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

Andrea CHECCHI
Hugo MADOUAS
Colin LEBLANC
Paul ROFIDAL
Dylan SAYARATH

1. Introduction et Objectifs

Les marchés financiers sont des systèmes dynamiques et complexes influencés par une multitude de facteurs économiques, sociaux et politiques. Dans ce contexte, l'analyse des prix des actifs financiers, tels que ceux du S&P 500, reste une tâche ardue mais cruciale pour les investisseurs et les économistes. L'objectif de ce projet est d'utiliser et tester plusieurs modèles de machine learning permettant de prédire les prix et les rendements des actifs du S&P 500, en tenant compte des données historiques et des variables exogènes telles que les indicateurs économiques, les taux d'intérêt et le VIX (indice de volatilité).

Dans ce cadre, nous allons explorer plusieurs modèles de machine learning pour prédire les prix: ARIMA, Random Forest et le Gradient Boosting.

Le projet se structure en quatre étapes principales :

- Collecte et préparation des données financières
- Développement d'un modèle prédictif
- Évaluation des performances
- Interprétation des résultats et recommandations d'investissement

En résumé, ce projet vise à utiliser le machine learning pour mieux comprendre les dynamiques du marché financier, prédire les évolutions des prix du S&P 500, et proposer des stratégies d'investissement éclairées à partir des résultats des modèles.

2. Méthodologie (préparation des données, choix des modèles)

<u>Sources des Données et Période d'observation</u>

Les données ont été collectées à partir des sources suivantes :

- **Yahoo Finance** : Utilisé pour obtenir les prix historiques du S&P 500, y compris les prix de clôture ajustés, les volumes de transactions et l'indice VIX.
- **FRED (Federal Reserve Economic Data)** : Utilisé pour récupérer des données macroéconomiques telles que l'inflation (CPI), le PIB, le taux de chômage aux États-Unis et les taux d'intérêts.

Les données couvrent une période de cinq ans (2019 à 2024) afin d'assurer une vue d'ensemble des tendances récentes et des variations économiques significatives.

Prétraitement des Données

a) Concaténation des Données

Toutes les données collectées ont été synchronisées sur l'axe des dates. Cela a été accompli en alignant les différentes sources de données selon une fréquence journalière, avec remplissage des valeurs manquantes lorsque nécessaire (voir section suivante).

b) Traitement des Valeurs Manquantes

Pour les données macroéconomiques de fréquence plus faible (mensuelle, trimestrielle), nous les avons transformées en fréquence journalière à l'aide de la fonction ***ffill*** afin de propager la dernière valeur connue.

Ensuite, les valeurs manquantes (« NAN ») ont été retirées pour garantir l'intégrité des analyses ultérieures. Cela a impliqué de filtrer les dates où des données critiques étaient absentes pour l'un ou l'autre indicateur.

c) Gestion des Valeurs Aberrantes

Les valeurs aberrantes, définies comme étant éloignées d'au moins 5 écarts types de la moyenne, ont été identifiées et supprimées. Cette procédure a permis de limiter l'impact de fluctuations extrêmes non représentatives des tendances générales.

