Rapport Écrit

Mathias Buée
Paul Consalès
Timothée Henriot
Marie Simon
Eliott Ollivier

Introduction

Le S&P 500 est souvent considéré comme le meilleur indicateur des performances du marché boursier américain, et par extension, un baromètre de l'économie américaine. La prédiction des prix et des rendements des actifs composant cet indice est d'une importance capitale pour les investisseurs, les gestionnaires de portefeuille, et les décideurs financiers, car elle aide à formuler des stratégies d'investissement plus informées et à minimiser les risques.

Dans le cadre de ce projet, nous visons à développer un modèle de machine learning capable de prédire les prix et les rendements des actifs du S&P 500. Ce travail s'inscrit dans une démarche académique, sous la supervision de nos professeurs de notre troisième année à l'École Centrale Marseille. Notre objectif est non seulement d'appliquer des concepts théoriques appris en cours, mais aussi de nous confronter à des données réelles et complexes pour aiguiser notre compréhension du marché et de modèles d'apprentissage.

L'application pratique de ce projet pourrait être multiple : elle permettrait aux investisseurs de mieux anticiper les mouvements de marché, aux gestionnaires de fonds d'optimiser la composition de leurs portefeuilles, et aux économistes d'analyser plus finement les tendances macroéconomiques influençant les marchés.

Les modèles que nous envisageons d'utiliser sont la régression linéaire, les Random Forests et le Gradient Boosting, chacun ayant des forces et des faiblesses en fonction des caractéristiques des données et des spécificités de notre objectif de prédiction.

Ce rapport détaillera la méthodologie adoptée pour la collecte et la préparation des données, le feature engineering, le développement et l'évaluation des modèles, ainsi que l'interprétation de nos résultats. Nous conclurons avec des recommandations basées sur nos prédictions et une discussion sur les implications potentielles de notre étude.

Méthodologie

Collecte des Données
Pour mener à bien ce projet, il était essentiel de collecter des données historiques fiables sur le S&P 500 ainsi que des données macroéconomiques pertinentes. Les données du S&P 500 ont été téléchargées via l'API de Yahoo Finance, qui fournit un accès gratuit aux données de clôture journalières, volumes de transaction, et autres indicateurs financiers. Nous avons choisi une période de cinq ans pour nos analyses, ce qui nous a permis d'avoir

une perspective à la fois récente et suffisamment longue pour identifier des tendances significatives.
En plus des données du S&P 500, des variables macroéconomiques telles que l'indice des prix à la consommation (CPI) et le produit intérieur brut (PIB) américain ont été récupérées. Ces données ont été téléchargées sous forme de fichiers CSV via des liens partagés par Google Drive, assurant que les mesures économiques étaient synchronisées avec les données de marché pour une analyse cohérente. Le PIB sera appelé “GDPC” dans notre étude et le CPI sera lui appelé “CPIAUCSL”. Nous avons extrait ces deux mesures depuis FRED (Federal Reserve Economic Data).

Préparation des Données
Une fois les données collectées, le processus de nettoyage et de préparation a été entrepris pour assurer la qualité et la précision des analyses. Cela inclut le traitement des valeurs manquantes, l'élimination des doublons, et la conversion des prix en rendements logarithmiques pour normaliser les données et réduire l'impact de la volatilité des prix élevés.
Le fusionnement des différents ensembles de données a été réalisé avec soin pour aligner les séries temporelles des données du marché avec celles des indicateurs économiques, en tenant compte des différences dans leurs fréquences de publication (quotidienne pour le S&P 500 et mensuelle ou trimestrielle pour les indicateurs économiques).

Feature Engineering
Pour enrichir notre modèle et améliorer la qualité des prédictions, divers indicateurs techniques ont été calculés à partir des données du S&P 500. Ces indicateurs incluent les moyennes mobiles, l'indice de force relative (RSI), et la convergence/divergence des moyennes mobiles (MACD). Chacun de ces indicateurs a été choisi pour sa capacité à fournir des signaux sur les tendances du marché et les points de retournement potentiel.
De plus, des analyses de corrélation et des méthodes de sélection ont été employées pour déterminer les variables les plus significatives à inclure dans le modèle final. Cela nous a permis de réduire la dimensionnalité des données tout en conservant les informations les plus critiques pour la prédiction des rendements.

