Centrale Méditerranée

Données et Décision Financière et Économique – 2025

PROJET DATA

MID TERM - Prédiction des Prix et Actifs du S&P 500 grâce à du Machine Learning simple

COMASSETO Vitória

ITURRALDE Martín

URTUBIA Carolina

VIOLA CARVALHO Henrique

Marseille, Janvier 2025

1. Introduction

Dans le domaine de la finance, anticiper le comportement futur des marchés est essentiel pour prendre des décisions stratégiques éclairées. Ce projet vise à appliquer des techniques de machine learning pour prédire les prix et les rendements de l'indice S&P 500, qui reflète la performance de 500 des plus grandes entreprises aux États-Unis. Cet indice est considéré comme un indicateur clé d'un marché hautement efficace et complexe.

Nous exploitons des données historiques du S&P 500 sur les cinq dernières années, incluant les prix de clôture et les volumes de transactions. Ces données sont enrichies par des variables exogènes, telles que l'indice de volatilité (VIX), les taux d'intérêt et divers indicateurs macroéconomiques (comme la taux de chômage). Ces informations seront soigneusement traitées et transformées afin d'assurer leur qualité, pertinence et intégration optimale dans les modèles d'analyse. Les données seront ensuite divisées en deux ensembles : un pour l'entraînement des modèles et un autre pour tester et valider les prédictions.

L'approche méthodologique inclut l'utilisation de modèles variés, allant des méthodes classiques comme la régression linéaire et le modèle ARIMA, à des algorithmes avancés comme les forêts aléatoires (Random Forest) ou les réseaux de neurones LSTM. Ces modèles seront évalués avec des métriques de performance telles que l'erreur absolue moyenne (MAE) et la racine carrée de l'erreur quadratique moyenne (RMSE), garantissant ainsi une analyse rigoureuse de leur précision.

Au-delà de la simple prédiction, ce projet cherchera à interpréter les résultats obtenus afin de formuler des recommandations d'investissement. L'objectif est de combiner des outils analytiques modernes et puissants avec des problématiques financières concrètes, offrant des solutions pratiques et des perspectives innovantes dans le domaine de la gestion des marchés financiers.

2. Objectifs

Les objectifs généraux de ce projet sont de développer un modèle prédictif basé sur le machine learning pour estimer les prix et les rendements de l’indice S&P 500, en intégrant des données historiques et des variables exogènes. Ce modèle vise également à générer des recommandations d’investissement pertinentes et précises, en se basant sur les résultats produits par le programme de machine learning.

Pour atteindre ces objectifs, plusieurs étapes spécifiques sont prévues. La première consiste à collecter et traiter des données pertinentes, incluant les prix historiques, les indicateurs techniques et les variables macroéconomiques. Ensuite, différents modèles prédictifs seront mis en œuvre et leurs performances comparées, tels que la régression linéaire, le modèle ARIMA, les forêts aléatoires (Random Forest) et les réseaux de neurones LSTM. Une analyse approfondie des résultats obtenus

sera réalisée pour identifier les sources d'erreur éventuelles et proposer des axes d'amélioration afin d'optimiser les performances du modèle.

En complément, ce projet s'inscrit dans une démarche visant à combiner une approche quantitative rigoureuse avec une interprétation qualitative des résultats. L'objectif est de développer des outils d'aide à la décision fiables. Les modèles créés permettront non seulement de prévoir avec précision les tendances du S&P 500, mais aussi d'identifier les facteurs exogènes les plus influents, offrant ainsi une meilleure compréhension des dynamiques du marché financier. Enfin, les recommandations finales mettront un accent particulier sur la gestion des risques et l'identification d'opportunités d'investissement.

3. Méthodologie et Analyse

3.1. Partie 1 : Collecte et Préparation des Données

La première étape de la méthodologie consiste à collecter et préparer les données nécessaires pour construire les modèles de Machine Learning. Cette phase est cruciale, car la qualité et la pertinence des données influencent directement la précision et l'efficacité des prédictions.

