Prédiction des prix et actifs du S&P 500

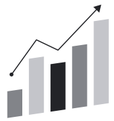
Eva Brixy, Hugo Gatti, Manon Berny, Quentin Doblado, Théo Leturcq

COLLECTE DES DONNÉES

Les données ont été collectées 2018 à 2023 et sont les données du S&P 500 :

- Prix de clôture quotidiens du S&P 500,
- Récupérés via Yahoo Finance

Des variables exogènes ont été rajoutées à nos données :



VIX : L'indice de volatilité, également récupéré via Yahoo Finance. Il reflète les attentes du marché en matière de volatilité future.

Taux d'intérêt : Le taux des fonds fédéraux des États-Unis, récupéré via Yahoo Finance.

CPI (Indice des prix à la consommation) : Un indicateur économique mesurant les variations des prix des biens et services dans l'économie (données du CPI disponibles via Federal Reserve Economic Data).

PRÉ-TRAITEMENT DES DONNÉES

Une fois les données collectées, nous avons effectué plusieurs étapes de prétraitement pour garantir leur qualité et leur préparation pour la modélisation :

Nettoyage des données :

- Fusion des données
- Détection des valeurs manquantes et doublons : lignes contenant des valeurs manquantes supprimées

Calcul des rendements log :

- Les rendements log ont été calculés pour chaque série de données (S&P 500, VIX, taux d'intérêt et CPI) selon la formule suivante :

$$\text{Rendement log} = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

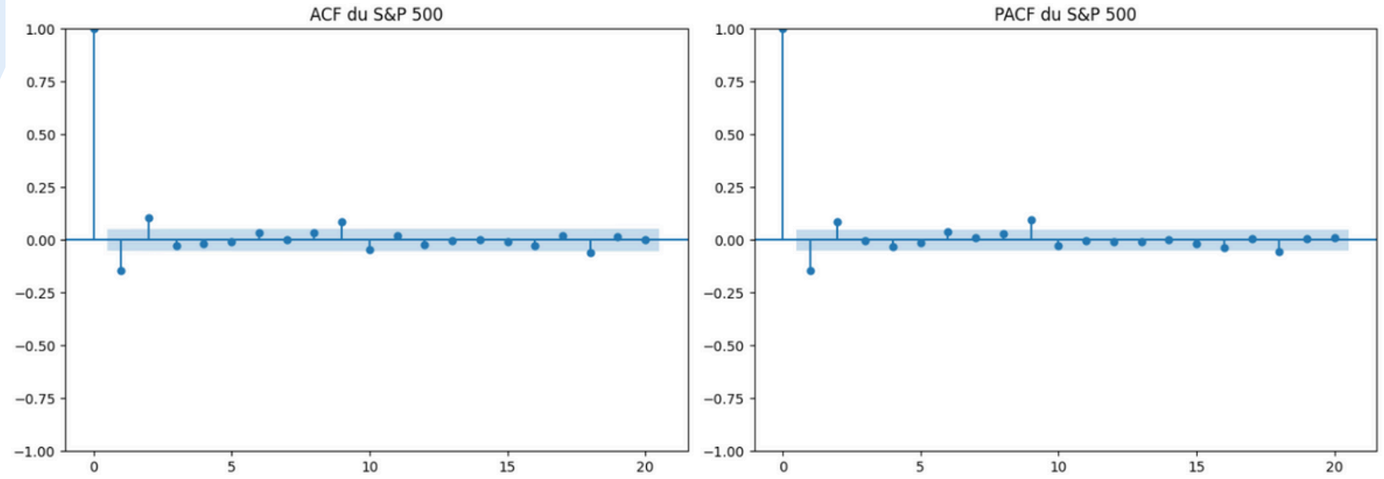
où P_t est le prix à l'instant t et P_{t-1} est le prix à l'instant $t-1$.

PRÉ-TRAITEMENT DES DONNÉES

Nous avons effectué une analyse de stationnarité et déterminé les décalages temporels.

