EL YAOUTI Chaimae
Data Science Engineering Student, INSEA Rabat
Github: https://github.com/chaimae-elyaouti
LinkedIn: https://www.linkedin.com/in/chaimae-el-yaouti-95897b320/

Rapport de Projet:

Bitcoin : De la Volatilité à la Prédiction

Introduction

Le Bitcoin, connu pour sa forte volatilité, pose un défi majeur en matière de prévision de son prix. Ce projet a pour objectif d'analyser l'évolution du prix du Bitcoin (USD) à l'aide de techniques de séries temporelles, afin de modéliser son comportement et d'estimer ses valeurs futures.
En appliquant la méthode de Box-Jenkins et le modèle ARIMA, nous cherchons à construire un modèle prédictif fiable, malgré l'instabilité des données. L'étude inclut les étapes de visualisation, de transformation, de stationnarisation, de modélisation et de validation, jusqu'à l'évaluation finale de la qualité des prévisions.

1. Description des Données

Les données analysées proviennent de Yahoo Finance, téléchargées automatiquement à l'aide de la bibliothèque Python yfinance, en utilisant le ticker BTC-USD, qui correspond aux prix journaliers du Bitcoin exprimés en dollars américains (USD).
Le DataFrame téléchargé contient plusieurs colonnes : Open, High, Low, Close, Adj Close et Volume. Nous nous concentrons principalement sur la colonne Close, représentant le prix de clôture quotidien du Bitcoin. Cette série temporelle est utilisée pour explorer, modéliser, puis prévoir les évolutions futures du prix.

---

2. Résultats Obtenus

a. Visualisation initiale des données

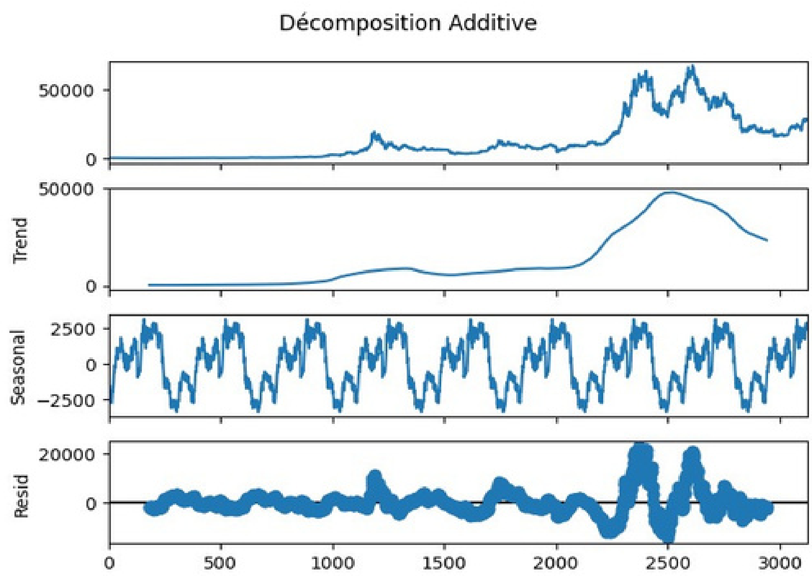
- Une première visualisation de la série a été effectuée en divisant les données en deux ensembles : données d'entrainement (train) qui sont à partir de 2014-09-17 jusqu'à 2023-04-05 et données de test (test) qui commence de la date 2023-04-06 jusqu'à 2025-05-26 .
- Cette séparation a permis de visualiser les tendances globales et de préparer le jeu de test pour évaluer la performance prédictive du modèle.

