



Rapport d'Analyse

Étude de Cas pour Stage Data Analyst/Scientist chez ID&A TECH

sous le thème

Prédiction de la Performance de l'Indice MASI

Réalisé par : Abdallahi DAH

Année Universitaire : 2023- 2024

Tables de Matières :

Introduction.....	3
2. Exploration Approfondie des Données (EDA).....	5
Exploration de l'Indice Général :.....	5
Exploration de Avoirs officiels de réserve :.....	6
Exploration des Autres Feuilles de Données :.....	8
3. Prétraitement des Données.....	9
3.1. Sélection de Variables.....	9
3.2. Traitement des Valeurs Manquantes.....	10
3.3. Fusion des Données pour Modélisation Exogène.....	10
3.5. Préparation des Données pour les Modèles de Série Temporelle.....	11
3.6. Standardisation des Données.....	12
4. Modélisation de Série Temporelle :.....	13
4.1. Modèles d'Apprentissage Automatique (ML) - RF et XGB :.....	13
4.2. Modèle Statistique ARIMA :.....	14
4.3. Modèle de Deep Learning - LSTM :.....	16
6. Modélisation avec Variables Explicatives :.....	18
6.1. Modélisation avec Machine Learning (ML) - RF et XGB :.....	18
6.2. Modélisation avec Deep Learning (DL) - LSTM :.....	19
7. Optimisation et Évaluation des Modèles :.....	20
7.1 Optimisation des modèles ML : Grid Search.....	20
7.2 Optimisation des modèles DL :.....	21
7.3 Évaluation des Modèles :.....	21
8. Résultats et Analyses :.....	22
8.1 Résultats sur la modélisation de Série Temporelle :.....	22
8.2 Résultats Modélisation avec Variables Explicatives.....	23
8.3 les facteurs les plus importants:.....	25
Conclusion Générale :.....	27



Introduction

L'environnement financier mondial évolue rapidement, marqué par des changements économiques, politiques et sociaux. Dans ce contexte, la capacité à anticiper et comprendre les tendances du marché financier revêt une importance cruciale pour les investisseurs et les décideurs financiers. Notre étude, centrée sur le marché financier marocain, se concentre sur la prédiction de la performance de l'Indice MASI (Moroccan All Shares Index), qui représente la performance globale du marché boursier au Maroc.

Contexte

Le marché financier est un écosystème complexe influencé par divers facteurs économiques et sectoriels. L'analyse de ces facteurs est cruciale pour anticiper les tendances et prendre des décisions éclairées. Dans ce contexte, notre étude s'est concentrée sur l'Indice MASI, une référence clé pour évaluer la performance boursière au Maroc.

Objectifs

L'objectif global de cette étude était de fournir des insights exploitables pour les décideurs financiers en utilisant des techniques avancées d'analyse de données. Notre objectif est double :

- Premièrement, nous cherchons à développer des modèles de prédiction robustes pour anticiper la performance de l'Indice MASI en se basant uniquement sur sa variation dans le temps.
- Deuxièmement, nous visons à identifier les facteurs clés influençant les rendements de cet indice. En intégrant d'autres variables exogènes comme variables explicatives représentées par une variété de données financières, macroéconomiques et sectorielles

Méthodologie Adoptée

Pour atteindre nos objectifs, nous avons suivi une méthodologie rigoureuse articulée autour de plusieurs étapes clés :

- Exploration approfondie des Données (EDA) : Analyse minutieuse des indices généraux, des avoirs officiels de réserve, et d'autres feuilles de données pour comprendre les tendances, les corrélations et les anomalies propres au marché boursier marocain.
- Prétraitement des Données : Nettoyage, sélection des variables pertinentes, traitement des valeurs manquantes et préparation des données pour la modélisation.
- Modélisation de Série Temporelle : Application de techniques de série temporelle pour anticiper la performance de l'Indice MASI en se basant uniquement sur le cours de clôture.
- Modélisation avec Variables Explicatives : Intégration de variables exogènes telles que les données macroéconomiques et sectorielles pour améliorer la prédiction et identifier les facteurs clés.
- Optimisation et évaluation des modèles : Réglage des hyperparamètres, validation croisée, et évaluation des performances des modèles.

Les résultats obtenus à travers ces étapes contribuent à notre compréhension du marché financier marocain, offrant des bases solides pour des décisions d'investissement éclairées.

Dans les sections suivantes, nous détaillerons chacune de ces étapes, mettant en lumière nos choix méthodologiques, les résultats obtenus et leurs implications pour le marché boursier marocain.

2. Exploration Approfondie des Données (EDA)

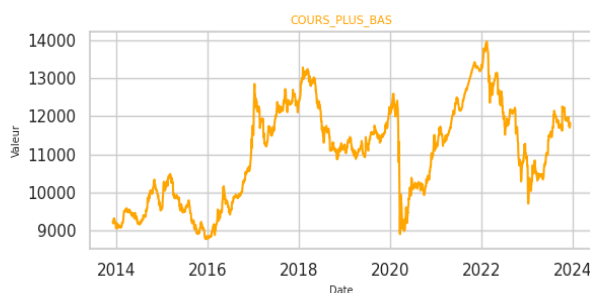
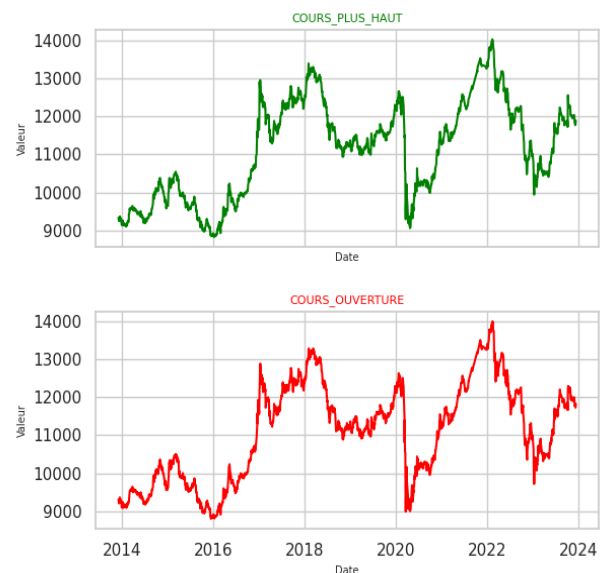
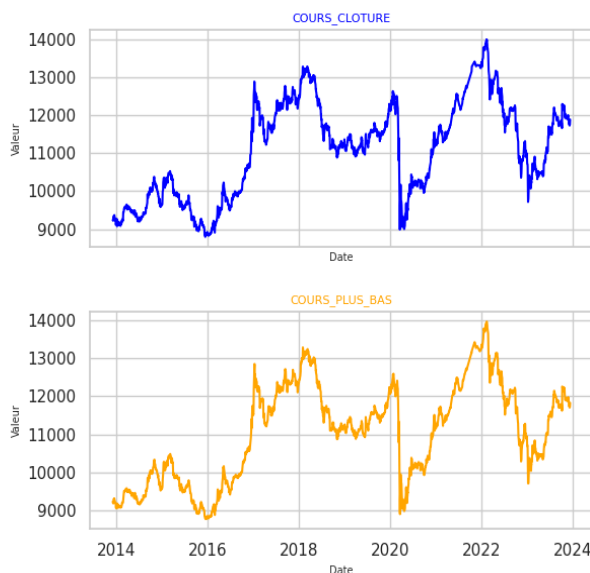
L'Exploration Approfondie des Données (EDA) constitue une étape préliminaire essentielle, nous permettant de forger des hypothèses initiales et de développer une compréhension approfondie de la structure de nos données et les relations entre les variables. Nous allons explorer chaque ensemble de données individuellement, en visualisant les variables présentes et en examinant leur évolution dans le temps.