Développement Initial du Modèle
Le développement du modèle a commencé par l'application de techniques de régression linéaire pour établir une base de comparaison pour les performances des modèles plus complexes. Les données ont été divisées en ensembles d'entraînement et de test selon un ratio de 80/20, et une validation croisée a été mise en place pour assurer la génération des résultats.
Les premiers résultats de cette modélisation linéaire ont servi de benchmark pour les étapes suivantes du développement de modèles, orientant les ajustements nécessaires pour aborder les modèles plus sophistiqués comme les Random Forests, qui seront discutés plus en détail dans les sections suivantes du rapport.
Ces modèles ont pu être analysés en termes de performances à l’aide de métriques adaptées, telles que le MAE (Mean Absolute Error), le RMSE (Root Mean Squared Error) ainsi que l’accuracy afin de prédire les directions de marché (hausse ou baisse).

Résultats et analyse

Une fois les modèles entraînés, il est crucial d'interpréter les résultats pour en tirer des conclusions exploitables.

Nous allons donc ici comparer les performances en identifiant le modèle le plus performant en fonction des métriques choisies et nous allons vérifier la présence d'éventuels écarts de prédictions entre les valeurs réelles et les valeurs de notre modèle. Il sera nécessaire d'identifier les sources de ces erreurs (sous-ajustement, surajustement, limitations liées aux données). Nous proposerons des solutions pour minimiser ces écarts. Enfin, nous analyserons les résultats du point de vue de la prise de décision d'investissement.

Analyse et comparaison des performances des modèles

Les modèles évalués sont la régression linéaire et le Random Forest. Ces modèles ont été évalués à l'aide des métriques suivantes :

MAE (Mean Absolute Error) : Mesure la moyenne des erreurs absolues, donnant une idée de l'écart moyen entre les prédictions et les valeurs réelles.

RMSE (Root Mean Squared Error) : Évalue les erreurs avec un poids plus important sur les grandes déviations, offrant une vision plus stricte des performances.

R^2 (coefficient de détermination) : Reflète la proportion de la variance expliquée par le modèle, où une valeur proche de 1 indique une excellente correspondance.

Les résultats obtenus montrent les performances suivantes :

	Modèle	MAE	RMSE	R^2
0	Régression Linéaire	0.001830	0.001873	0.922304
1	Random Forest	0.000582	0.001873	0.977799

La Random Forest se distingue par un MAE nettement inférieur à celui de la régression linéaire, démontrant une meilleure précision des prédictions.