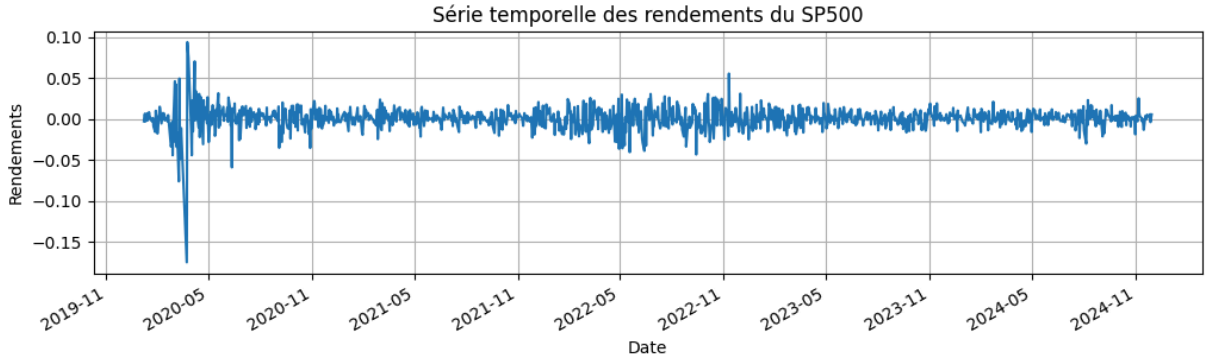
d) Calcul des Rendements

Les rendements journaliers ont été calculés de deux manières (Rendements simples, Log-rendements) à la à l'aide de la colonne "Adj Close", qui représente les prix successifs du SP500, ajustés des dividendes versés.

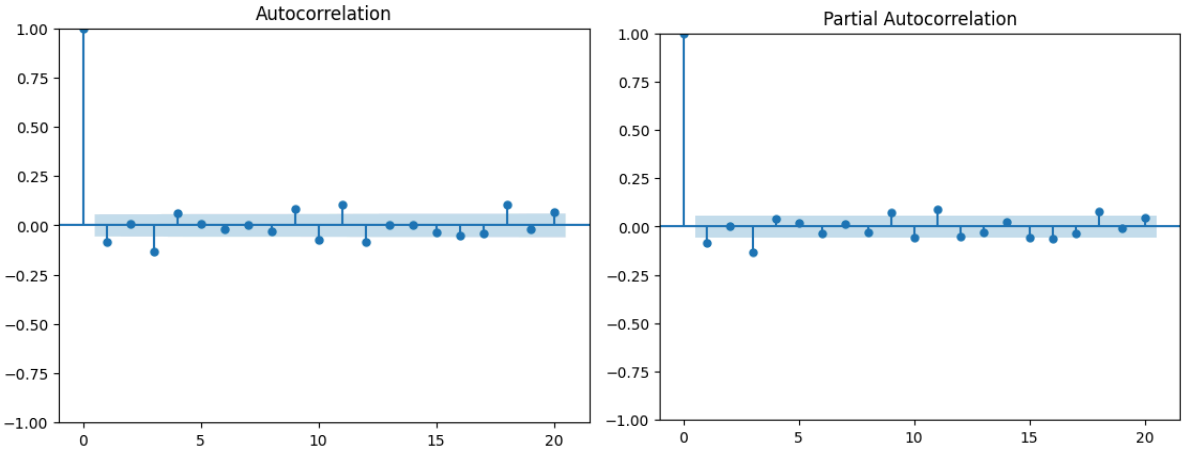
e) Stationnarité de la série temporelle des rendements

Pour évaluer la stationnarité de la série temporelle décrite par les rendements nous avons effectué le test de Dickey Fuller augmenté. La P-value du test est inférieure à

0.05, on rejette l'hypothèse nulle, la série est donc stationnaire. Le graphique ci-dessous confirme cette analyse.



