*Berny, Brixy, Doblado, Gatti, Leturcq*

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

Lien vers le code :

https://colab.research.google.com/drive/1X-MglmVaHlC7OLfTIfXiG_g5AKrNdHQt?usp=sharing

## Partie 1. Collecte et Préparation des Données

### 1. Collecte des Données

Les données ont été collectées sur la période de 5 ans de **2018 à 2023**. Les principales sources de données utilisées sont :

- **S&P 500** : Les prix de clôture quotidiens du S&P 500, récupérés via **Yahoo Finance**.
- **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. Nous avons utilisé les données du **CPI** disponibles via **FRED** (Federal Reserve Economic Data).

| | SP500_Close | SP500_Volume | VIX_Close | Interest_Rate | CPI_Close |
|---|---|---|---|---|---|
| **2018-01-01** | NaN | NaN | NaN | NaN | 248.859 |
| **2018-01-02** | 2695.810059 | 3.397430e+09 | 9.770000 | 1.378 | NaN |
| **2018-01-03** | 2713.060059 | 3.544030e+09 | 9.150000 | 1.370 | NaN |
| **2018-01-04** | 2723.989990 | 3.697340e+09 | 9.220000 | 1.370 | NaN |
| **2018-01-05** | 2743.149902 | 3.239280e+09 | 9.220000 | 1.370 | NaN |

### 2. Préparation des Données

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

1. *Nettoyage des données :*
    - Nous avons vérifié les données pour détecter les valeurs manquantes et les doublons. Toutes les lignes contenant des valeurs manquantes ont été supprimées.

2. *Calcul des rendements log :*

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

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

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

| | SP500_Close | SP500_Volume | VIX_Close | Interest_Rate | CPI_Close | SP500_Log_Return | VIX_Log_Return | Interest_Rate_Log_Return |
|---|---|---|---|---|---|---|---|---|
| **2018-01-03** | 2713.060059 | 3.544030e+09 | 9.150000 | 1.370 | 248.859 | 0.006378 | -0.065563 | -0.005822 |
| **2018-01-04** | 2723.989990 | 3.697340e+09 | 9.220000 | 1.370 | 248.859 | 0.004021 | 0.007621 | 0.000000 |
| **2018-01-05** | 2743.149902 | 3.239280e+09 | 9.220000 | 1.370 | 248.859 | 0.007009 | 0.000000 | 0.000000 |
| **2018-01-08** | 2747.709961 | 3.246160e+09 | 9.520000 | 1.380 | 248.859 | 0.001661 | 0.032020 | 0.007273 |
| **2018-01-09** | 2751.290039 | 3.467460e+09 | 10.080000 | 1.415 | 248.859 | 0.001302 | 0.057158 | 0.025046 |

