Prédiction des Prix et Actifs du S&P 500 grâce à du Machine Learning simple

Introduction

1.1 Contexte du projet

Le projet de prédiction vise à construire des modèles prédictifs capables d'aider sur les stratégies d'investissement quant aux actions **S&P 500**.

Le **S&P 500** (Standard & Poor's 500) est un indice boursier qui regroupe les 500 plus grandes entreprises cotées en bourse aux États-Unis. Cet indice est souvent utilisé comme un baromètre de l'économie américaine et un indicateur clé des performances du marché boursier global.

Il inclut des entreprises de divers secteurs comme la technologie (Apple, Microsoft), la santé (Johnson & Johnson), la finance (JPMorgan Chase), et bien d'autres. Les investisseurs et analystes s'appuient sur le S&P 500 pour évaluer les tendances du marché, formuler des stratégies d'investissement et mesurer les performances de leurs portefeuilles par rapport à cet indice de référence.

Ainsi, prédire les prix des actions du S&P 500 offre un potentiel stratégique important pour prendre des décisions éclairées dans un environnement économique compétitif.

En s'appuyant sur les données historiques de S&P500 fournies par yahoofinance, mais en rajoutant d'autres données micr-économiques et macro économiques pour enrichir nos analyses et rendre les futurs modèles plus robustes.

1.2 Objectifs principaux

Diverses techniques de **machine learning** peuvent servir à prédire le prix de l'action en question. Les modèles qui vont être mis en place par la suite sont : **régression linéaire**, qui est un modèle de base pas très compliqué mais se distingue par le fait que ses résultats sont facilement interprétables et fournissent plus d'éclairage sur les effets des variables du modèle. Ensuite, vu que la variable dépendante est une série de prix qui évolue selon une échelle temporelle, cela encourage aussi à ajuster un modèle **ARIMA** sur les données historiques. D'autres modèles basés sur les arbres de décision comme **Random Forest** ou **Gradient Boosting** seront testés aussi, et à la fin sera entraîné un modèle de type réseau de neurones **LSTM (Long Short-Term Memory)**.

Cette étude sera menée selon la méthodologie classique adoptée pour les problèmes de machine learning, commençant par une **collecte de données**, suivie d'un **prétraitement**, de la **transformation** et de la **génération de variables**, alignés avec le but final qui est de prédire le **Close price** : il s'agit du prix de clôture de l'action, autrement dit, le prix de l'action en fin de journée de trading.

Les modèles seront entraînés et leurs performances évaluées via des métriques comme le **R^2 score** ou le **Mean Absolute Error (MAE)**. Le meilleur modèle sera sélectionné, et une pipeline résumant les étapes clés du prétraitement sera construite. Des prédictions seront

faites avec le modèle retenu sur des données de test afin de quantifier son pouvoir de généralisation.

Méthodologie

2.1 Préparation des données

Dans le cadre de ce projet, plusieurs APIs ont été utilisées pour collecter les données nécessaires à la prédiction des prix du S&P 500. Ces données proviennent de différentes sources fiables et permettent de capturer à la fois des informations financières et des indicateurs économiques. Les tables créées à partir de ces données contiennent les variables explicatives essentielles pour le modèle prédictif.

Tout d'abord, l'API **Yahoo Finance** a été exploitée pour télécharger les données historiques de l'indice **S&P 500** (`^GSPC`) ainsi que celles de l'indice de volatilité **VIX** (`^VIX`). Ces données couvrent une période de 5 ans et incluent des informations cruciales comme le prix de clôture (`Close`), le prix d'ouverture (`Open`), le prix le plus haut (`High`), le prix le plus bas (`Low`) et le volume des transactions (`Volume`). Ces variables fournissent une base solide pour analyser la dynamique de l'indice S&P500.

Ensuite, l'API **FRED (Federal Reserve Economic Data)** a été utilisée pour accéder à des données macroéconomiques importantes. Les taux directeurs de la Réserve fédérale (**`FedFundsRate`**) ont été collectés pour étudier leur influence sur les mouvements du marché S&P500. De plus, le PIB réel américain (**`GDPC1`**) a été inclus aussi dans l'analyse. Ces données macroéconomiques sont essentielles pour intégrer des facteurs externes aux analyses financières.

Enfin, l'API **Alpha Vantage** a fourni le taux d'inflation, un indicateur économique pertinent pour évaluer les pressions économiques influant les marchés financiers.