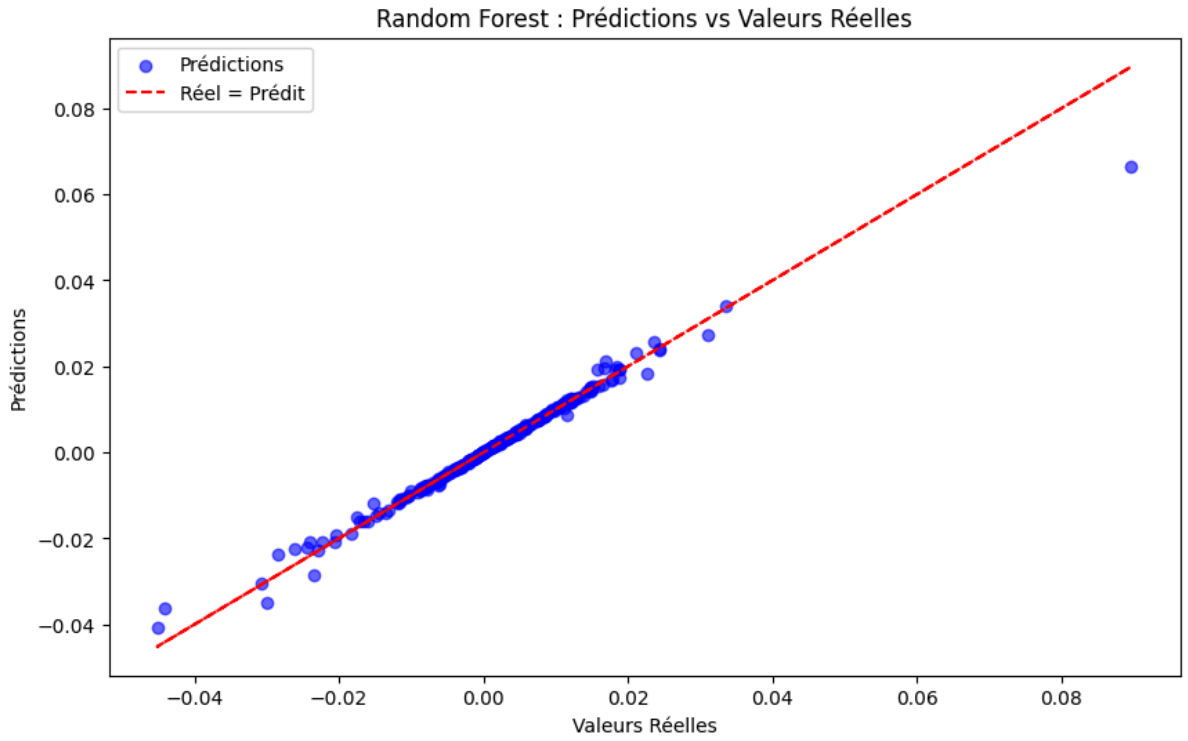
Le RMSE des deux modèles est équivalent, mais cela peut être dû à des erreurs systématiques non corrigées.

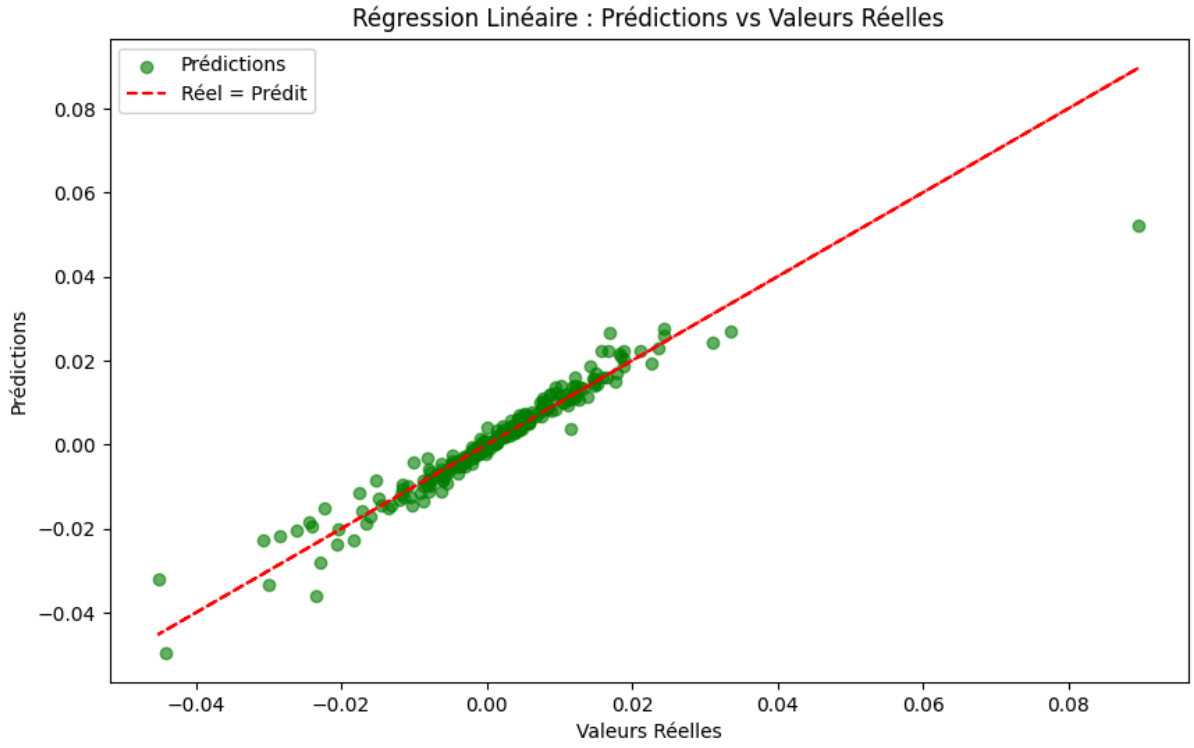
Le R^2 montre que les deux modèles expliquent une grande part de la variance des données, la Random Forest ayant une performance légèrement supérieure.