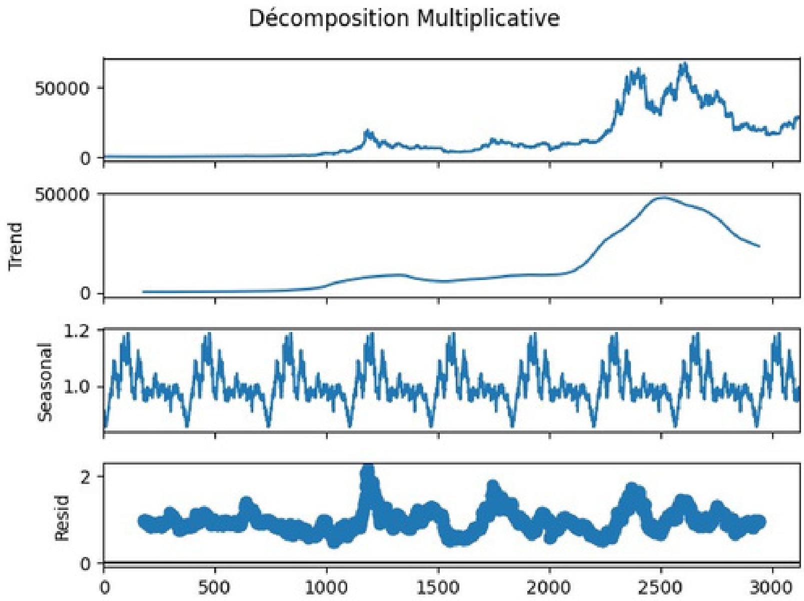
EL YAOUTI CHAIMAE
ATIF IMANE
Data Science

---

b. Décomposition de la série

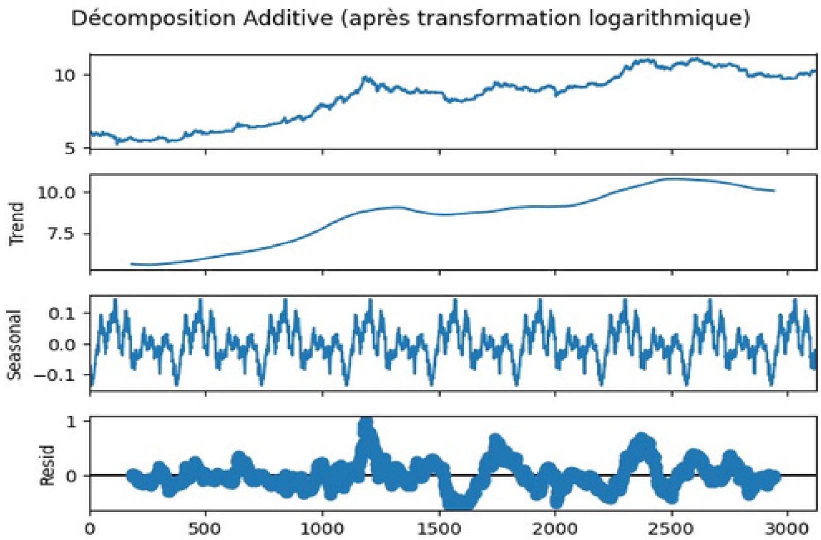
- Une décomposition additive a d'abord été testée. Cependant, cette approche s'est révélée inadaptée : la variabilité de la série est importante, et aucune saisonnalité claire ne ressort.
- En conséquence, une décomposition multiplicative a été réalisée. Ce modèle est mieux adapté, mais nécessite que la série soit stationnaire.