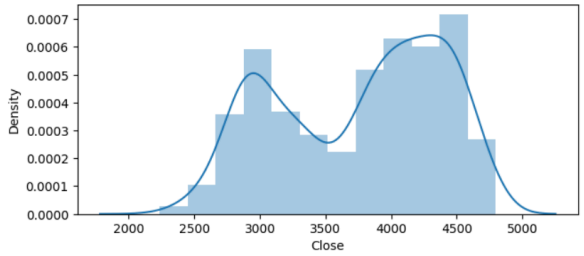
À partir de ces données, plusieurs tables ont été construites et fusionnées par la suite dans une table unique, `data`, combinant à la fois les données financières et économiques. Ce processus a permis de construire un jeu de données complet et cohérent, prêt à être utilisé pour entraîner des modèles prédictifs dans le cadre de notre analyse des prix de clôture du S&P 500.

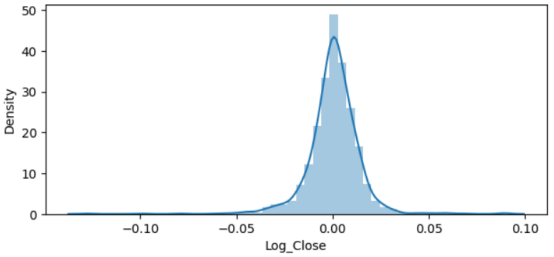
2.2 Transformation des variables

Pour améliorer la qualité des données et maximiser les performances des modèles prédictifs, des transformations et des généralisations de variables ont été effectuées. Ces étapes permettent à la fois de rendre les distributions des données plus adaptées à l'analyse

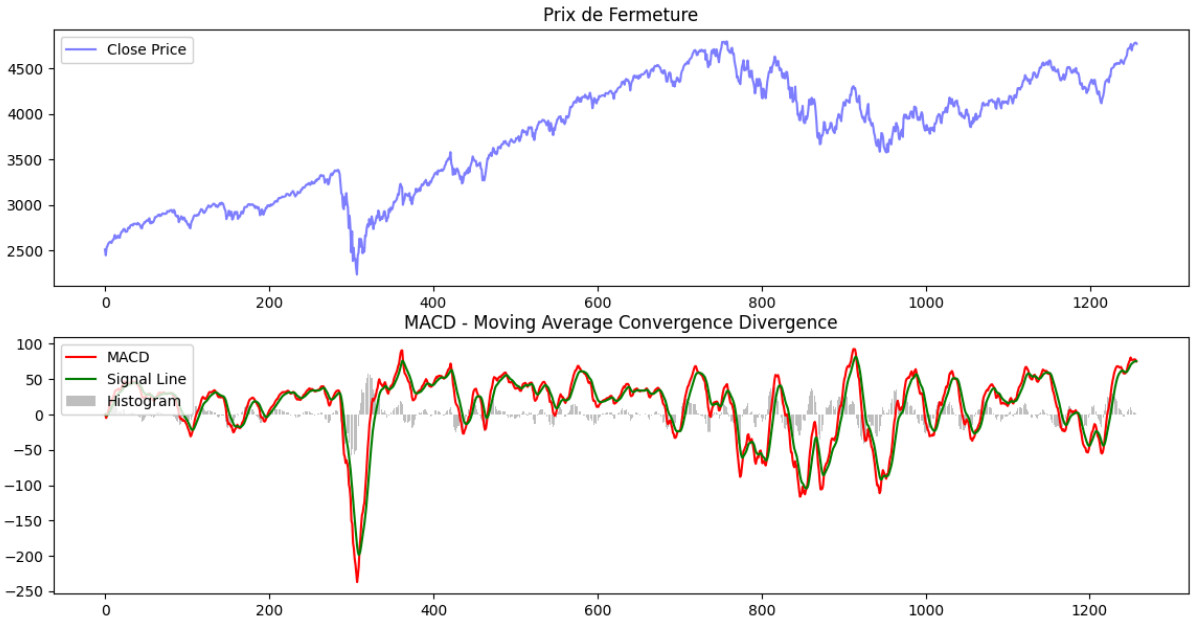
et de fournir des variables explicatives supplémentaires susceptibles de capturer des tendances ou des relations complexes dans les séries temporelles.