Visualisation des résultats

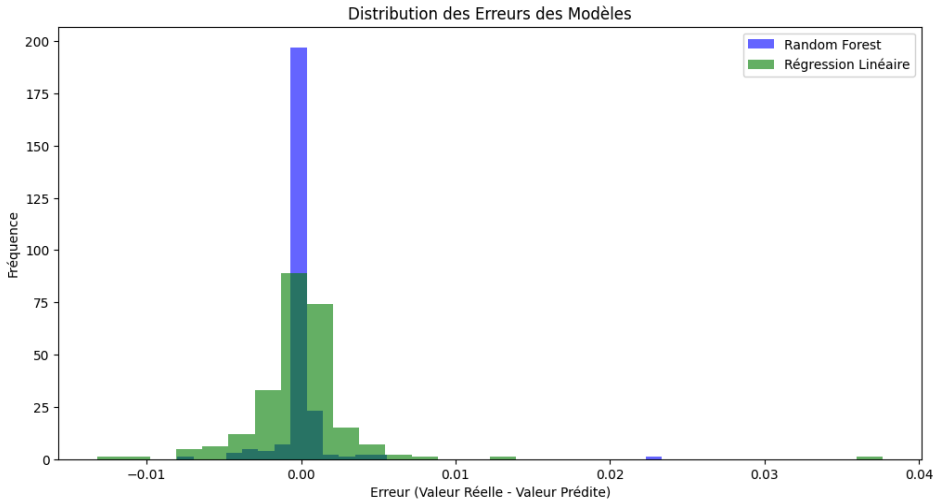
Graphique des prédictions vs valeurs réelles :

Les prédictions de Random Forest suivent de manière plus étroite les valeurs réelles, avec moins de dispersion autour de la ligne idéale (y = x).


Random Forest : Prédictions vs Valeurs Réelles
Prédictions
Réel = Prédit
Prédictions
Valeurs Réelles
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08


Régression Linéaire : Prédictions vs Valeurs Réelles
Prédictions
Réel = Prédit
Prédictions
Valeurs Réelles
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08

Distribution des erreurs :



Random Forest produit une distribution d'erreurs bien centrée autour de zéro avec une variance plus faible par rapport à la régression linéaire.

Discussion des erreurs, des limites et des améliorations envisageables

Nous allons maintenant analyser les résultats et discuter des éventuelles sources d'erreurs de nos modèles. Les performances des modèles, bien qu'acceptables, peuvent être influencées par plusieurs facteurs dont notamment :

- La Qualité des données : les données financières (S&P 500, PIB, inflation) peuvent parfois être bruitées et cela réduit la précision.
- La complexité des relations économiques : Les relations entre les variables explicatives (PIB, inflation) et la variable cible (rendements) peuvent être influencées par d'autres facteurs exogènes que ceux pris en compte. En effet, les annonces économiques ou les chocs externes sont des éléments qui peuvent agir sur ces relations. Les prendre en compte pourrait donc être une source d'amélioration si nous voulions utiliser notre modèle au jour le jour.
- Un surajustement dans le Random Forest : Avec un nombre d'arbres élevé et des profondeurs importantes, le modèle peut s'adapter trop spécifiquement aux données d'entraînement, ce qui risque de poser un problème si nous voulions l'utiliser pour d'autres datasets.

Ces sources d'erreurs prises en compte, il est ensuite nécessaire d'étudier les limites de nos modèles.