En mettant en œuvre des techniques de visualisation et en calculant les corrélations entre les variables, nous cherchons à identifier des tendances, des modèles saisonniers et des relations significatives. Ces découvertes informeront nos choix ultérieurs en matière de sélection de variables, de prétraitement des données et de modélisation.

Exploration de l'Indice Général :

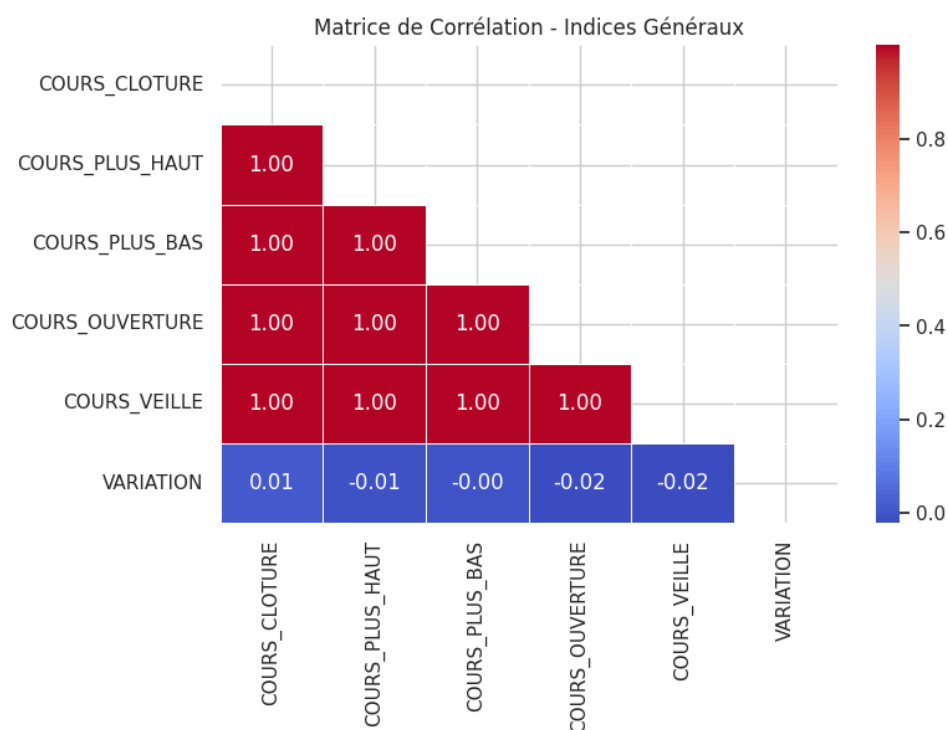
Dans cette section, nous avons procédé à une exploration approfondie des données en examinant l'évolution temporelle des prix de l'indice MASI. Les figures ci-dessous présentent les variations au fil du temps pour les différentes mesures, notamment COURS_CLOTURE, COURS_PLUS_HAUT, COURS_PLUS_BAS et COURS_OUVERTURE.

Cours des Indices en fonction du temps



F

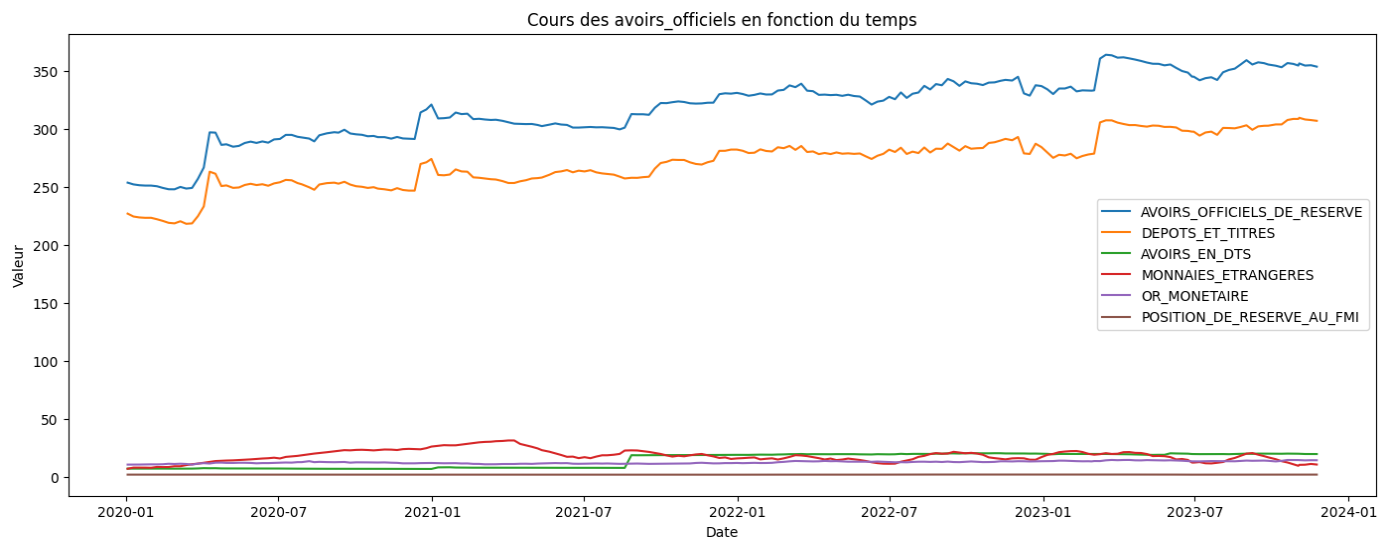
La matrice de corrélation, que vous trouverez également ci-dessous, met en lumière une forte corrélation entre ces variables. Cette corrélation étroite entre les différentes mesures nous conduit à prendre une décision stratégique pour la suite de notre analyse : afin de simplifier et d'optimiser notre modèle, nous choisirons de travailler uniquement avec la variable COURS_CLOTURE dans les étapes suivantes.



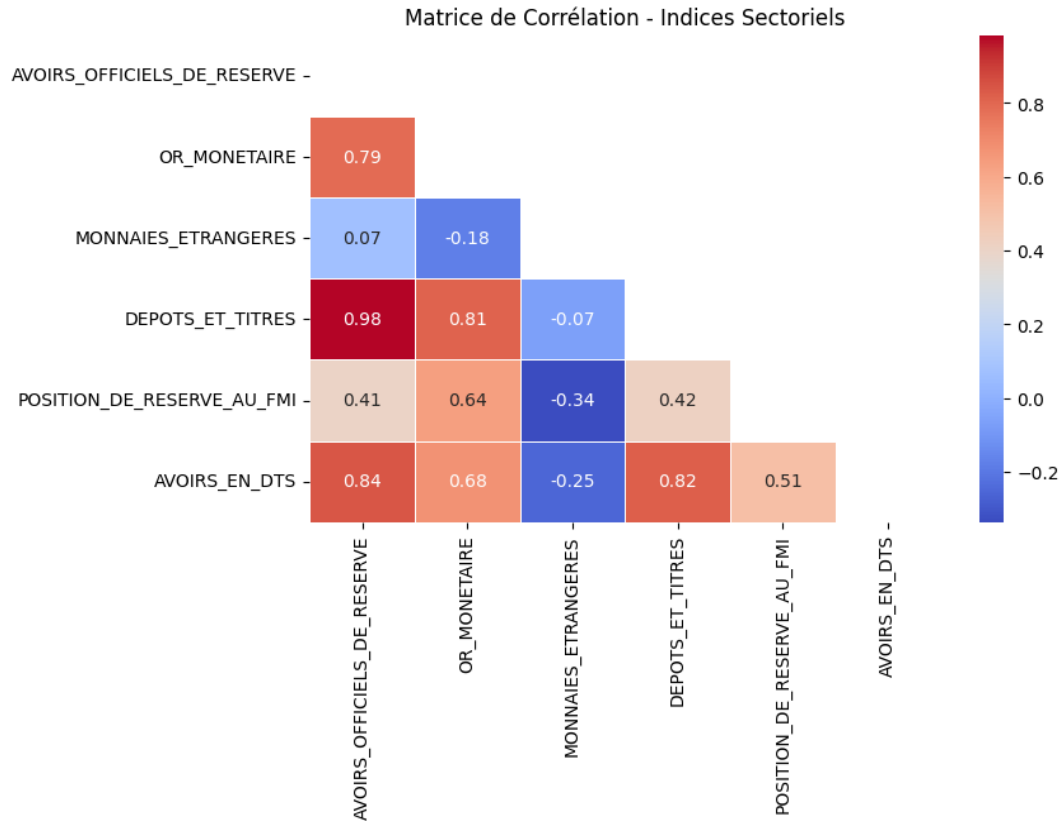
Exploration de Avoirs officiels de réserve :