3. *Test :*

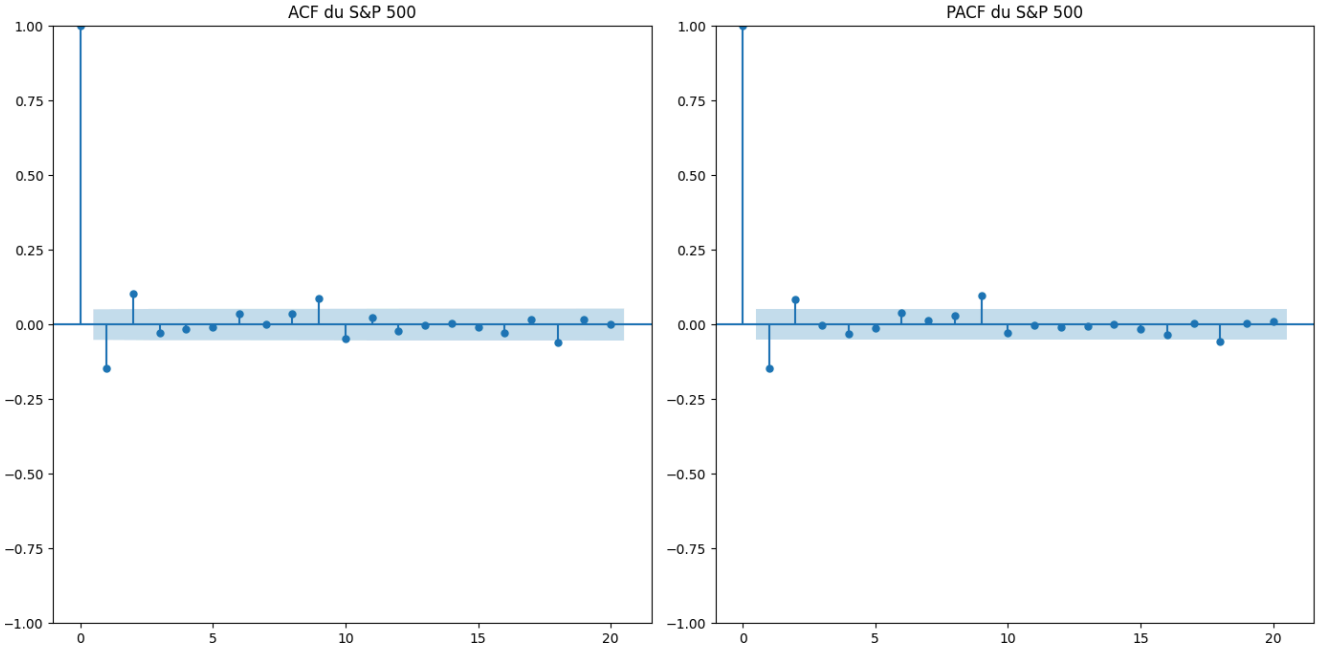
***ADF***

| Série | SP500_Close | SP500_Volume | VIX_Close | Interest_Rate |
|---|---|---|---|---|
| p-value | 0.80670085 | 3.06472e-05 | 0.00018371 | 0.19577398 |

| Série | CPI_Close | SP500_Log_Return | VIX_Log_Return | Interest_Rate_Log_Return |
|---|---|---|---|---|
| p-value | 0.9945430 | 5.1827355e-20 | 0.0 | 1.9298006e-15 |

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

### *ACF et PACF*




L'analyse a révélé des **dépendances temporelles** fortes, particulièrement pour les rendements immédiats (lag 1), ce qui suggère que des modèles comme **AR(1)** ou des modèles de machine learning (Random Forest, Régression Linéaire) sont bien adaptés pour cette tâche.

## 3. Feature engineering

Calcul d'indicateurs techniques :

1. *Moyennes mobiles simples (SMA) et exponentielles (EMA)*

La SMA à court terme (10 périodes) suit de plus près les variations récentes du prix, a 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.

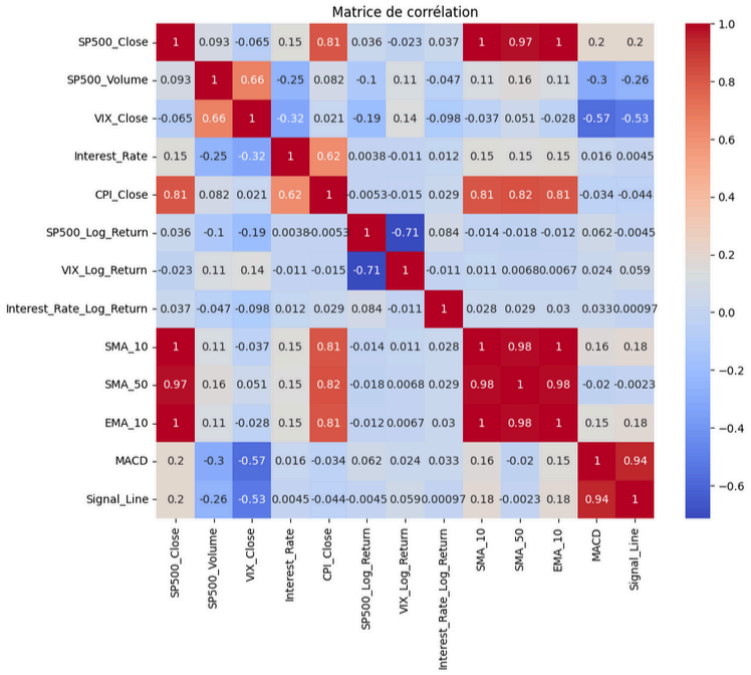
2. *Moyenne mobile de convergence / divergence (MACD)*