Tous les tests ADF ont montré que les séries en log return sont stationnaires, ce qui nous permet d'utiliser les données sans transformations supplémentaires.

```
Test ADF pour SP500_Log_Return:
Statistique ADF: -11.047358285865887
p-value: 5.1827355070440374e-20
Hypothèse nulle: La série a une racine unitaire (non stationnaire)
Conclusion: La série est stationnaire si p-value < 0.05
```



ACF & PACF : fortes dépendances temporelles, surtout pour les rendements immédiats (lag 1)

↓

Des modèles comme AR(1) ou des modèles de ML (Random Forest, Régression Linéaire) seront bien adaptés

FEATURE ENGINEERING

Calcul d'indicateurs techniques

Les moyennes mobiles simples (**SMA**) et exponentielles (**EMA**) :

- la SMA à court terme (10 périodes) suit de plus près les variations récentes du prix.
- la SMA à long terme (50 périodes) est plus lisse et reflète la tendance générale sur une période plus longue.
- l'EMA donne plus de poids aux données récentes, ce qui la rend plus réactive aux changements récents des prix.

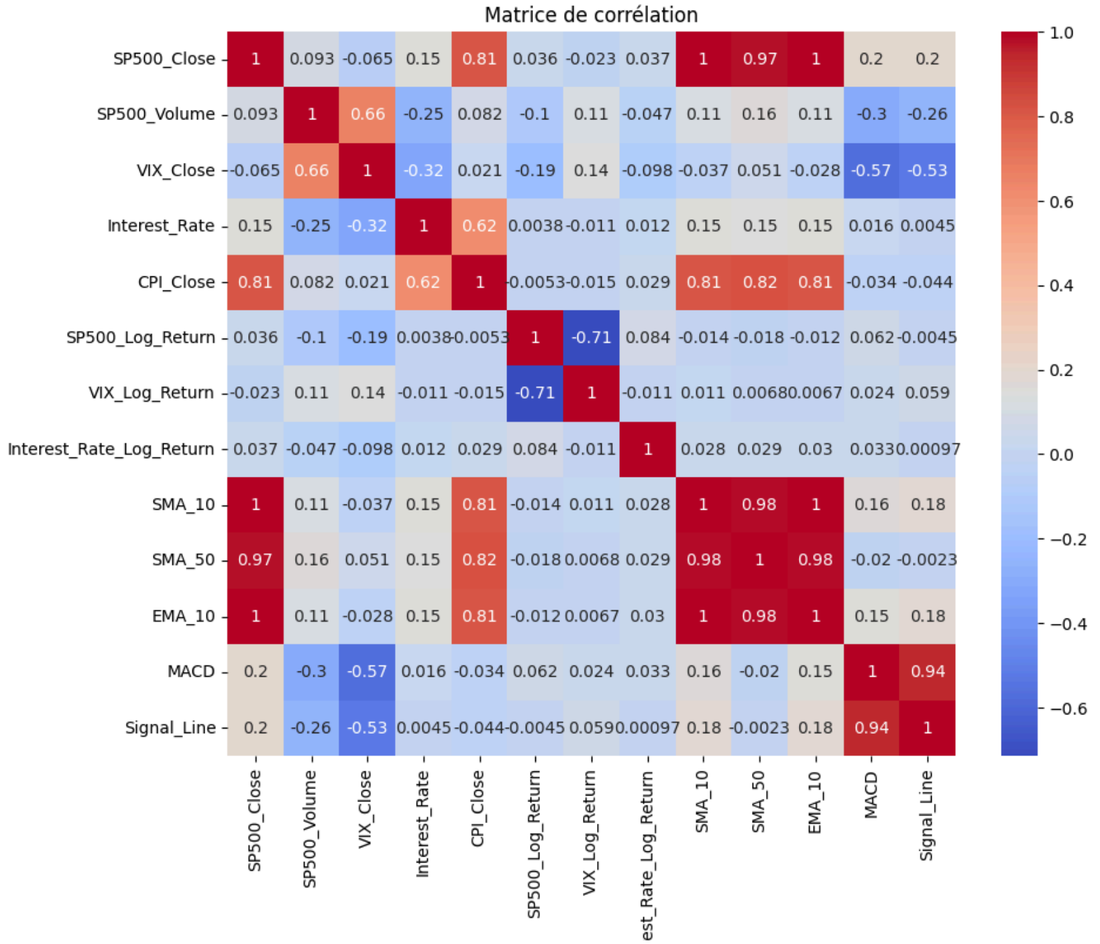
La Moyenne Mobile de Convergence / Divergence (**MACD**)