Dans cette étape, nous avons effectué une exploration détaillée des données relatives aux Avoirs Officiels de Réserve, en mettant particulièrement l'accent sur les variables DEPOTS_ET_TITRE et AVOIR_OFFICE_DE_RESERVE. La table de statistiques descriptives ci-dessous offre un aperçu quantitatif de ces données, tandis que les figures suivantes présentent l'évolution temporelle de ces variables.

	AVOIRS_OFFICIELS_DE_RESERVE	OR_MONETAIRE	MONNAIES_ETRANGERES	DEPOTS_ET_TITRES	POSITION_DE_RESERVE_AU_FMI	AVOIRS_EN_DTS
count	207.000000	207.000000	207.000000	207.000000	207.000000	207.000000
mean	318.721258	12.583768	18.090473	271.568289	1.957151	14.527534
std	28.083077	1.070565	5.140158	22.118081	0.058049	6.034526
min	247.751000	10.560000	6.992000	218.044000	1.871077	6.852960
25%	300.747593	11.672304	15.006020	255.733000	1.903109	7.510500
50%	324.153140	12.590958	17.863413	274.850257	1.957449	18.945310
75%	338.558395	13.455218	21.087087	284.093164	1.994542	19.695408
max	363.692490	14.600335	31.386899	309.361444	2.076412	20.403429



La matrice de corrélation, présentée ci-dessous, révèle une corrélation significative entre les variables DEPOTS_ET_TITRE et AVOIR_OFFICE_DE_RESERVE. Cette observation nous conduit à la décision stratégique de ne conserver qu'une seule de ces variables dans la suite de notre analyse, afin de simplifier notre modèle tout en préservant la représentation essentielle de ces données.



Exploration des Autres Feuilles de Données :

Pour maintenir une concision appropriée dans ce rapport d'analyse, nous avons décidé de présenter deux exemples représentatifs du processus d'exploration de données que nous avons suivi. Cependant, il est important de souligner que la même méthodologie a été appliquée à chaque feuille de données, y compris TMP, l'Inflation, les Indices Sectoriels, etc.

Pour chaque feuille, nous avons commencé par visualiser les variations temporelles des variables pertinentes, fournissant une compréhension visuelle de l'évolution de ces données au fil du temps. Par la suite, nous avons examiné les matrices de corrélation pour identifier les relations significatives entre les variables.

Toutes ces analyses détaillées, y compris les graphiques, les statistiques descriptives et les matrices de corrélation, sont disponibles dans le Notebook Jupyter associé à ce rapport.

3. Prétraitement des Données

Dans cette section, nous nous attaquons au prétraitement de chaque feuille de données, visant à garantir la qualité et la pertinence des informations utilisées dans nos modèles. Chaque sous-section est dédiée à une feuille spécifique, mettant en œuvre des stratégies de prétraitement adaptées à la nature des données.

3.1. Sélection de Variables

La sélection de variables revêt une importance cruciale pour réduire la redondance et se concentrer sur les informations les plus pertinentes. En se basant sur les matrices de corrélation de chaque dataframe, nous avons choisi de conserver une seule variable parmi celles fortement corrélées. De plus, nous avons standardisé le nom des variables, en renommant la variable de date "DATE" dans toutes les Dataframes. Voici un aperçu des choix effectués pour chaque dataframe :

- *indices_generaux_cleaned* : Nous avons conservé uniquement la variable "COURS_CLOTURE" avec "DATE".
- *avoirs_officiels* : Nous avons supprimé la variable "DEPOTS_ET_TITRES" et conservé les autres variables avec "DATE".
- *change* : Nous avons conservé les variables "Change_Minimum" et "Change_Maximum" avec "DATE".
- *monia* : Nous avons conservé les variables "val_Indice_MONIA" et "vol_Volume_JJ" avec "DATE".
- *tmp* : Nous avons conservé uniquement la variable "TMP" avec "DATE".
- *inflation* : Nous avons gardé uniquement la variable "INFLATION" avec "DATE".
- *indices_taux* : Nous avons conservé les variables "NOMINAL 1 AN", "NOMINAL 2 ANS", "NOMINAL 3 ANS", "NOMINAL 10 ANS" et "NOMINAL 20 ANS" avec "DATE".
- *indices_sectoriels_cleaned* : Nous avons éliminé les variables fortement corrélées et conservé plusieurs autres, notamment "BANQUES", "BOISSONS", "CHIMIE", "DISTRIBUTEURS", "ELECTRICITE", etc., avec "DATE".

Ces choix stratégiques dans la sélection de variables nous fourniront des ensembles de données plus précis et éviteront les redondances inutiles.

3.2. Traitement des Valeurs Manquantes

Après avoir effectué la sélection de variables, nous avons examiné la présence de valeurs manquantes dans chaque dataframe. En général, la majorité des dataframes ne présentait aucune lacune, à l'exception de la dataframe `indices_taux_cleaned` qui contenait seulement quatre valeurs manquantes.

Conscients de l'importance de maintenir l'intégrité des données, nous avons opté pour une approche pragmatique en supprimant ces quelques lignes. Cette décision découle du faible nombre de valeurs manquantes par rapport à la taille totale des données, minimisant ainsi l'impact sur la qualité globale des informations.

3.3. Fusion des Données pour Modélisation Exogène

Afin de préparer notre jeu de données pour la modélisation prenant en compte des variables exogènes, nous avons fusionné la dataframe `indices_generaux_cleaned` avec d'autres dataframes, qu'elles soient d'ordre macroéconomique ou sectoriel. Cependant, il est important de noter que les données de certaines variables, telles que `indices_taux_cleaned`, `inflation_cleaned`, `tmp_cleaned`, et `monia_cleaned`, ne sont disponibles que depuis 2021 ou à partir du 22 décembre 2020.

Pour garantir la cohérence des données, nous avons décidé de conserver uniquement les enregistrements de notre variable cible à partir de l'année 2021. Par la suite, nous avons procédé à la fusion des dataframes en utilisant la fonction `merge_asof()` de la bibliothèque Pandas. Cette méthode a permis d'associer les dates aux dates les plus proches.