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 :

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

3. *Heat map*

Nous avons ensuite réalisé une heatmap afin de mettre en avant la corrélation des variables entre elles :




## Partie 2. Développement du modèle prédictif

Nous avons choisi les deux modèles de ML suivant pour prédire les rendements du S&P 500 :

1. **Régression Linéaire** : La régression linéaire a été choisie pour sa simplicité et sa capacité à modéliser des relations linéaires entre les variables. Dans le cadre de la prédiction des rendements financiers, ce modèle permet d'identifier rapidement les tendances générales sans nécessiter un volume important de ressources de calcul. De plus, il est particulièrement adapté aux données financières stationnaires, comme celles utilisées ici.

2. **Random Forest** : Le Random Forest a été sélectionné pour sa capacité à capturer des relations non linéaires et complexes dans les données. En combinant plusieurs arbres de décision, ce modèle offre une robustesse accrue face au bruit et réduit le risque de surajustement. Ce choix est justifié par la nature multi-dimensionnelle et potentiellement non linéaire des rendements du S&P 500, où des interactions complexes entre les variables exogènes (VIX, taux d'intérêt, etc.) peuvent jouer un rôle significatif.

Puis nous avons séparé les données en deux ensembles :

- **80% pour l'entraînement** des modèles.

- **20% pour les tests** afin d'évaluer la capacité de généralisation des modèles.

## Partie 3. Analyse des résultats et interprétation

1. *Performance des modèles*

Nous avons utilisé une validation croisée temporelle (rolling window) et nous avons évalué les performances des modèles grâce aux métriques **MAE, RMSE** et Accuracy. Ces métriques permettent de mesurer l'erreur des prédictions des modèles :

- **MAE** (Mean Absolute Error) mesure la moyenne des erreurs absolues entre les valeurs prédites et les valeurs réelles. Un MAE faible indique un modèle précis.
- **RMSE** (Root Mean Squared Error) mesure la racine carrée de la moyenne des carrés des erreurs. Le RMSE est plus sensible aux grandes erreurs, ce qui signifie qu'il pénalise davantage les grandes différences entre les valeurs réelles et les prédictions.
- **Accuracy** : L'Accuracy, ou taux de précision directionnelle, est une métrique qui mesure la proportion de prédictions correctement orientées par rapport aux tendances réelles. Autrement dit, elle évalue si le modèle prédit correctement les périodes de hausse ou de baisse des rendements. Cette métrique est particulièrement pertinente dans le cadre de la finance, où une prédiction directionnelle correcte peut souvent être plus critique que la précision exacte du rendement.

| **Régression linéaire** | | | |
|---|---|---|---|
| Itération | MAE | RMSE | Accuracy |
| 1 | 0.0031 | 0.0041 | 0.8243 |
| 2 | 0.0106 | 0.0171 | 0.6987 |
| 3 | 0.0064 | 0.0076 | 0.7364 |
| 4 | 0.0087 | 0.0113 | 0.7531 |
| 5 | 0.0047 | 0.0063 | 0.7197 |