La MACD mesure la force et la direction d'une tendance. Si elle est positive, elle indique une dynamique positive du marché. Elle est calculée selon la formule :

$$\text{MACD} = \text{EMA rapide (12)} - \text{EMA lente (26)}$$

FEATURE ENGINEERING

Pour identifier les variables clés, nous avons tracé une heatmap



Variable cible : **SP500_Log_Return**

Variables explicatives (features) :

- VIX_Close
- Interest_Rate
- CPI_Close
- SMA_10
- SMA_50
- VIX_Log_Return
- Interest_Rate_Log_Return
- MACD
- Signal_Line

DÉVELOPPEMENT D'UN MODÈLE PRÉDICTIF

Choix des modèles

Régression linéaire

- Simplicité et capacité à modéliser des relations linéaires entre les variables
- Identifie rapidement les tendances générales sans nécessiter un volume important de ressources de calcul

Random Forest

- Capacité à capturer des relations non linéaires et complexes dans les données
- Nature multi-dimensionnelle et potentiellement non linéaire des données,

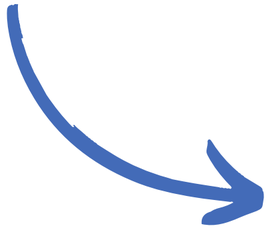
PERFORMANCE DES MODÈLES

Evaluation de métriques

MAE (Mean Absolute Error) : moyenne des erreurs absolues entre les valeurs prédites et les valeurs réelles

RMSE (Root Mean Squared Error) : racine carrée de la moyenne des carrés des erreurs

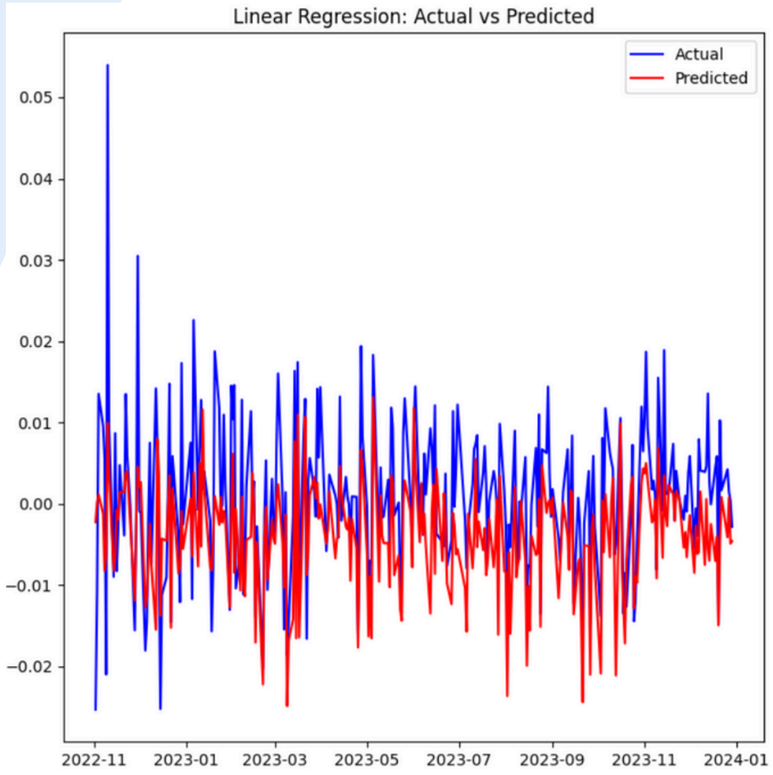
Accuracy : proportion de prédictions correctement orientées par rapport aux tendances réelles