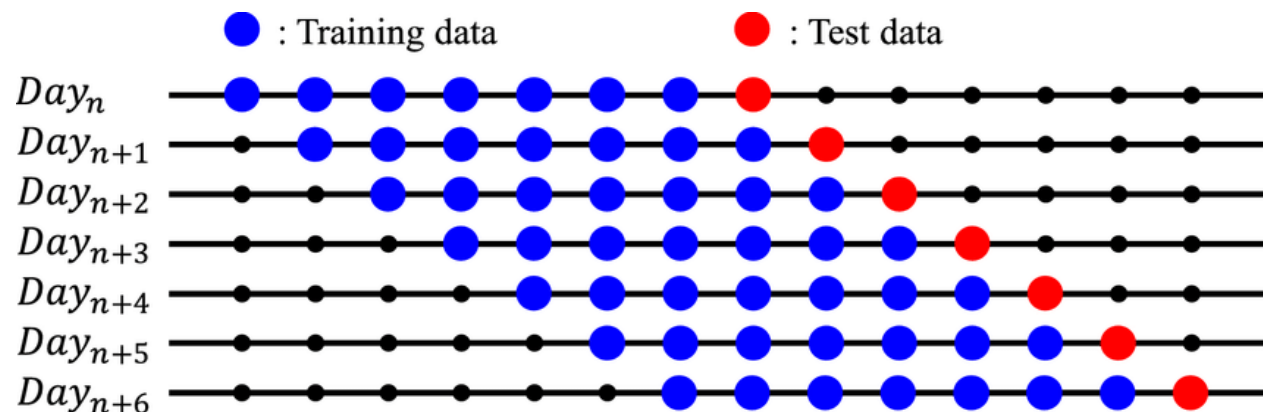
Il est à noter que les données d'inflation ne sont disponibles qu'à partir de 2022. Pour combler ces valeurs manquantes, nous avons opté pour une valeur de remplacement réaliste, en utilisant le taux d'inflation moyen au Maroc en 2021, soit 1.4%.

Suite à cette fusion, nous avons examiné les données résultantes et constaté, grâce à la matrice de corrélation, une forte corrélation entre notre variable cible, représentant les prix de clôture des cours, et certaines variables sectorielles telles que l'indice de BANQUE et l'indice de sociétés financières, ce qui est cohérent du point de vue économique.

3.5. Préparation des Données pour les Modèles de Série Temporelle

Dans cette section, nous avons adopté une approche spécifique pour préparer nos données en vue de l'application de modèles de série temporelle. Le choix de cette méthode est justifié par le caractère séquentiel et temporel des données, en particulier pour la variable cible représentant les prix de clôture des cours de l'indice MASI.

Nous avons décidé de ne conserver que la variation de notre variable cible au fil du temps. Pour ce faire, nous avons appliqué la technique de la fenêtre glissante (sliding window) avec une fenêtre de taille 7. Concrètement, cela signifie que pour chaque point temporel, nous avons considéré les six points précédents pour définir la variation actuelle. Cette méthode est particulièrement pertinente car elle permet de capturer les tendances et les motifs à court terme tout en maintenant une certaine réactivité aux changements.



L'application de la fenêtre glissante présente l'avantage d'offrir une vision dynamique de l'évolution de la variable cible, permettant aux modèles de capturer les variations fréquentes et potentiellement de mieux s'adapter aux fluctuations du marché.

Cette approche offre une manière robuste de traiter les séries temporelles en mettant l'accent sur les variations récentes pour prédire les tendances futures. Elle peut également aider à atténuer l'impact des valeurs aberrantes ou des fluctuations temporaires, mettant davantage l'accent sur la tendance générale du marché.

3.6. Standardisation des Données

Afin d'assurer la cohérence et la stabilité des résultats de nos modèles, nous avons appliqué une étape cruciale de standardisation des données. Cette démarche s'est avérée essentielle compte tenu de la diversité des échelles entre nos variables explicatives et de la variable cible, en particulier.

Pour les variables explicatives, nous avons opté pour la méthode de Min-Max Scaling en utilisant le `MinMaxScaler` de la bibliothèque `scikit-learn`. Cette technique permet de ramener toutes les valeurs à une plage commune entre 0 et 1, préservant ainsi les relations de proportion entre les différentes caractéristiques.

Quant à notre variable cible, représentant les prix de clôture des cours de l'indice MASI, nous avons constaté des valeurs numériques significatives avec un ordre de grandeur relativement élevé (ex. : 10000, 11000, etc.). Pour homogénéiser ces valeurs, nous avons appliqué une normalisation en utilisant la formule standard :

$$y_{scaled} = \frac{y - \min(y)}{\max(y) - \min(y)}$$

Cela a permis de ramener nos valeurs de la variable cible à une échelle normalisée entre 0 et 1. Il est important de noter que cette normalisation sera inversée lors de l'évaluation des performances des modèles, ramenant ainsi les prédictions à l'échelle d'origine pour une interprétation plus intuitive.

4. Modélisation de Série Temporelle :

Pour appréhender la dynamique temporelle des données, nous avons mis en œuvre trois approches distinctes : les modèles de Machine Learning (ML) avec Random Forest (RF) et XGBoost (XGB), l'approche statistique avec ARIMA, et enfin, l'utilisation de réseaux de neurones profonds avec Long Short-Term Memory (LSTM).

4.1. Modèles d'Apprentissage Automatique (ML) - RF et XGB :

Dans cette phase, nous avons mis en œuvre deux puissants modèles d'apprentissage automatique, à savoir le Random Forest (RF) et le XGBoost (XGB), pour modéliser la série temporelle des variations des prix de clôture de l'indice MASI.

Après un processus approfondi de prétraitement des données, nous avons soigneusement divisé notre ensemble de données en ensembles d'entraînement (x_{train_st} , y_{train_st}) et de test (x_{test_st}). Ces modèles ont été ensuite entraînés sur l'ensemble d'entraînement pour comprendre les tendances et les motifs temporels.



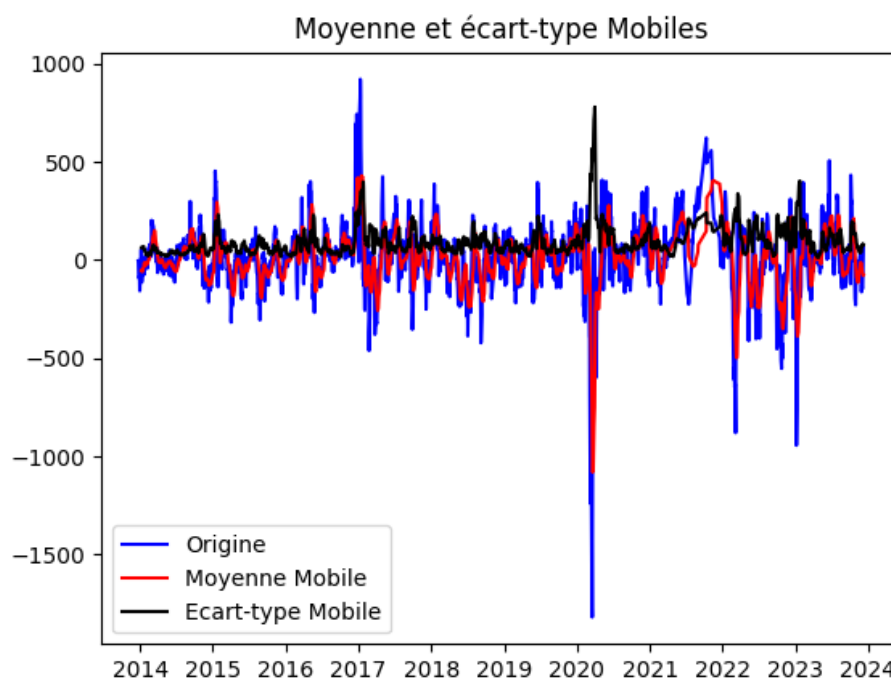
Le graphe ci-dessus montre une comparaison entre les prédictions générées par les modèles ont aux données réelles du jeu de test, offrant ainsi une visualisation graphique de la performance. Ces visualisations nous permettront de tirer des conclusions préliminaires sur la performance des modèles

Il est important de noter que l'optimisation des paramètres de ces modèles sera détaillée dans la section de l'optimisation et l'évaluation des modèles.