Le modèle prédictif est structuré autour de deux types de variables. La variable dépendante, qui constitue la cible principale à prédire, est le rendement logarithmique de l'indice S&P 500. Les variables indépendantes, quant à elles, se divisent en deux catégories principales : les indicateurs techniques et les variables exogènes. Les indicateurs techniques comprennent des outils tels que les moyennes mobiles simples (SMA), l'indice de force relative (RSI) et le MACD (Moving Average Convergence Divergence). Ces indicateurs ont été choisis pour leur capacité à identifier des schémas ou signaux concernant les tendances futures possibles du marché, des aspects souvent invisibles à partir des seuls prix ou rendements historiques. Les variables exogènes, en revanche, incluent des facteurs externes tels que le taux d'intérêt et l'indice de volatilité (VIX). Ces variables offrent une perspective élargie en intégrant des éléments économiques extérieurs qui influencent significativement le comportement du S&P 500.

Les indicateurs techniques et les variables exogènes jouent des rôles complémentaires. Les premiers permettent d'analyser le comportement du marché en se basant sur ses mouvements historiques, tandis que les seconds apportent une vue d'ensemble des dynamiques macroéconomiques. Par exemple, le taux d'intérêt peut influencer les investissements institutionnels, tandis que l'indice de volatilité reflète les anticipations des acteurs du marché face aux risques.

Les données nécessaires pour le projet sont extraites de plusieurs sources fiables. Les informations sur l'indice S&P 500, comme la valeur de clôture journalière et le volume d'actions échangées, sont récupérées via l'API de Yahoo Finance. Ces données fournissent une base essentielle pour analyser l'activité du marché. En parallèle, les variables exogènes, telles que l'inflation, le taux de chômage et les démarrages de constructions de logements, sont collectées à partir de bases de

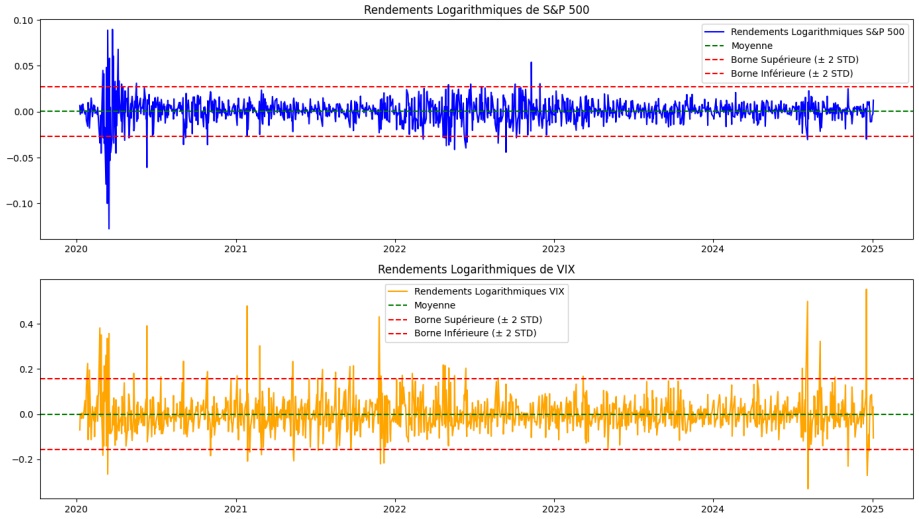
données économiques telles que FRED (Federal Reserve Economic Data). Ces données externes permettent d'enrichir le modèle en intégrant des indicateurs macroéconomiques pertinents.

Après la collecte, les données brutes subissent un traitement rigoureux pour garantir leur qualité. Cela inclut le resampling des données journalières à une fréquence mensuelle, si nécessaire, afin de maintenir une cohérence temporelle entre les différentes séries. Les doublons et les données manquantes sont supprimés pour assurer l'intégrité des informations. Enfin, les données sont fusionnées dans un jeu de données unique, alignant les dates et intégrant les variables dépendantes et indépendantes. Cette consolidation permet une exploitation optimale des données par les modèles de Machine Learning. Cette approche garantit une base solide pour la modélisation, en intégrant à la fois les caractéristiques intrinsèques du marché et les facteurs externes qui influencent son comportement.