Tout d'abord, une transformation logarithmique a été appliquée aux variables de prix telles que `Close`, `High`, `Low` et `Open`. Cette transformation vise à rendre les distributions de ces variables plus symétriques et à stabiliser la variance, ce qui est particulièrement utile lorsque les données présentent une forte asymétrie ou des fluctuations importantes. En travaillant sur les rendements logarithmiques, cette transformation permet également de mieux analyser les variations relatives des prix dans le temps.($r_i = log(\frac{P_t}{P_{t-1}})$)





Des indicateurs techniques avancés ont également été ajoutés pour enrichir les données. Le **Relative Strength Index (RSI)** mesure la force des gains et des pertes récents pour détecter des situations de surachat ou de survente. De même, le **Moving Average Convergence Divergence (MACD)**, basé sur la différence entre deux moyennes mobiles exponentielles, permet de capturer les changements dans la dynamique des prix et d'identifier les retournements de tendance.



2.3 Encodage et normalisation

Pour construire un modèle prédictif réaliste, l'objectif est de prédire `Log_Close`, et les prédictions sont effectuées au début de la journée. Par conséquent, les variables `High`, `Low`, `Open` et leurs transformations logarithmiques sont exclues car elles ne sont pas disponibles au moment de la prédiction. Les colonnes `Date` et `Close` sont également supprimées.

Ensuite, la variable catégorique `Day_of_Week` est encodée en numérique à l'aide de **Label Encoding**, produisant une nouvelle variable `Day_of_Week_Encoded`. Pour garantir une comparabilité entre les variables, la matrice X est standardisée, car les variables présentent des échelles différentes.

Enfin, les données sont divisées en ensembles d'entraînement et de test à l'aide de `train_test_split` avec un ratio de 80%-20%.

Résultats et Analyse

3.1 Régression linéaire

Dans cette étape, une régression linéaire a été utilisée pour analyser les relations entre les variables explicatives et la variable cible `Log_Close`. Le modèle présente un R^2 de 0.177, ce qui indique que seulement 17.7 % de la variance de la variable cible est expliquée par les variables indépendantes. Ce résultat met en évidence une capacité explicative limitée du modèle et suggère qu'il pourrait manquer des facteurs clés influençant les variations de `Log_Close`.

OLS Regression Results

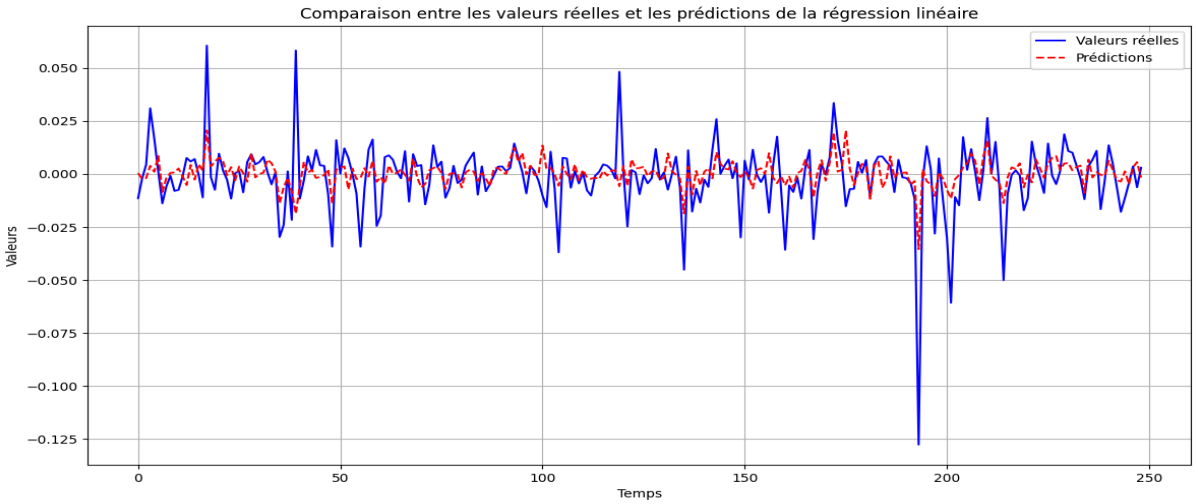
Dep. Variable:	Log_Close	R-squared:	0.177
Model:	OLS	Adj. R-squared:	0.167
Method:	Least Squares	F-statistic:	17.67
Date:	Sun, 22 Dec 2024	Prob (F-statistic):	9.19e-35
Time:	19:27:22	Log-Likelihood:	3030.4
No. Observations:	996	AIC:	-6035.
Df Residuals:	983	BIC:	-5971.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0008	0.000	2.223	0.026	9.6e-05	0.002
Volume	0.0010	0.001	1.940	0.053	-1.16e-05	0.002
VIX	-0.0030	0.001	-4.435	0.000	-0.004	-0.002
FedFundsRate	-0.0007	0.001	-1.174	0.241	-0.002	0.000
RealGDP	-0.0002	0.001	-0.270	0.787	-0.002	0.001
inflation	-0.0005	0.001	-0.723	0.470	-0.002	0.001
Log_RealGDP	0.0013	0.000	3.884	0.000	0.001	0.002
Month	0.0003	0.000	0.805	0.421	-0.000	0.001
Day_of_Month	0.0001	0.000	0.274	0.784	-0.001	0.001
10_Day_Moving_Avg	0.0056	0.001	10.391	0.000	0.005	0.007
RSI	0.0009	0.001	1.572	0.116	-0.000	0.002
MACD	-0.0052	0.001	-8.537	0.000	-0.006	-0.004
Day_of_Week_Encoded	-0.0001	0.000	-0.274	0.784	-0.001	0.001

