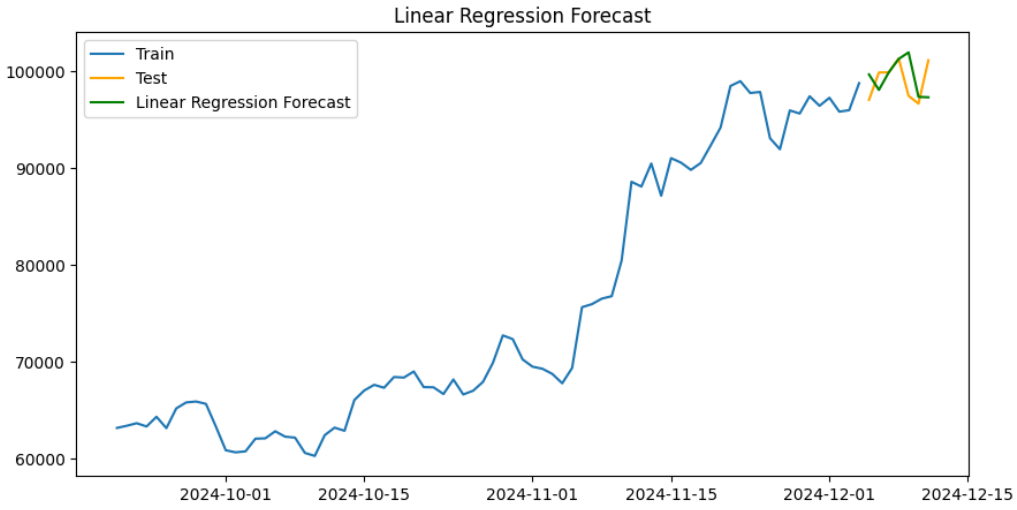
SARIMA :

- ✓ Extension d'ARIMA adaptée aux séries temporelles présentant une saisonnalité.
- ✓ Intègre des variables exogènes afin d'améliorer les prédictions.



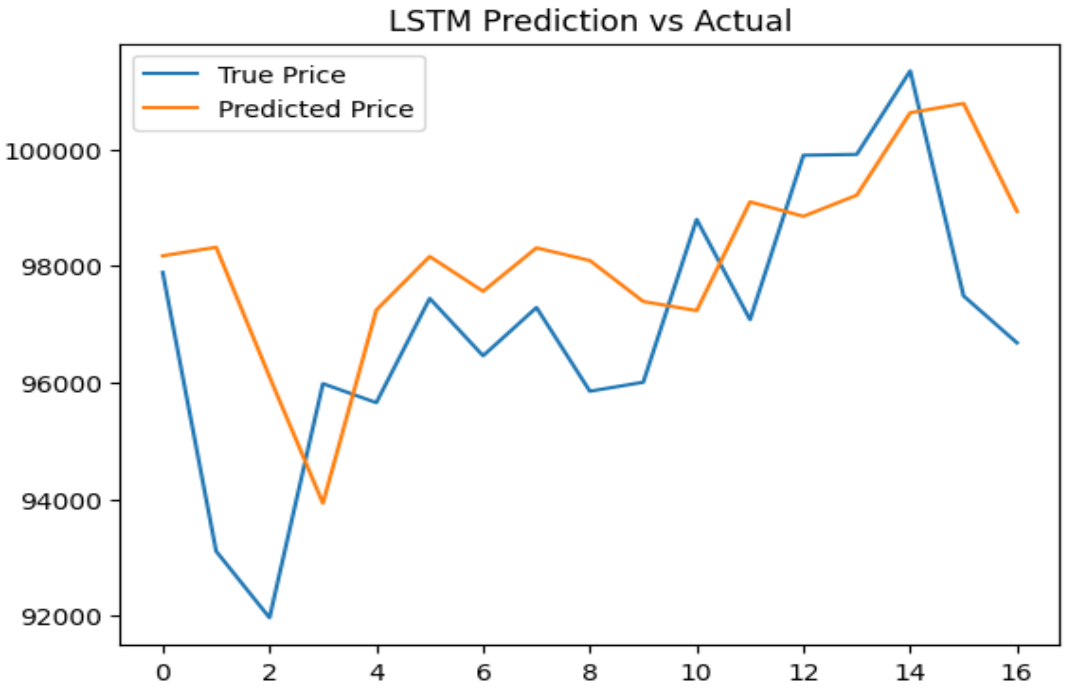
Régression :

- ✓ Approche simple utilisant des variables explicatives dérivées des prix historiques.
- ✓ Convient pour des analyses de base, mais manque de capacité à capturer des schémas complexes.



LTSM :

- ✓ Réseau neuronal récurrent conçu pour capturer les dépendances à long terme.
- ✓ Requiert des ressources significatives en données et calcul.
- ✓ Donne des résultats prometteurs pour les séries temporelles complexes



Choix de l'Approche et points à améliorer :

Les modèles ARIMA et SARIMA sont bien adaptés aux séries stationnaires et saisonnières, mais leur performance a été limitée par la non-stationnarité et le bruit présents dans les données. La régression logistique, bien qu'intéressante pour des relations simples, ne capture pas la dynamique temporelle. LSTM, conçu pour des dépendances complexes, a été pénalisé par un prétraitement insuffisant, le bruit, et des hyperparamètres sous-optimaux. Pour améliorer les résultats, il serait pertinent de décomposer la série en tendance, saisonnalité et résidu (via STL), d'utiliser des modèles hybrides combinant SARIMA et LSTM, et d'enrichir les données avec des variables explicatives tout en ajustant soigneusement les hyperparamètres.