f) Autocorrélation de la série des rendements



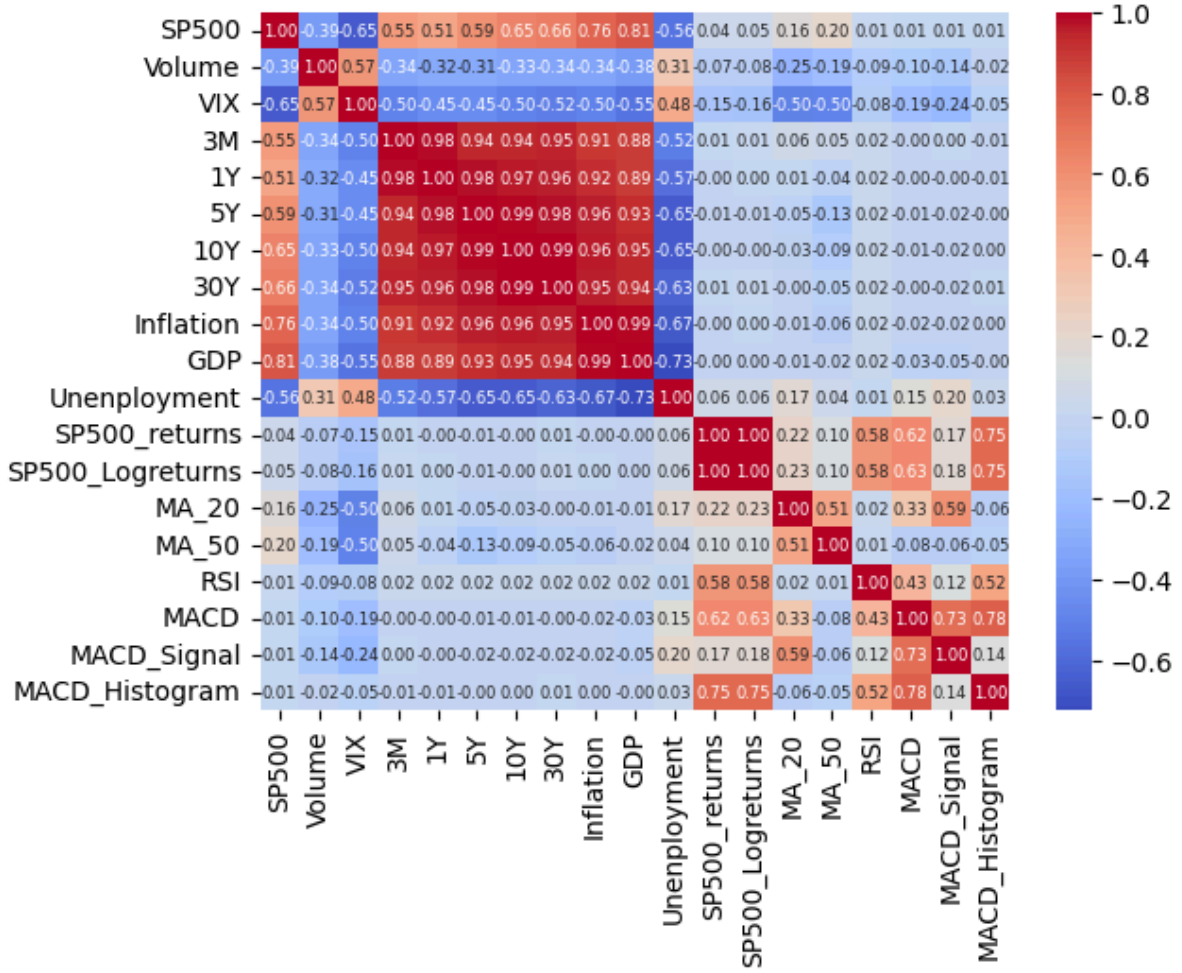
Dans les deux graphiques, seul le premier lag est significatif. Cela suggère dans un premier temps un AR(1)

<u>Features Engineering</u>

A partir des données collectés nous avons créé plusieurs variables dérivées:

- Moyenne Mobile 20 (MA20)
- Moyenne Mobile 50 (MA50)
- RSI
- MACD
- MACD signal
- MACD histogram

Nous analysons grâce à une matrice de corrélation les features que nous allons exploiter dans les modèles de machine learning.



Nous utilisons les variables les plus corrélées pour réaliser entraîner nos modèles et réaliser les prédictions.

A partir des données nous utilisons la loi de pareto (80/20) pour les séparer en un train set et un test set qui seront utilisés respectivement pour entraîner et tester la performance des modèles de machine learning.

Choix des modèles

Tout d'abord, nous avons élaboré un modèle de régression linéaire simple.