Omnibus:	302.683	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6276.921
Skew:	0.866	Prob(JB):	0.00
Kurtosis:	15.176	Cond. No.	5.10

Certaines variables apparaissent statistiquement significatives. En revanche, plusieurs variables n'ont pas montré d'impact significatif sur `Log_Close`. Par exemple, `FedFundsRate`, `RealGDP`, et `Day_of_Week_Encoded` ont des p-values respectives de 0.241, 0.787, et 0.784, ce qui indique qu'elles n'expliquent pas de manière notable les variations de la variable cible.

Ce modèle pourrait être amélioré en supprimant les variables non significatives afin de réduire le bruit et d'optimiser les performances.



3.1 Modèle ARIMA

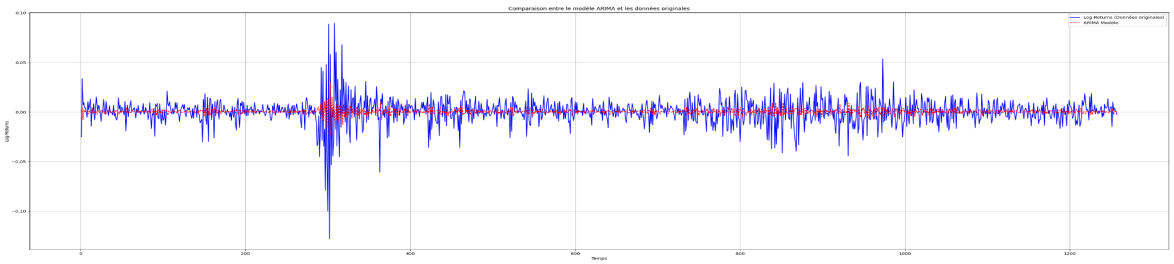
L'analyse ARIMA a été menée pour modéliser la série temporelle de `Log_Close`, en suivant une méthodologie rigoureuse pour déterminer les paramètres optimaux.

La première étape a consisté à vérifier la stationnarité de la série à l'aide du **test ADF (Augmented Dickey-Fuller)**. Les résultats montrent une statistique ADF de **-10.509**, avec une p-value de **1.03e-18**, bien en dessous du seuil de 0.05. Cela indique que la série est stationnaire, rejetant ainsi l'hypothèse nulle. Par conséquent, aucune différenciation supplémentaire (d=0) n'est nécessaire pour modéliser cette série.

Le graphique PACF a été utilisé pour identifier les décalages temporels pertinents. Le lag 1 présente une corrélation très forte, et les lags jusqu'à 10 montrent des corrélations légèrement significatives. Sur cette base, il a été recommandé d'inclure les lags pertinents pour capturer les dépendances temporelles.

Afin d'automatiser la recherche des meilleurs paramètres, l'algorithme **auto_arima** a été appliqué. Celui-ci a testé plusieurs combinaisons de paramètres et a identifié **ARIMA(2, 0, 0)** comme étant le meilleur modèle, basé sur des critères comme l'AIC.

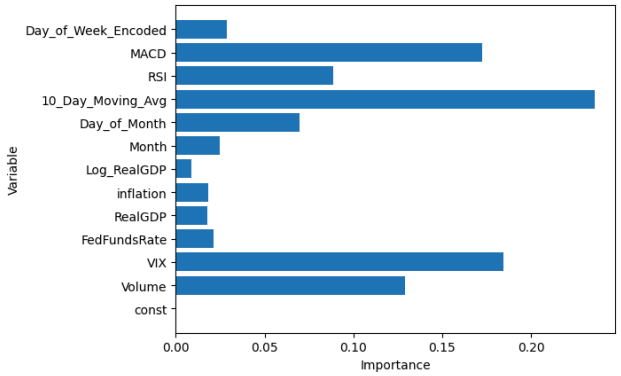
Le modèle ARIMA(2, 0, 0) a ensuite été ajusté et comparé aux données originales. Les résultats montrent une bonne capacité du modèle à suivre les tendances principales des données.