Pour la régression linéaire, le modèle peut s'avérer incapable de capturer les relations non linéaires entre les variables explicatives et la cible et il peut être susceptible à la multicolinéarité, rendant les coefficients instables.

La Random Forest quant à elle est un modèle qui, bien qu'efficace pour des relations non linéaires, est moins interprétable, rendant difficile l'identification des relations spécifiques entre les variables.

Enfin, lorsque l'on s'intéresse à la distribution temporelle des erreurs, l'analyse des erreurs résiduelles montre que la régression linéaire génère des pics d'erreurs significatifs autour

des périodes de forte volatilité. La Random Forest réduit ces pics, mais certaines erreurs subsistent, notamment lors des retournements de tendance rapide.

Il est cependant possible d'envisager des améliorations pour les modèles. Pour la Random Forest, nous pouvons agir sur les hyperparamètres et effectuer une recherche en grille pour optimiser les paramètres tels que le nombre d'arbres, la profondeur maximale, et le critère d'évaluation. Il est également envisageable de tester des variantes comme Gradient Boosting ou XGBoost, qui peuvent offrir des améliorations supplémentaires. En termes d'enrichissement des données, il est concevable d'ajouter de nouvelles variables exogènes et d'intégrer des indicateurs avancés tels que les flux de capitaux, les données sectorielles, ou les annonces macroéconomiques. Nous pouvons aussi réduire le bruit en utilisant des techniques de débruitage pour améliorer la qualité des données d'entrée, par exemple via des filtres ou des transformations spécifiques.
Il reste enfin évidemment l'option d'utiliser des modèles alternatifs comme le LSTM (Long Short-Term Memory) pour capturer les dépendances temporelles complexes dans les séries historiques ou d'exploiter davantage la Ridge Regression (pour la régression linéaire), afin de limiter les effets de la multicolinéarité.

Conclusion et recommandations

Ce projet avait pour objectif d'explorer les capacités de prédiction des modèles de machine learning appliqués aux rendements et prix des actifs du S&P 500, en intégrant des variables macroéconomiques et techniques. À travers une méthodologie rigoureuse de collecte, préparation, modélisation et analyse, nous avons montré que le Random Forest est le modèle le plus performant, grâce à sa capacité à capturer des relations non linéaires complexes dans les données. D'après nos résultats, elle a affiché une erreur moyenne absolue (MAE) inférieure à celle de la régression linéaire, et un coefficient de détermination (R^2) supérieur. La régression linéaire, quant à elle, est interprétable mais s'est avérée limitée dans le contexte de notre étude, notamment en raison de son incapacité à modéliser des relations non linéaires. Elle a montré des performances acceptables mais inférieures à celles du Random Forest. De plus, les données exogènes évoquées (l'inflation et le PIB) ont enrichi la modélisation, mais d'autres variables pertinentes pourraient encore améliorer les prédictions. Enfin, l'étude des métriques de performance (MAE, RMSE, R^2) ont permis de comparer objectivement les modèles, et les visualisations des erreurs résiduelles ont montré que des ajustements supplémentaires sont possibles.

Sur la base de ces résultats, nous sommes en mesure de proposer des recommandations pour améliorer les performances des modèles. Comme nous l'avons expliqué dans les solutions, il est possible d'optimiser les modèles en jouant avec des hyperparamètres ou même d'utiliser de nouveaux modèles (Gradient Boosting ou réseaux neuronaux). Il est possible d'enrichir les données (données sectorielles, événementielles, variables macroéconomiques en plus etc). L'automatisation est aussi à travailler avec la possibilité de mettre en place un flux de travail automatisé pour la collecte des données, le prétraitement, l'entraînement des modèles et la génération de prédictions en temps réel. Enfin, pour valider et généraliser les résultats, il est bien de penser à une validation croisée temporelle rigoureuse en testant les modèles sur des périodes de validation qui reflètent la volatilité

historique du marché et d’évaluer les données externes en testant les modèles sur des indices boursiers différents (par exemple, le NASDAQ ou le Dow Jones) pour évaluer leur robustesse.