4.2. Modèle Statistique ARIMA :

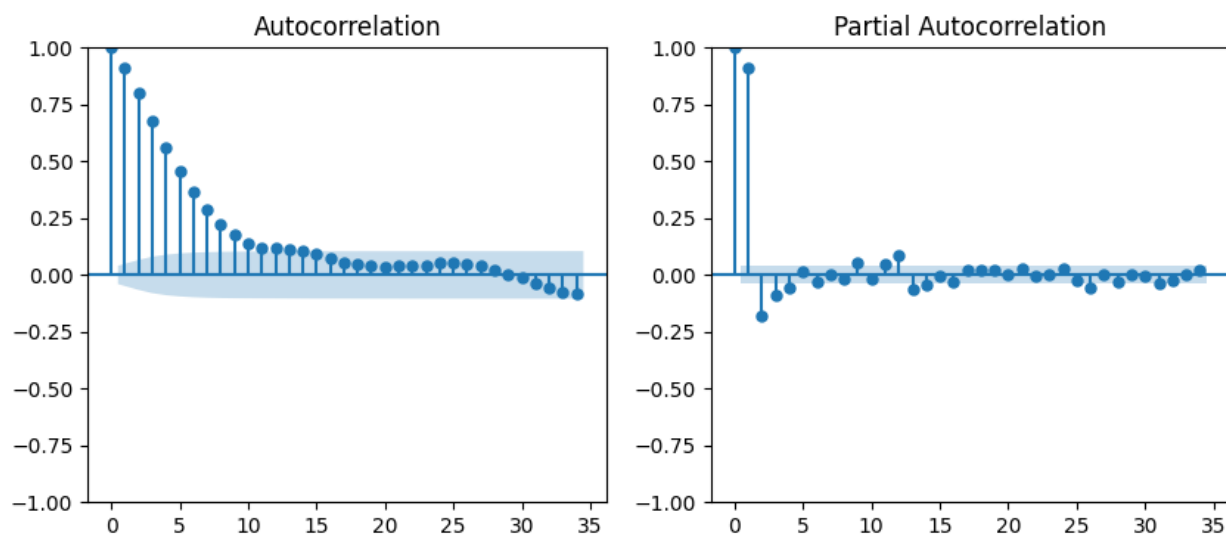
Dans cette phase, nous avons exploré l'utilisation d'un modèle statistique AutoRegressive Integrated Moving Average (ARIMA) pour capturer les motifs temporels dans la série chronologique des prix de clôture de l'indice MASI.

Avant d'appliquer le modèle ARIMA, nous avons évalué la stationnarité de la série chronologique à l'aide du test Dickey–Fuller. Ensuite, nous avons entrepris une transformation pour rendre la série stationnaire en éliminant les tendances et les saisons. Le résultat de cette transformation est illustrés dans le graphique suivant qui montre la série chronologique stationnaire avec les résultats de test de Dickey–Fuller après la stationnarité de données.



Statistiques ADF	-9.758105634726261
p-value	7.692141963652814e-17
Valeurs Critiques	1%: -3.433146448071043 5%: -2.8627754190239156 10%: -2.5674276676268972

Les fonctions d'autocorrélation (ACF) et de corrélation partielle (PACF) ont été analysées pour déterminer les ordres optimaux p et q du modèle ARIMA. Le graphique ci-dessous montre la visualisation de ces analyses qui montre une chute après le deuxième pic de courbe de corrélation partielle ce qui est une signe de stationnarité de notre série.



En se basant sur ces résultats nous avons décidé de choisir les ordre de notre modèle ARIMA comme le suivant $p=3$, $q= 3$ et un une différenciation qui est égale a 0. Le modèle a été ajusté avec les paramètres optimaux. Voici une résumé sur le modèle ARIMA fitted :


```

=====
SARIMAX Results
=====
Dep. Variable:    COURS_CLOTURE    No. Observations:    2123
Model:           ARIMA(4, 0, 3)    Log Likelihood       -12040.275
Date:            Sun, 17 Dec 2023    AIC                  24098.550
Time:            18:47:18            BIC                  24149.495
Sample:          0                  HQIC                 24117.200
                    - 2123
Covariance Type:    opg
=====
              coef    std err          z      P>|z|      [0.025    0.975]
-----
const          3.9876     14.647      0.272    0.785    -24.720     32.696
ar.L1           0.1171      0.169      0.693    0.488     -0.214      0.448
ar.L2           0.9491      0.161      5.908    0.000      0.634      1.264
ar.L3           0.0590      0.144      0.409    0.682     -0.224      0.342
ar.L4          -0.3222      0.127     -2.538    0.011     -0.571     -0.073
ma.L1           0.9394      0.171      5.503    0.000      0.605      1.274
ma.L2          -0.0454      0.095     -0.480    0.631     -0.231      0.140
ma.L3          -0.2094      0.118     -1.777    0.076     -0.440      0.022
sigma2         5017.2203     53.729     93.379    0.000    4911.913    5122.528
=====
Ljung-Box (L1) (Q):      0.00    Jarque-Bera (JB):      33141.63
Prob(Q):                 0.96    Prob(JB):              0.00
Heteroskedasticity (H):   4.47    Skew:                  -0.78
Prob(H) (two-sided):      0.00    Kurtosis:              22.29
=====

```

4.3. Modèle de Deep Learning - LSTM :

Dans cette section, nous avons exploré l'application d'un modèle de réseau neuronal à mémoire à court terme (LSTM) pour la modélisation de la série temporelle des prix de clôture de l'indice MASI.

Après avoir remodelé nos données d'entraînement pour convenir à l'architecture LSTM, avec une forme de (2123, 7, 1), nous avons construit notre modèle en utilisant la bibliothèque Keras. Notre modèle LSTM comprend deux couches LSTM avec 50 unités chacune, suivies d'une couche dense de 25 unités. Enfin, une couche dense avec une seule unité est ajoutée pour la prédiction.

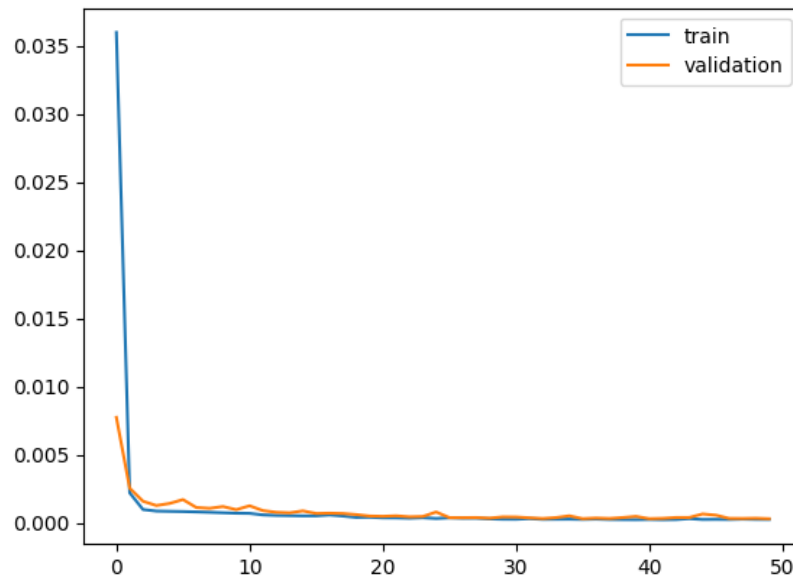
Nous avons compilé le modèle en utilisant la fonction de perte MSE (Mean Squared Error) et l'optimiseur Adam. Le modèle a ensuite été ajusté aux données d'entraînement avec un batch_size de 64, une validation_split de 0.1 et pendant 50 époques. La figure ci-dessous montre la sortie du modèle, y compris les détails sur les différentes couches, pour une compréhension approfondie de l'architecture LSTM utilisée.

Model: "sequential_9"

Layer (type)	Output Shape	Param #
lstm_15 (LSTM)	(None, 7, 50)	10400
lstm_16 (LSTM)	(None, 50)	20200
dense_16 (Dense)	(None, 25)	1275
dense_17 (Dense)	(None, 1)	26

=====
Total params: 31901 (124.61 KB)
Trainable params: 31901 (124.61 KB)
Non-trainable params: 0 (0.00 Byte)

Pour évaluer la performance du modèle, nous avons tracé les courbes d'entraînement et de validation des pertes au fil des époques. Cela nous a permis de surveiller la convergence du modèle et de détecter d'éventuels problèmes de surajustement ou de sous-ajustement.