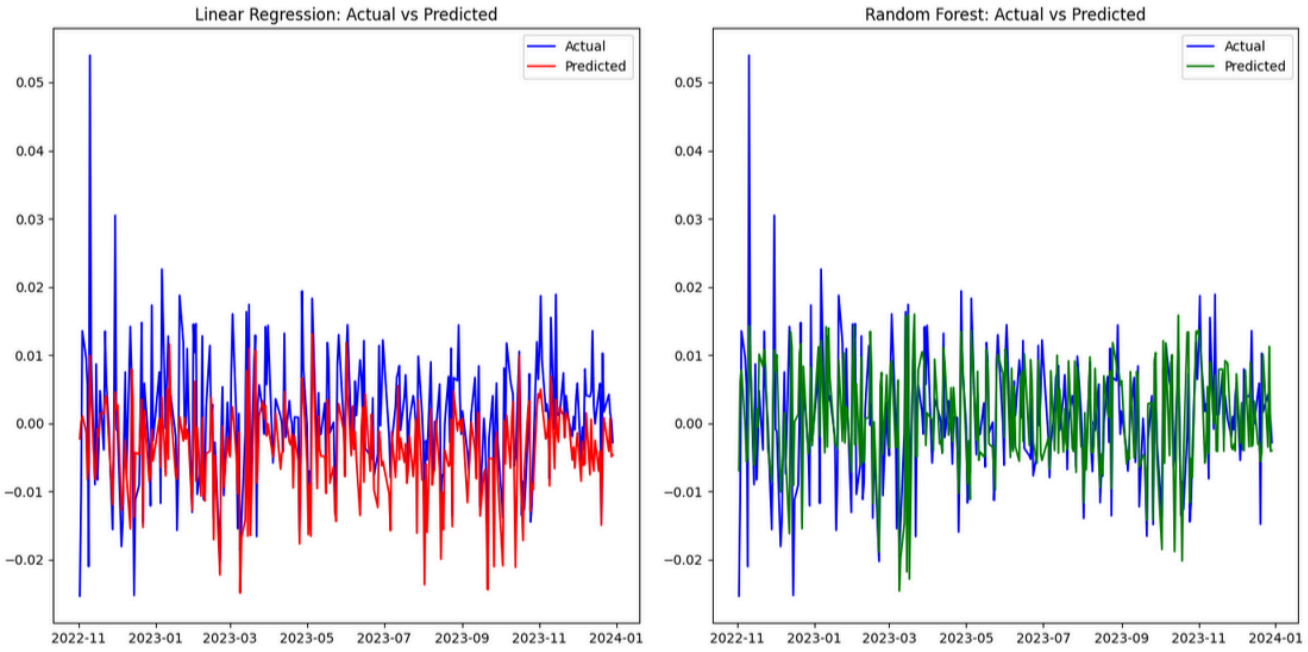
| Random forest | | | |
|---|---|---|---|
| Itération | MAE | RMSE | Accuracy |
| 1 | 0.0031 | 00043 | 0.7950 |
| 2 | 0.0097 | 0.0168 | 0.7741 |
| 3 | 0.0038 | 0.0049 | 0.7992 |
| 4 | 0.0085 | 0.0110 | 0.7322 |
| 5 | 0.0051 | 0.0065 | 0.7280 |

Interprétation :

- Première itération : la régression linéaire a une précision meilleure (+3 %) et une légère réduction du RMSE, mais les MAE sont identiques.
- Deuxième itération : Random Forest surpasse la régression linéaire en termes de MAE (-0.0009), RMSE (-0.0003), et surtout Accuracy (+7,5 %).
- Troisième itération : Random Forest est meilleure sur tous les plans avec des MAE et RMSE plus faibles et une meilleure précision directionnelle (+6,3 %).
- La régression linéaire fait légèrement mieux en termes d'Accuracy (+2 %), mais les erreurs (MAE et RMSE) sont légèrement plus élevées.
- La régression linéaire  a des erreurs légèrement plus faibles (MAE et RMSE), mais une Accuracy inférieure (-0,8 %).