Nous avons ensuite utilisé un Random Forest Classifier et un Gradient Boosting Classifier pour prédire s'il y aura une hausse ou baisse du SP500 le jour suivant.

Pour calculer le return attendu le lendemain (ie le niveau du SP500 le jour suivant) nous utilisons un Gradient Boosting Classifier

3. Résultats et Analyses

Régression linéaire

Voici ce qu'il en est sorti :

```
MSE: 5.7681203387994996e-05
R²: 0.0466612775291817
         Feature  Coefficient
5             5Y     0.035698
7            30Y     0.034241
9            GDP     0.003540
10  Unenployment     0.003308
4             1Y     0.000751
1         Volume     0.000607
8      Inflation    -0.000246
0          SP500    -0.001175
2            VIX    -0.005259
3             3M    -0.008634
6            10Y    -0.064505
```

Le MSE est faible. Cependant, cela ne permet pas de juger réellement notre modèle de prédiction, car les log-rendements journaliers sont par nature très petits, et donc les écarts à la réalité également. Le R2 nous permettra d'avoir un avis plus précis sur notre modèle.

R2 = 0.0467, ce qui signifie que seulement 4.67% de la variance de nos rendements est expliquée par nos variables indépendantes. Cela est très faible.

On peut donc conclure que le modèle linéaire n'est pas précis et ne permet pas de prédire efficacement nos rendements. Ce qui est logique, car en finance les rendements sont très volatiles et ne peuvent pas être décrits par des modèles linéaires simples.

Random Forest classifier

```
Accuracy (simple score): 0.4255
Accuracy: 0.4255

Confusion Matrix:
[[28 75]
 [60 72]]

Precision: 0.4898
Recall: 0.5455
F1-Score: 0.5161

Classification Report:
              precision    recall  f1-score   support

           0       0.32      0.27      0.29       103
           1       0.49      0.55      0.52       132

    accuracy                           0.43       235
   macro avg       0.40      0.41      0.40       235
weighted avg       0.41      0.43      0.42       235
```

Performance globale plutôt faible:

- L'accuracy de 49.58% est proche du hasard, ce qui indique que le modèle a besoin d'améliorations.
- Le modèle semble légèrement favorisé pour prédire des hausses (classe 1) au détriment des baisses (classe 0).

Sources potentielles d'erreurs:

- Données non représentatives: Les indicateurs utilisés peuvent être insuffisants pour différencier efficacement les classes.
- Corrélation des variables: Des corrélations entre les variables pourraient limiter la capacité du modèle.

Pistes d'amélioration

Ajout de variables explicatives en intégrant davantage de facteurs macroéconomiques ou des indicateurs techniques plus spécifiques.

Optimisation des hyperparamètres: en ajustant les paramètres du Random Forest (n estimators , max depth , min samples split n estimators ,max depth,min samples split) pour trouver une meilleure configuration.

Test d'autres modèles comme le Gradient Boosting qui pourrait mieux capter les variations des données.

<u>Gradient Boosting classifier</u>

```
Accuracy: 0.5021

Confusion Matrix:
[[34 69]
 [48 84]]

Precision: 0.5490
Recall: 0.6364
F1-Score: 0.5895

Classification Report:
              precision    recall  f1-score   support

           0       0.41      0.33      0.37       103
           1       0.55      0.64      0.59       132

    accuracy                           0.50       235
   macro avg       0.48      0.48      0.48       235
weighted avg       0.49      0.50      0.49       235
```

L'accuracy (50.21%) et les scores faibles (F1-score, précision, rappel) montrent que le modèle ne capture pas correctement les patterns des données. Le modèle a une légère préférence pour prédire la hausse (classe 1), mais reste imprécis dans les deux cas.

Bien qu'il y ait un équilibre relatif entre les classes (103 baisses vs 133 hausses), le modèle semble biaisé. De plus, la complexité des interactions avec des variables explicatives utilisées (comme les indicateurs techniques et la volatilité) pourraient ne pas être suffisantes pour capturer toute la complexité des données.