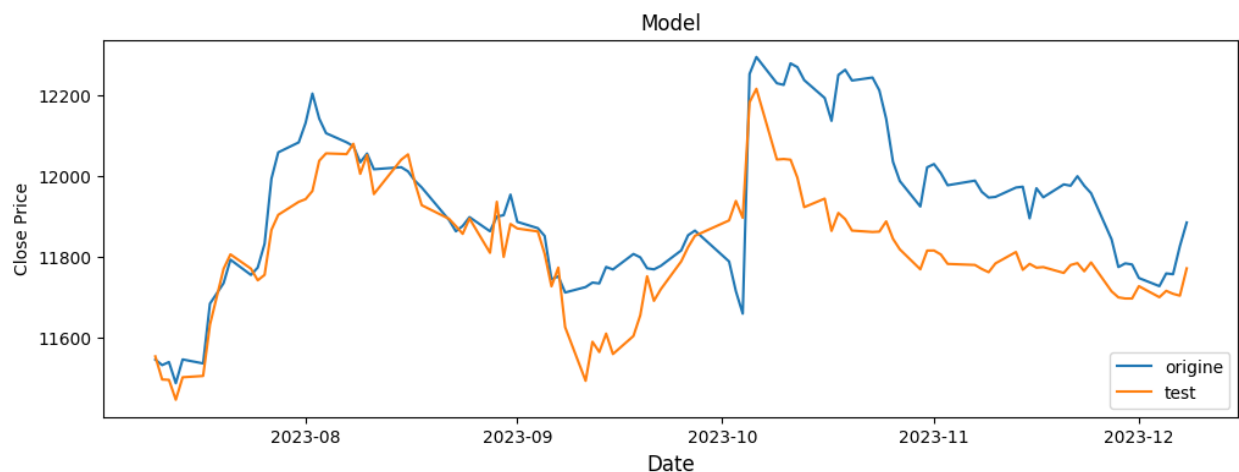
6. Modélisation avec Variables Explicatives :

Dans cette section, nous avons enrichi notre approche en intégrant des variables explicatives externes pour améliorer la prédiction des performances de l'indice MASI. Nous avons fusionné les données de l'indice général avec celles d'autres indices macroéconomiques et sectoriels, créant ainsi une table de données complète contenant 31 variables explicatives en plus de notre variable cible, le prix de clôture de l'indice MASI.

6.1. Modélisation avec Machine Learning (ML) - RF et XGB :

Pour cette phase cruciale de notre analyse, nous avons capitalisé sur les données soigneusement prétraitées, intégrant avec finesse les variables explicatives externes dans notre modèle. Cette approche a été stratégiquement orchestrée pour capturer l'influence des facteurs extérieurs sur les performances de l'indice MASI.

Nous avons appliqué deux modèles de Machine Learning de pointe, à savoir Random Forest (RF) et XGBoost (XGB), aux données résultantes. Ces modèles ont été minutieusement entraînés sur les données d'entraînement, profitant de la richesse des variables explicatives désormais intégrées. Le graphique ci-dessous montre que les prédictions générées par ces modèles comparées aux données de test, offrent une vision préliminaire sur la performance de ces modèles et ces capacités de prédire la performance de l'indice MASI.



Les résultats de cette modélisation seront examinés en profondeur dans la section des résultats, révélant les nuances des performances relatives de RF et XGB et mettant en lumière l'influence discernable des variables explicatives sur la prédiction de l'indice MASI.

6.2. Modélisation avec Deep Learning (DL) - LSTM :

Dans cette étape, nous avons exploité la puissance du Deep Learning, plus précisément Long Short-Term Memory (LSTM), pour modéliser les données en tenant compte des variables explicatives externes.

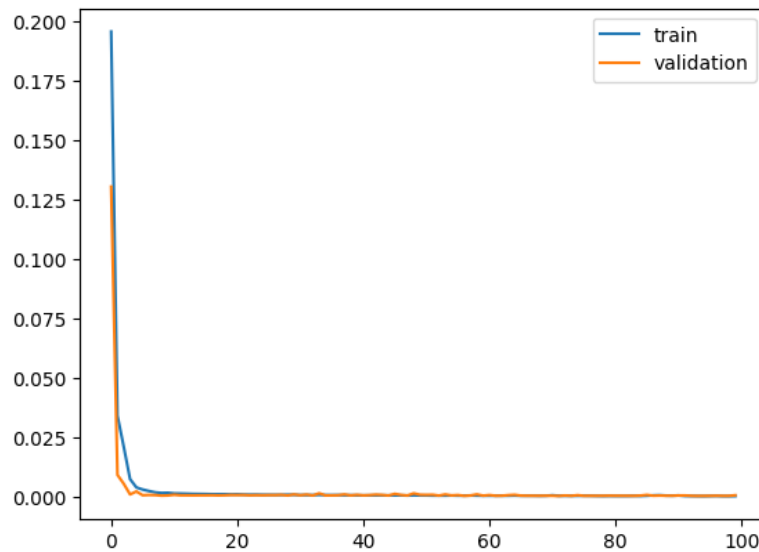
Nous avons construit un modèle LSTM séquentiel avec deux couches LSTM et une couche Dense. La première couche LSTM est configurée pour retourner des séquences, tandis que la deuxième a un retour de séquence désactivée. Nous avons également inclus une couche Dense pour la sortie du modèle. La figure ci-dessous montre la sortie du modèle, y compris les détails sur les différentes couches, pour une compréhension approfondie de l'architecture LSTM utilisée.

```
Model: "sequential_7"
```

Layer (type)	Output Shape	Param #
lstm_11 (LSTM)	(None, 7, 50)	16800
lstm_12 (LSTM)	(None, 50)	20200
dense_12 (Dense)	(None, 10)	510
dense_13 (Dense)	(None, 1)	11

```
=====
Total params: 37521 (146.57 KB)
Trainable params: 37521 (146.57 KB)
Non-trainable params: 0 (0.00 Byte)
```

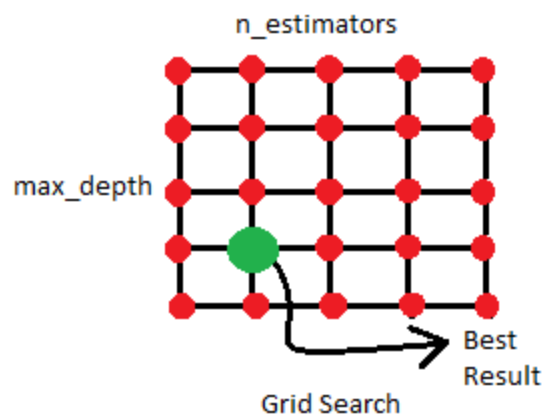
Le modèle est compilé avec la fonction de perte "mean_squared_error" et l'optimiseur "adam". Ensuite, le modèle est entraîné sur les données d'entraînement avec un ensemble d'hyper paramètres spécifiés, et l'historique d'entraînement est tracé pour évaluer les performances du modèle au fil des époques.




7. Optimisation et Évaluation des Modèles :

7.1 Optimisation des modèles ML : Grid Search

Pour maximiser les performances de nos modèles, nous avons utilisé la technique de Grid Search, une approche systématique permettant d'explorer différentes combinaisons d'hyper paramètres pour chaque modèle de machine learning. Cette technique revêt une importance cruciale, car elle nous permet de trouver la configuration optimale des hyperparamètres, conduisant ainsi à des modèles plus performants et robustes.





La technique de Grid Search fonctionne en définissant un ensemble de valeurs possibles pour chaque hyperparamètre, puis en évaluant le modèle pour chaque combinaison de ces valeurs. Cela permet d'identifier les paramètres qui conduisent à une performance optimale du modèle.