c. Stationnarisation

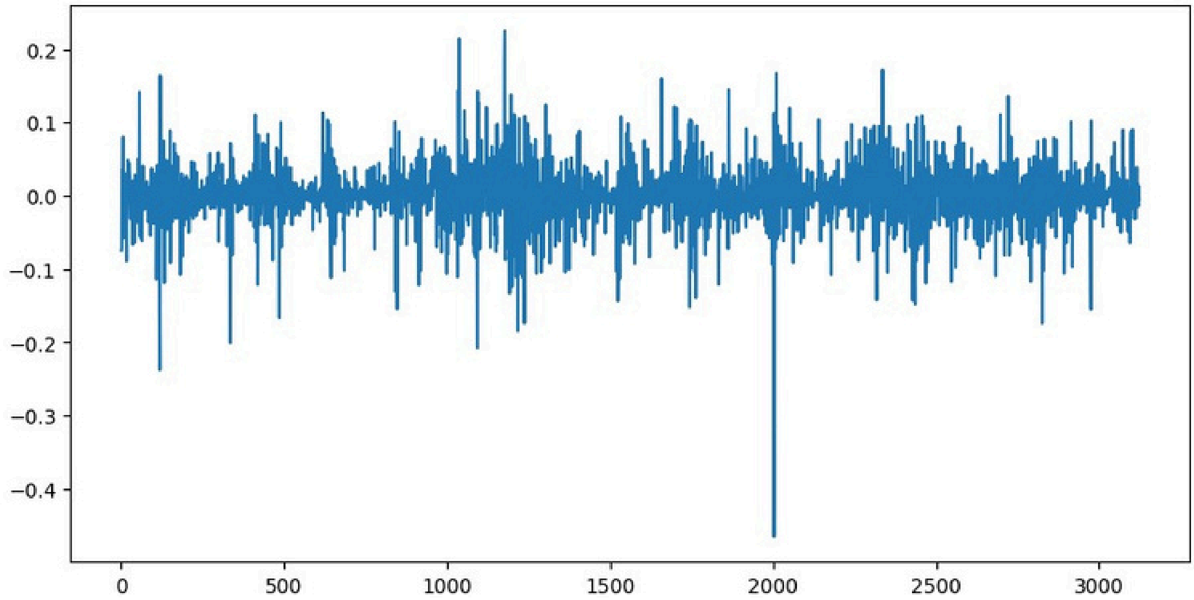
Une transformation logarithmique a été appliquée pour réduire la variance.
Un test de racine unitaire (ADF) a été réalisé pour évaluer la stationnarité. Résultat : la série n'était pas encore stationnaire : p_value supérieure à 5%
Une différenciation (logarithme différencié) a donc été nécessaire. Une nouvelle visualisation a confirmé un comportement plus stationnaire.




```
ADF Statistic : -0.877010
p_value : 0.795449
```

EL YAOUTI CHAIMAE
ATIF IMANE
Data Science

Courbe de train_log data après différenciation



---

d. Identification du modèle : Méthode de Box-Jenkins

A l'aide des corrélogrammes ACF et PACF, les valeurs maximales de p et q ont été estimées (p max = 1, q max = 1).
Plusieurs modèles ARIMA candidats ont été testés :

- ARIMA(0,1,1)
- ARIMA(1,1,0)
- ARIMA(1,1,1)

Le modèle ARIMA(0,1,1) a été retenu, car il présente les meilleurs critères AIC et BIC.
Avant de procéder manuellement à l'identification du modèle ARIMA optimal, nous avons utilisé la fonction ***auto_arima*** de la bibliothèque pmdarima, qui propose automatiquement la meilleure combinaison de paramètres (p, d, q) en se basant sur des critères d'information comme ***AIC*** (Akaike Information Criterion) et ***BIC*** (Bayesian Information Criterion) .
Résultat : ***Le modèle ARIMA(0,1,1)*** a été recommandé par auto_arima comme étant le plus performant. Ce résultat a confirmé notre propre analyse issue des correlogrammes ACF et PACF.

```
...  ARIMA(1,1,1) AIC: -11468.240979814396
     ARIMA(1,1,1) BIC: -11450.101331742548
     ARIMA(1,1,0) AIC: -10212.554426377033
     ARIMA(1,1,0) BIC: -10200.461327662468
     ARIMA(0,1,1) AIC: -11469.032368049036
     ARIMA(0,1,1) BIC: -11456.93926933447
```

```
Best model:  ARIMA(0,1,1)(0,0,0)[12]
Total fit time: 3.558 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  129
Model:               SARIMAX(0, 1, 1)   Log Likelihood               -1248.199
Date:                Wed, 28 May 2025   AIC                           2500.398
Time:                        00:56:50   BIC                           2506.102
Sample:                    09-30-2014   HQIC                          2502.715
                         - 05-31-2025
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.5685      0.053     10.822      0.000       0.466       0.671
sigma2      1.836e+07    1.29e+06     14.202      0.000    1.58e+07    2.09e+07
===================================================================================
Ljung-Box (L1) (Q):                   0.16   Jarque-Bera (JB):               152.96
Prob(Q):                              0.69   Prob(JB):                         0.00
Heteroskedasticity (H):              15.18   Skew:                             1.21
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.78
===================================================================================
```