3.2 Random Forest

Pour le modèle Random Forest, un ensemble initial d'entraînement a été utilisé avec 100 estimateurs pour évaluer les performances du modèle. Le **R²** obtenu est de **0.136**, ce qui indique une capacité explicative relativement faible pour ce modèle sur les données testées. Cela signifie que le modèle explique seulement 13.6% de la variance de la variable cible.

Pour optimiser les hyperparamètres, une recherche en grille (**GridSearchCV**) a été appliquée pour trouver la meilleure valeur du nombre d'estimateurs (`n_estimators`). Les tests ont montré que le meilleur paramètre est **n_estimators = 55**, avec un score moyen de validation croisée (R2) de **0.057**. Cependant, ce score reste faible, suggérant que le modèle Random Forest n'est pas optimal dans ce contexte, potentiellement en raison d'un manque de relations non linéaires suffisamment fortes ou de la faible importance explicative de certaines variables.



Une analyse des importances des variables montre que certaines, comme `10_Day_Moving_Avg` et `VIX`, contribuent significativement au modèle, tandis que d'autres, comme `Month` ou `FedFundsRate`, ont un impact négligeable. Cela met en évidence l'importance de sélectionner ou d'ingénier les variables explicatives pertinentes pour améliorer les performances du modèle.

En conclusion, bien que le modèle Random Forest fournisse des informations utiles sur l'importance des variables, ses performances globales en prédiction restent limitées.

3.3 Gradient Boosting

Pour le modèle Gradient Boosting, un premier entraînement avec 100 estimateurs a été effectué. Cependant, le **R² obtenu sur le jeu de test est négatif (-0.025)**, ce qui signifie que le modèle performe moins bien qu'une simple moyenne des valeurs cibles. Cela indique que le modèle n'est pas adapté aux données actuelles ou que les relations non linéaires qu'il est censé capturer sont mal représentées.

3.4 LSTM

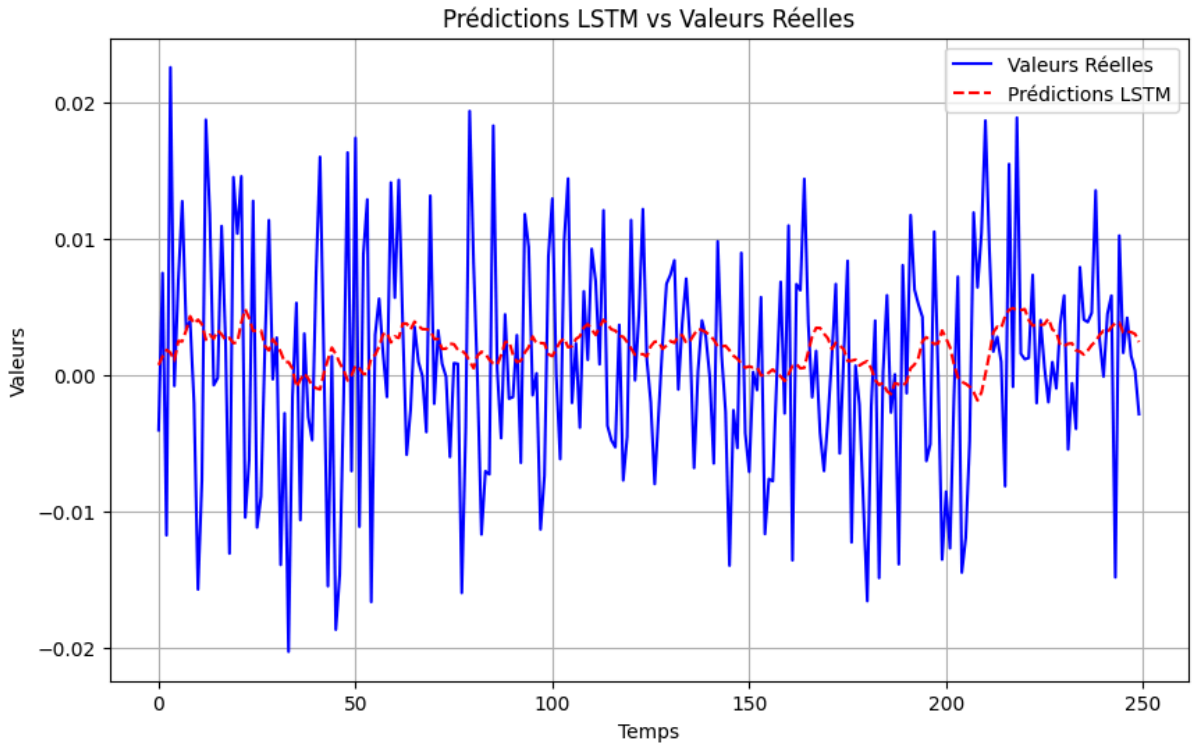
Pour le modèle **LSTM** (Long Short-Term Memory), les données ont été normalisées à l'aide de **MinMaxScaler** pour les transformer entre 0 et 1. Cette étape est essentielle pour garantir une convergence optimale de l'entraînement du modèle. Ensuite, des séquences