7.2 Optimisation des modèles DL :

En ce qui concerne la Deep Learning (DL) avec le modèle LSTM, nous avons effectué des ajustements minutieux sur les hyperparamètres tels que l'optimiseur, le nombre d'époques, et la taille du lot (batch_size). Ces paramètres jouent un rôle crucial dans la formation du modèle et leur optimisation contribue à obtenir des résultats plus précis.

7.3 Évaluation des Modèles :

Pour la mesure de la performance des modèles, nous avons choisi le Root Mean Squared Error (RMSE) comme critère d'évaluation principal. Le RMSE mesure la dispersion des erreurs entre les valeurs prédites et les valeurs réelles. Plus spécifiquement, l'équation du RMSE est définie comme suit :

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_{true,i} - Y_{pred,i})^2}{n}}$$

Où $Y_{true,i}$ représente la valeur réelle, $Y_{pred,i}$ représente la valeur prédite, et n est le nombre total d'observations. Le RMSE est une mesure utile pour évaluer la précision d'un modèle de prédiction, car il pénalise davantage les erreurs importantes. Ainsi, un RMSE plus bas indique une meilleure adéquation du modèle aux données réelles.

8. Résultats et Analyses :

8.1 Résultats sur la modélisation de Série Temporelle :

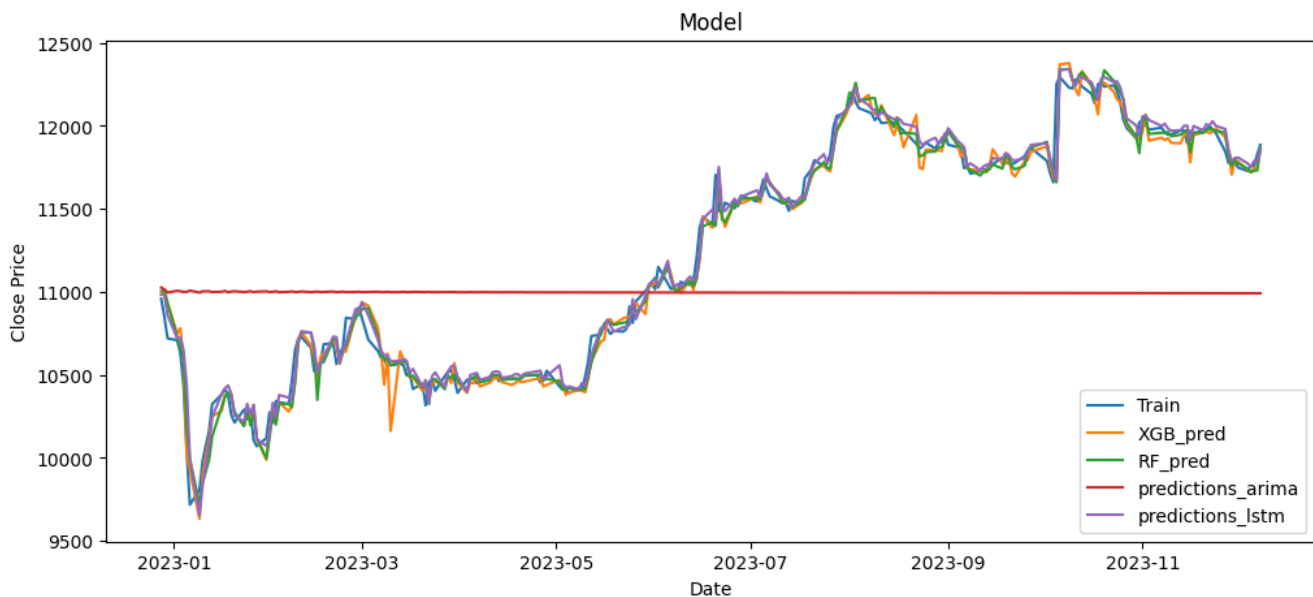
Nous avons évalué la performance de nos modèles appliqués uniquement à la variation de l'indice de MASI dans le temps en utilisant la métrique RMSE (Root Mean Squared Error). Voici les résultats regroupés dans le tableau ci-dessous :

Modèles	RMSE
LSTM	86.60
XGBoost	87.99
Random Forest Régresser	88.00
ARIMA	183.70

Nous observons que le modèle LSTM affiche la meilleure performance avec un RMSE de 86.6, indiquant une excellente adéquation de ses prédictions par rapport aux valeurs réelles. Les modèles XGB et RF présentent également des performances remarquables, avec des RMSE respectifs de 87.99 et 88.00, soulignant leur capacité à prédire la fluctuation de l'indice MASI.

Cependant, l'approche statistique ARIMA affiche une performance relativement plus faible avec un RMSE de 183.70. Cela suggère que les modèles d'apprentissage automatique (XGB, RF, LSTM) ont une capacité supérieure à capturer la complexité de la variation des prix de clôture de l'indice MASI par rapport à l'approche statistique.

Le graphique ci-dessous illustre visuellement la comparaison entre les prédictions de ces quatre modèles et les données réelles de test. On peut clairement observer que le modèle LSTM suit de près la variation des prix de clôture de l'indice MASI, démontrant ainsi sa capacité à capturer les tendances temporelles de manière précise.



8.2 Résultats Modélisation avec Variables Explicatives

Nous avons étendu notre analyse en intégrant des variables explicatives externes pour modéliser la performance de l'indice MASI. Voici les résultats obtenus en utilisant la métrique RMSE (Root Mean Squared Error) pour évaluer la performance des modèles :

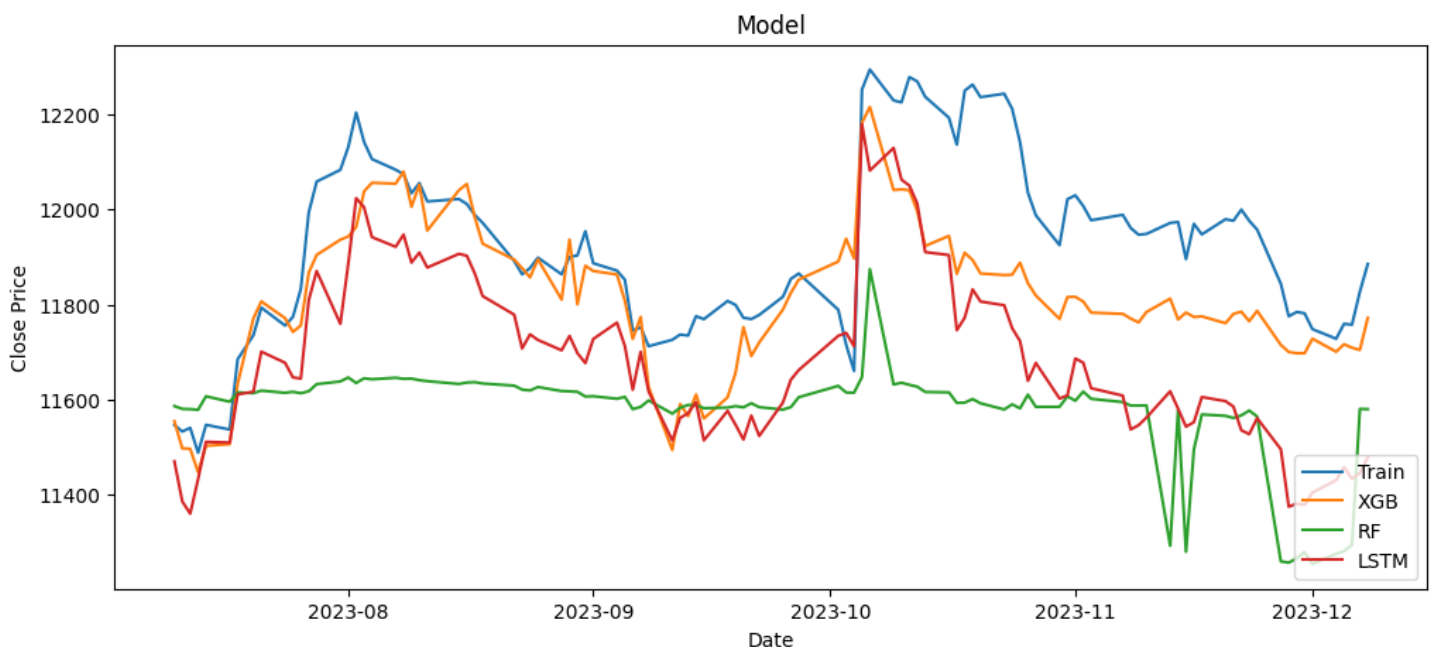
Modèles	RMSE
XGBoost	157.47
LSTM	272.80
Random Forest Régresser	303.70

Les résultats montrent que le modèle XGB affiche le RMSE le plus bas parmi les modèles, avec une valeur de 157.47. Cela indique que le modèle XGB est capable de prédire la performance de l'indice MASI de manière plus précise par rapport aux autres modèles.