3.2. Partie 2 : Développement du Modèle Prédictif

Les données historiques du S&P 500, du VIX et des taux d'intérêt ont été collectées via Yahoo Finance pour une période de 5 ans. Elles ont été transformées en rendements logarithmiques (log-returns) pour stationnariser les séries temporelles, une condition nécessaire pour plusieurs modèles, notamment ARIMA. Des indicateurs techniques (SMA, RSI, MACD) ont été ajoutés pour enrichir les données et fournir des informations supplémentaires sur les tendances du marché. Les données ont ensuite été divisées en ensembles d'entraînement (80 %) et de test (20 %).

Graphique 1: Évolution des Rendements Logarithmiques du S&P 500 et du VIX avec Bornes de Confiance



3.2.1. Régression Linéaire

La régression linéaire est un modèle statistique qui établit une relation entre une variable cible (dépendante) et une ou plusieurs variables explicatives (indépendantes) en ajustant une droite. Elle est

utilisée pour prédire des valeurs continues (comme les rendements financiers) et analyser l'impact de chaque variable explicative sur la cible.

La régression linéaire présente des résultats satisfaisants avec un MAE faible (0.0045), montrant que les prédictions sont globalement proches des valeurs réelles, et un RMSE modéré (0.0057), qui met davantage l'accent sur les grandes erreurs. Avec une précision de 76.89 %, le modèle est capable de prédire correctement la direction des rendements dans une majorité des cas. Toutefois, malgré sa simplicité et son efficacité pour les tendances générales, il reste limité dans la capture des relations complexes ou non linéaires.

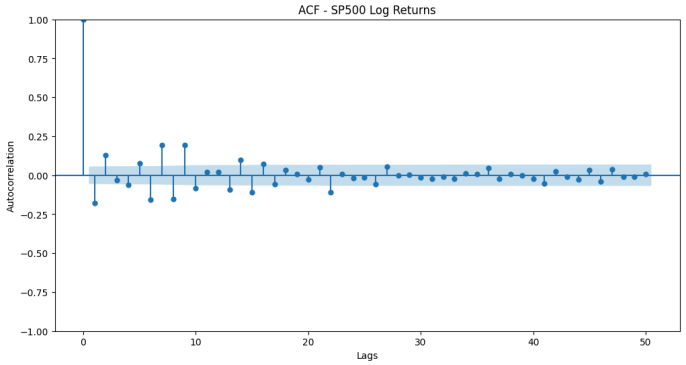
3.2.2. ARIMA (AutoRegressive Integrated Moving Average)

Le modèle AutoRegressive Integrated Moving Average (ARIMA) est utilisé pour analyser et prédire les séries temporelles stationnaires en capturant les relations auto-régressives, les différences pour stationnariser les données et les moyennes mobiles. Ce modèle est particulièrement adapté pour identifier des tendances linéaires à court terme et prévoir des valeurs futures.

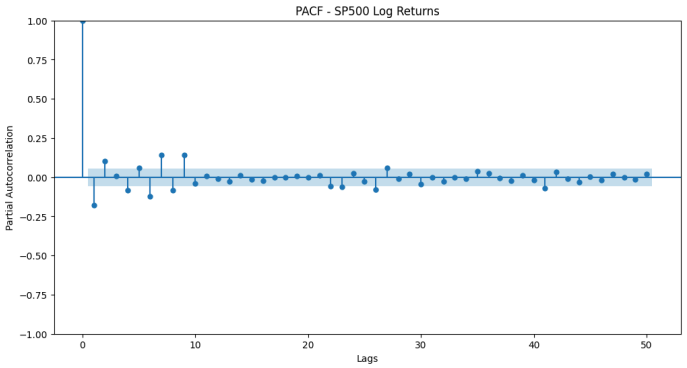
Pour modéliser efficacement une série temporelle, il est essentiel d'identifier les décalages temporels pertinents (lags) qui capturent les relations entre les valeurs passées et actuelles. L'analyse des fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) permet de déterminer ces lags. Les lags pertinents sont ceux dont les valeurs dépassent l'intervalle de confiance (zone ombragée en bleu sur les graphiques ACF et PACF). Ces lags sont utilisés pour déterminer les paramètres du modèle ARIMA : p (terme auto-régressif, AR) et q (terme de moyenne mobile, MA).