temporelles ont été créées avec une longueur fixe de 10 jours pour capturer les dépendances à court terme dans les données de `Log_Close`.

Le modèle LSTM a été construit avec une architecture simple comprenant une couche **LSTM** de 50 unités suivie d'une couche dense pour la prédiction. Le modèle a été compilé avec l'optimiseur **Adam** et la fonction de perte **Mean Squared Error (MSE)**. L'entraînement s'est déroulé sur 20 époques avec un batch size de 16, en utilisant une validation sur l'ensemble de test pour surveiller la performance.

Après l'entraînement, les prédictions ont été effectuées sur l'ensemble de test. Les résultats montrent que les prédictions capturent correctement les tendances globales, bien que les fluctuations rapides soient moins bien représentées. Cela peut être dû à la structure simplifiée du modèle ou aux limitations des données disponibles.

Enfin, les valeurs prédites et réelles ont été inversées pour revenir à leur échelle originale et ont été comparées visuellement. La visualisation montre que le modèle suit les tendances des prix, ce qui est conforme à l'objectif de capturer les dynamiques principales de la série étudiée. Cependant, des améliorations, telles que l'ajout de couches supplémentaires ou l'ajustement des hyperparamètres, pourraient encore améliorer la précision des prédictions.



Conclusion et Recommandations

4.1 Synthèse des performances des modèles

Au cours de ce projet, plusieurs modèles ont été testés pour prédire le prix logarithmique de clôture (`Log_Close`) de l'indice S&P 500, chacun offrant des perspectives uniques sur les dynamiques des données financières.

La **régression linéaire** a permis une interprétation claire des variables explicatives, avec un R^2 de 0.177. Ce résultat montre une capacité limitée à expliquer la variance de la target, mais il a révélé des relations significatives entre certaines variables, telles que `VIX` et `Log_RealGDP`, et les fluctuations des prix.

Le modèle **ARIMA**, identifié comme ARIMA(2, 0, 0) via `auto_arima`, a capturé efficacement les dépendances temporelles à court terme. Bien que ses prédictions suivent les tendances principales des données historiques, il reste limité pour les variations complexes et rapides.

Les modèles basés sur les arbres, comme le **Random Forest** et le **Gradient Boosting**, n'ont pas réussi à produire des résultats satisfaisants. Le Random Forest a obtenu un R^2 de 0.136, tandis que le Gradient Boosting a montré des performances négatives sur le jeu de test avant optimisation. Ces résultats suggèrent que ces modèles nécessitent des variables plus pertinente autres que celles disponibles.

Enfin, le modèle **LSTM** a montré de bons résultats pour capturer les tendances à long terme dans les données. Bien qu'il nécessite un temps d'entraînement plus long, il a réussi à modéliser efficacement les dépendances temporelles.

4.2 Recommandations pour les analyses futures

Les modèles prédictifs développés dans ce projet offrent des perspectives intéressantes pour orienter des stratégies d'investissement basées sur l'analyse des données financières. En particulier, l'utilisation du **VIX** comme indicateur principal permet d'identifier des périodes de forte volatilité, offrant ainsi la possibilité de détecter des risques potentiels ou des opportunités d'arbitrage sur les marchés.

Par ailleurs, les prédictions issues des modèles **LSTM** et **ARIMA** peuvent être exploitées pour identifier les tendances des prix à long terme. Ces informations permettent d'optimiser les décisions d'achat ou de vente en fonction des fluctuations anticipées, renforçant ainsi la capacité à prendre des décisions éclairées dans un environnement de marché complexe.