Cependant, il est important de noter que le modèle RF présente un RMSE relativement plus élevé, atteignant 393.70. Cette différence peut s'expliquer par la complexité des relations entre les variables explicatives et la performance de l'indice MASI, que le modèle RF pourrait ne pas capter aussi efficacement que le modèle XGB.

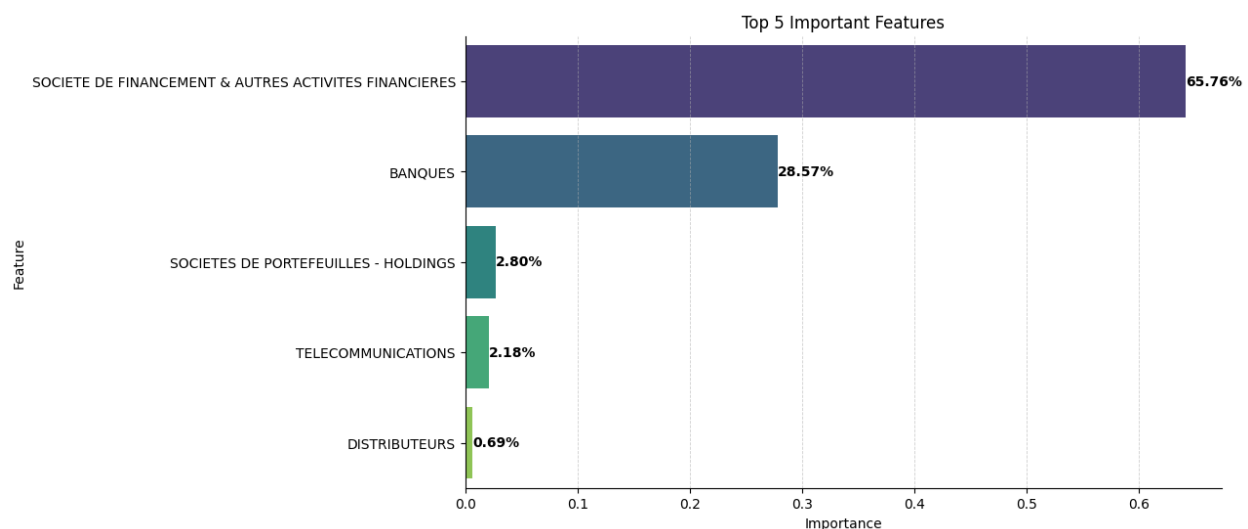
Le modèle LSTM affiche un RMSE intermédiaire de 272.80, montrant une performance robuste dans la prédiction de la performance de l'indice MASI lorsqu'il est alimenté avec des variables explicatives externes.

Le graphique ci-dessous illustre la comparaison entre les prédictions de ces trois modèles (XGB, RF, LSTM) avec des variables explicatives et les données réelles de test. On peut observer les tendances prévues par chaque modèle par rapport aux fluctuations réelles de l'indice MASI.



8.3 les facteurs les plus importants:

Dans le cadre de notre étude, il est crucial d'identifier les variables qui ont le plus d'impact sur la performance de l'indice MASI. En analysant les importances des variables obtenues à partir du modèle XGB, voici les résultats significatifs :



SOCIETE DE FINANCEMENT & AUTRES ACTIVITES FINANCIERES (SF) Importance : 64.11%

Cette variable semble jouer un rôle prépondérant dans l'explication des variations de l'indice MASI. Son importance élevée suggère que les activités financières de sociétés de financement ont un impact majeur sur la performance globale du marché boursier au Maroc.

BANQUES Importance 27.85% : Les institutions bancaires, représentées par la variable "BANQUES", occupent également une position significative. Leur influence substantielle souligne l'importance du secteur bancaire dans la détermination des mouvements de l'indice MASI.



SOCIETES DE PORTEFEUILLES - HOLDINGS Importance 2.73% : Bien que moins impactante que les deux premières, la catégorie des "SOCIETES DE PORTEFEUILLES - HOLDINGS" joue un rôle non négligeable dans l'explication des variations de l'indice MASI.

TELECOMMUNICATIONS Importance 2.13% : Le secteur des télécommunications, bien que moins prépondérant que les deux premières catégories, contribue également à l'explication de la performance de l'indice MASI.

DISTRIBUTEURS : Importance : 0.68%

La variable "DISTRIBUTEURS" représente une influence plus modérée, mais elle participe néanmoins à la compréhension des variations de l'indice.

Ces résultats soulignent l'importance de ces secteurs spécifiques dans la dynamique du marché boursier au Maroc. Les investisseurs et les décideurs peuvent tirer des insights utiles de ces données pour comprendre les forces motrices du marché et prendre des décisions éclairées.

Conclusion Générale :

Suite à une analyse approfondie des données boursières du marché marocain, notre étude visait à élaborer des modèles prédictifs pour anticiper la performance de l'indice MASI. Nous avons adopté deux approches distinctes pour cette tâche, chacune offrant des perspectives uniques.

1. Modélisation de la Variation Temporelle (Approche Série Temporelle) :


Après avoir exploré les données et sélectionné le modèle LSTM, nous avons constaté que ce dernier offre une robustesse remarquable pour la prédiction de la performance de l'indice MASI au fil du temps. Son RMSE réduit (86.6) témoigne de sa capacité à capturer les tendances temporelles du marché boursier.

2. Modélisation avec les Variables Exogènes :

En intégrant des variables exogènes, XGB s'est démarqué comme un modèle de choix avec un RMSE de 157.47. Cela confirme sa pertinence dans la prédiction en tenant compte de facteurs externes.

3. Principaux Facteurs d'Influence sur l'Indice MASI :

La composante la plus déterminante est représentée par la catégorie "SOCIETE DE FINANCEMENT & AUTRES ACTIVITES FINANCIERES (SF)" avec une prépondérance notable de 64.11%. Les activités financières jouent un rôle prédominant, exerçant une influence substantielle sur le marché. Les institutions bancaires, sous la désignation "BANQUES," détiennent également une importance significative, avec une contribution évaluée à 27.85%. Ceci met en évidence le poids substantiel du secteur bancaire dans la dynamique du marché.



L'utilisation combinée des modèles LSTM pour la prédiction du prix du MASI sans prendre en compte les variables exogènes, et de XGB en considérant les autres variables explicatives, constitue une approche robuste pour anticiper les performances de l'indice MASI. Ces stratégies se révèlent être des indicateurs de qualité, fournissant des informations essentielles aux investisseurs et aux analystes financiers.

La mise en évidence particulière des secteurs financiers et bancaires dans cette étude souligne leur rôle prépondérant en tant que moteurs essentiels du marché marocain. Cette analyse établit un socle fiable pour les prises de décision, identifiant clairement les secteurs clés à suivre dans le paysage boursier du Maroc.