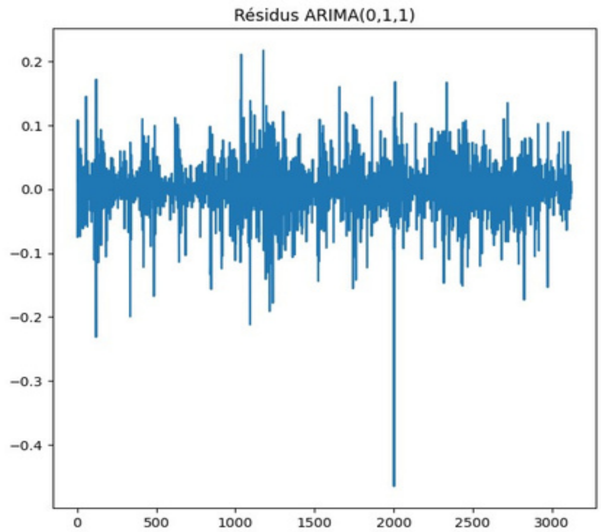
---

e. Estimation et validation du modèle

Le modèle retenu a été estimé à l'aide de .summary() de statsmodels, fournissant les paramètres du modèle.
Une analyse des résidus a été effectuée :

- Visuellement, les résidus sont centrés autour de zéro, avec une variance constante.
- Le test de Ljung-Box, cependant, indique que les résidus ne sont pas parfaitement un bruit blanc (p_value est inférieure à 5% ; donc on rejette l'hypothèse : le résidus est un bruit blanc.)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 3124
Model:                 ARIMA(0, 1, 1)   Log Likelihood                5749.213
Date:                Tue, 27 May 2025   AIC                         -11494.426
Time:                        22:59:55   BIC                         -11482.333
Sample:                             0   HQIC                        -11490.085
                               - 3124
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.0164      0.011     -1.458      0.145      -0.038       0.006
sigma2         0.0015   1.48e-05     99.355      0.000       0.001       0.002
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):             15867.70
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               1.05   Skew:                            -0.77
Prob(H) (two-sided):                  0.41   Kurtosis:                        13.93
===================================================================================
```