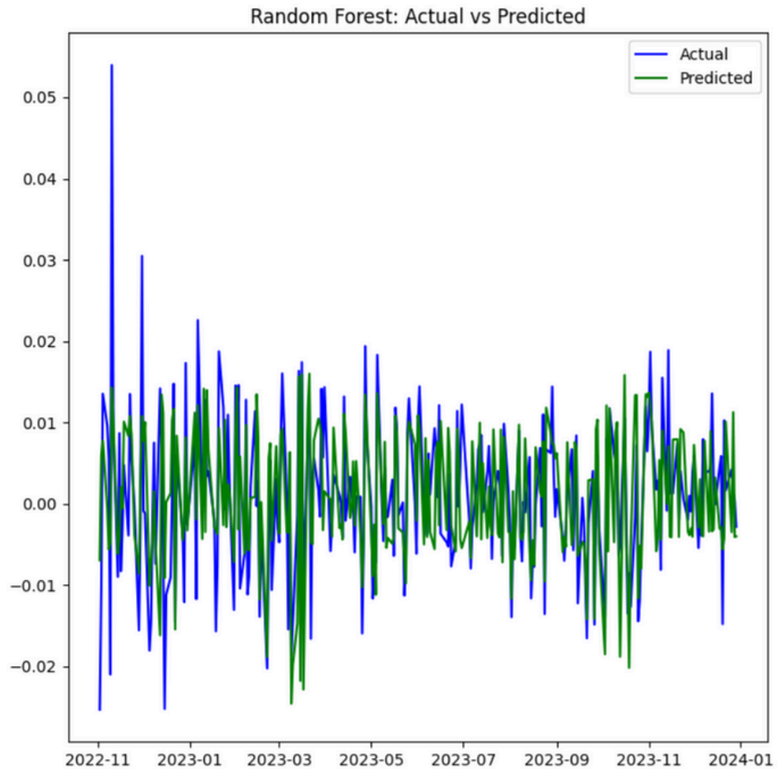


La régression linéaire a globalement des erreurs plus élevées que la Random Forest : le modèle suit moins bien les rendements réels du S&P 500

PERFORMANCE DES MODÈLES

Etude graphique





Random Forest capture mieux les variations et est globalement plus performant

La capacité de Random Forest à modéliser des relations non linéaires en fait un meilleur choix pour nos données

SOURCES D’ERREUR

Multicolinéarité

- Causes : Features comme SMA_10, SMA_50, MACD fortement corrélées, rendant les coefficients instables.
- Signes : Coefficients non intuitifs, performances incohérentes entre splits.

Endogénéité

- Causes : Utilisation de données futures ou features dépendantes de la cible.
- Signes : Précision élevée sur test mais faible généralisation sur données réelles

Non stationnarité

- Causes : Séries temporelles avec tendances ou volatilité non corrigées.
- Signes : Fluctuations importantes des erreurs (MAE, RMSE) entre splits.

SOURCES D’ERREUR

Améliorations proposées :

Optimiser les hyperparamètres du modèle via GridSearchCV

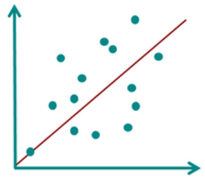
Ajouter des décalages temporels (lags) pour éviter l'endogénéité

Tester des modèles avancés comme Gradient Boosting ou XGBoost ou de Deep Learning

Utiliser une méthode de type PCA pour réduire la colinéarité et sélectionner nos variables

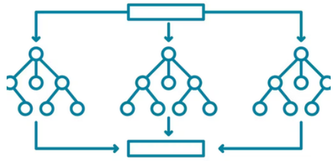
CONCLUSION

Prédiction des rendements du S&P 500 en tirant parti de leur interaction avec des variables exogènes



Régression linéaire

La meilleure en termes de MAE et RMSE



Random forest

La meilleure pour les relations complexes et non linéaires

INVESTMENT STRATEGIES

Efficient frontier

Portefeuille optimal pour un rendement maximisé à un niveau de risque donné

Stratégie de momentum

Parier sur l'inertie