Les résultats montrent que la régression linéaire a globalement des erreurs plus élevées que la Random Forest, ce qui suggère que ce modèle suit moins bien les rendements réels du S&P 500. La régression linéaire a des erreurs légèrement plus élevées, ce qui pourrait indiquer que le modèle n’a pas réussi à capturer efficacement les relations complexes dans les données ou qu’il souffre de surajustement.

Evaluation des prédictions des modèles :




Pour la régression linéaire : les valeurs réelles montrent une variabilité importante avec des pics positifs et négatifs. Les valeurs prédites par la régression linéaire sont beaucoup plus lisses et ne capturent pas bien la variabilité des données réelles. Cela indique que la régression linéaire a tendance à sous-estimer les valeurs extrêmes et ne représente pas correctement les fluctuations.

Pour Random Forest : les valeurs prédites sont plus proches des valeurs réelles, et le modèle semble mieux capturer les variations. Même si le modèle ne suit pas parfaitement toutes les fluctuations, il offre une meilleure correspondance globale comparée à la régression linéaire.

Random Forest est un meilleur choix pour ce jeu de données, probablement grâce à sa capacité à modéliser des relations non linéaires et des interactions complexes. Ce résultat est en adéquation avec les métriques étudiées précédemment.

2. *Sources d'Erreur et Pistes d'Amélioration*

**Surajustement (Overfitting)** : Le modèle **Random Forest** pourrait souffrir de surajustement, ce qui signifie qu'il a capturé des détails spécifiques aux données d'entraînement qui ne se généralisent pas bien sur les données de test. Une meilleure validation croisée et un réglage des hyperparamètres pourraient améliorer cette situation.

**Multicolinéarité et Variables Exogènes** : L'ajout de **variables supplémentaires** (telles que d'autres indicateurs économiques ou des indices boursiers internationaux) pourrait

potentiellement améliorer la précision des modèles.

**Modèles plus complexes** : D'autres modèles comme les **réseaux de neurones récurrents (LSTM)**, adaptés aux séries temporelles, pourraient mieux capturer les relations temporelles et améliorer les prédictions.

3. *Conclusions et recommandations*

Ce projet a démontré la faisabilité de la prédiction des rendements du S&P 500 à l'aide de leur corrélation aux variables exogènes telles que le VIX, les taux d'intérêt, etc. Les modèles de machine learning utilisés, notamment la régression linéaire et la Random Forest, ont permis de capturer des tendances significatives dans les données. Bien que la régression linéaire ait affiché de meilleures performances en termes de MAE et RMSE, la Random Forest reste utile pour explorer des relations plus complexes et non linéaires.

Les résultats mettent en évidence l'importance de la stationnarité des données et des dépendances temporelles dans la modélisation des séries financières. Certaines limitations, comme le risque de surajustement et la sensibilité à la volatilité, soulignent la nécessité d'améliorer les modèles par des approches plus complexes, telles que les réseaux neuronaux récurrents (LSTM). Il semble primordial de développer en parallèle de ces prédictions des indices de risques et de volatilités poussées afin de pouvoir se plier aux exigences des normes financières sur les marchés.

**Stratégies d'investissements**

- Stratégie Momentum : acheter les actions qui ont récemment été haussières et vendre celles qui ont récemment été baissières est parfaitement calibrée pour notre projet. En effet, cette stratégie propose d'exploiter l'inertie des prix. Ceci repose sur le comportement des investisseurs comme l'effet moutonnier ou encore le retard dans la réaction aux nouvelles. Cette stratégie fonctionnera très bien avec notre notre travail car elle repose notamment sur la moyenne mobile et c'est l'indicateurs que nous avons le fiable.

- Efficient Frontier : Une fois que notre programme fonctionne pour le SP500, on peut l'intégrer à tout autre indice comme le cac40 ou MSCI Emerging Market. L'idée sera ensuite de calculer les corrélations entre ces indices pour appliquer la méthode de l'efficient frontier pour obtenir la répartition de notre portefeuille. Ceci nous permettra d'obtenir un portefeuille d'actions de notre choix avec un risque et un rendement maîtrisés.