Les résultats actuels ne semblent pas produire des performances satisfaisantes afin de les exploiter de manière fiable.

Gradient Boosting regressor

```
Mean Absolute Error (MAE): 0.0057
Mean Squared Error (MSE): 0.0001
Root Mean Squared Error (RMSE): 0.0074
R-squared (R²): 0.5260
```

L'erreur moyenne de 0.0056 suggère que les prédictions du modèle sont assez proches des valeurs réelles en moyenne, ce qui est un bon résultat avec les données cibles de faible amplitude.

Pour ce qui est de notre faible valeur du MSE, cela indique que les grandes erreurs sont bien contrôlées et que le modèle est globalement performant.

L'erreur moyenne quadratique est de 0.75%, ce qui est faible pour des retours financiers.

Le R^2 indique que 51.08% de la variance des rendements du SP500 est expliquée par le modèle. Bien qu'un score supérieur à 0.5 montre une certaine qualité du modèle, cela laisse une part importante de la variance inexpliquée (environ 49%).

On peut donc en conclure que ce modèle semble bien capter les tendances globales des rendements, mais il ne capture pas complètement la variabilité des retours, en partie parce que les données financières ont tendance à être assez bruitées.

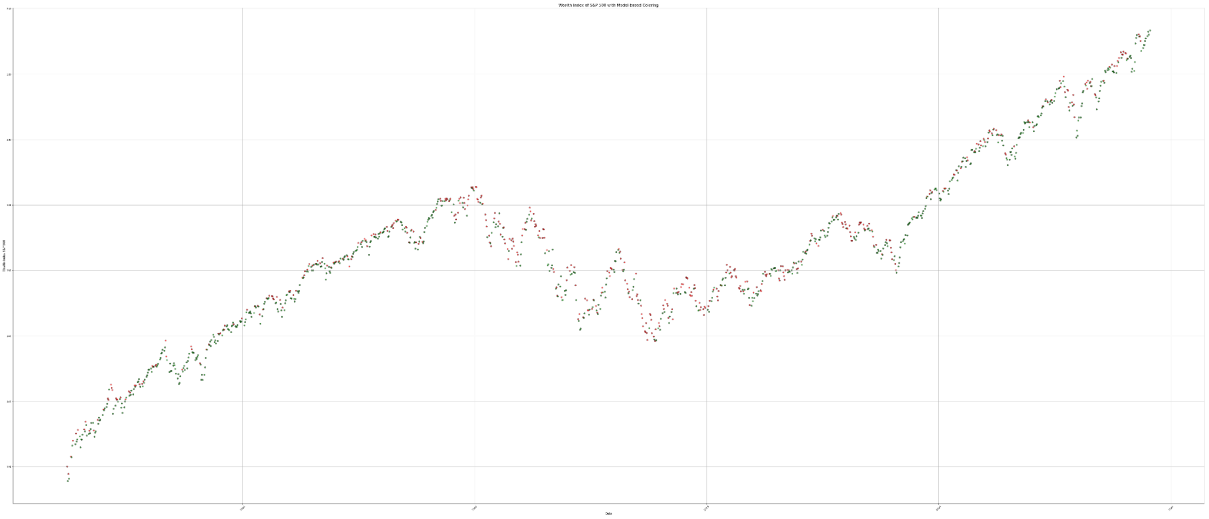
4. Stratégie d'investissement

Construction d'une stratégie d'investissement à l'aide d'un des modèles:

En utilisant les prédictions du Gradient Boosting Classifier nous pouvons mettre en place la stratégie suivante:

- Achat si une hausse est prédite
- Vente si une baisse est prédite

Sur le graphique ci dessous nous pouvons observer les moments la stratégie vend (en rouge) et achète en vert



Sur le graphique ci dessous nous pouvons observer l'évolution de 100€ investis sur le S&P500 (bleue) et 100€ dans la stratégie (orange)