```
Résultat du test de Ljung-Box :
    lb_stat  lb_pvalue
10  42.303145   0.000007
```

---

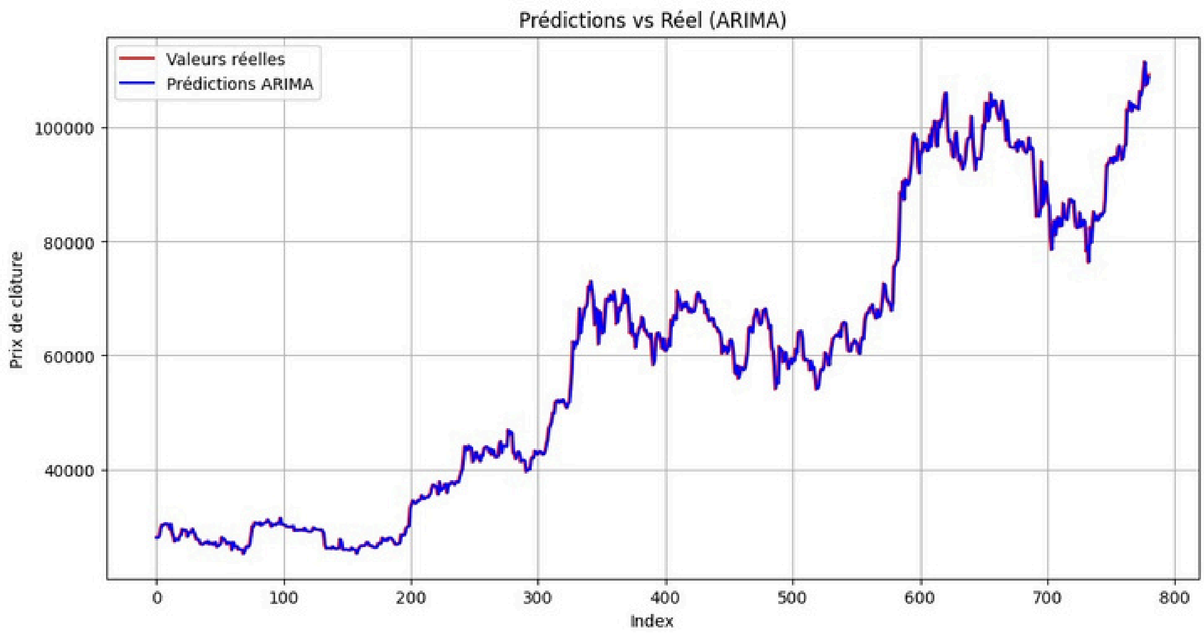
f.Prévision et évaluation

Pour la phase de prévision, nous avons adopté une approche améliorée d'ARIMA en utilisant une ***stratégie à fenêtre glissante (rolling forecast)***.

Étant donné la ***forte variabilité des données du Bitcoin***, il n'est pas pertinent de prédire toutes les valeurs futures d'un seul coup à partir d'un modèle figé. À la place, nous avons procédé comme suit :

Le modèle ARIMA(0,1,1) est d'abord entraîné sur les données d'apprentissage.
Pour chaque nouvelle prévision, la ***valeur prédite précédente est ajoutée dynamiquement aux données d'entraînement***.
Le modèle est ensuite ***réentraîné à chaque itération***, permettant une ***mise à jour constante*** qui tient compte des nouvelles dynamiques du marché.

Cette méthode améliore significativement la qualité des prédictions sur des séries aussi instables que celle du Bitcoin.
Les résultats montrent que les ***valeurs prévues sont très proches des valeurs réelles***, et l'erreur de prévision reste très faible avec un ***MAPE = 1.76%***, ce qui indique une ***excellente précision*** du modèle.




```
... MAPE : 1.758606543003495%
```

---

3. Discussion et Interprétation des Résultats

a. Difficultés rencontrées

- La première difficulté a été de choisir la bonne forme de décomposition : le modèle additif a été écarté rapidement à cause de la forte variabilité.

Uneautrecomplexité importante était la stationnarisation. Malgré une transformation logarithmique, la série n’était pas encore stationnaire, ce qui a nécessité une différenciation.

- Enfin, bien que le modèle ARIMA(0,1,1) ait de bonnes performances, le test de Ljung-Box n’a pas validé complétement l’hypothèse de résidus blancs, ce qui montre que le modèle n'est pas parfait statistiquement, mais reste très performant empiriquement.

b. Choix méthodologiques justifiés

Le choix d’un ARIMA non saisonnier est justifié, car aucune saisonnalité claire n’apparait dans les données.
La méthode Box-Jenkins a permis d’identifier un modèle sobre, mais efficace.
L’approche a bien équilibré complexité et performance.

c. Conclusion

Malgré des imperfections statistiques (résidus non totalement blancs), le modèle ARIMA(0,1,1) offre une précision remarquable sur les prévisions à court terme. Le MAPE très faible valide ce choix. Ce travail montre l’importance de la transformation, de l’analyse exploratoire et des critères d’évaluation dans la modélisation de séries temporelles financières.