Pour déterminer les paramètres du modèle ARIMA dans ce projet, les graphiques ACF et PACF ont été analysés. Le graphique ACF montre une autocorrélation significative au premier lag, suivie d'une diminution rapide des valeurs vers l'intervalle de confiance. Cela suggère un p = 1 pour le terme auto-régressif. Le graphique PACF présente un comportement similaire, avec un premier lag significatif suivi d'une chute rapide des valeurs. Ces observations confirment l'utilisation d'un modèle ARIMA(1, 0, 1), avec p = 1 et q = 1, car les autres lags n'indiquent pas de dépendance forte. Ce choix de paramètres permet de capter les relations temporelles tout en maintenant une simplicité dans la structure du modèle.

Graphique 2: ACF - SP500 Log Returns



Graphique 3: PACF - SP500 Log Returns



Le modèle ARIMA présente un MAE de 0.0059 et un RMSE de 0.0081, indiquant des erreurs légèrement plus élevées que la régression linéaire. Avec une accuracy de 56.13%, il montre des performances faibles pour prédire correctement les directions (hausse/baisse). Cela reflète sa limite à capturer les fluctuations complexes des rendements. ARIMA est plus adapté aux séries stationnaires et moins efficace pour ces données volatiles.

3.2.3. Random Forest

Random Forest est un modèle d'apprentissage supervisé d'ensemble basé sur de multiples ensembles d'arbres de décision. Plusieurs arbres sont créés indépendamment, sur la base de sous-échantillons de données, et leurs prédictions sont combinées soit par la moyenne, soit par le vote. Ce modèle est idéal pour capturer les relations non linéaires et éviter les risques de surajustement.

Dans ce projet, le modèle Random Forest a été entraîné sur les données d'entraînement (X_train, y_train) pour établir une relation entre les variables explicatives et les rendements log du S&P 500. Une fois entraîné, le modèle a généré des prédictions (y_test_pred_rf) sur l'ensemble de test (X_test). Les performances ont été mesurées à l'aide des métriques suivantes :

Table 1: MAE, RMSE et Accuracy du Random Forest

MAE (Mean Absolute Error)	0.0041	Précision globale satisfaisante
RMSE (Root Mean Squared	0.0052	Accent sur les grandes erreurs

Error)		
Accuracy	75.47 %	Le modèle prédit correctement la direction des rendements dans la majorité des cas

3.2.4. LSTM

Les LSTM sont des réseaux neuronaux récurrents développés pour apprendre les dépendances temporelles dans les séries chronologiques. Elles peuvent traiter efficacement des données séquentielles car elles sont conçues pour mémoriser des informations sur une très longue période. Ils peuvent servir de modèle idéal pour détecter des modèles et des tendances complexes dans des séries temporelles volatiles.

Le code configure un modèle LSTM pour analyser les données temporelles et prédire les rendements log du S&P 500. Il ajoute une couche LSTM avec 50 neurones pour capturer les dépendances temporelles, suivie d'une couche dense pour produire une seule sortie. Le modèle utilise l'optimiseur Adam et la fonction de perte MSE pour minimiser les erreurs.

Le modèle LSTM produit un MAE de 0,0055 et un RMSE de 0,0070, deux erreurs modérées mais légèrement plus élevées que celles de Random Forest. La précision de 77,35 % indique également une bonne performance dans la prédiction de la direction des rendements (haut/bas), bien qu'un peu plus faible par rapport à Random Forest. Cela signifie que le modèle réussit à capturer des dépendances temporelles complexes, mais que la sensibilité aux données peut être l'une des raisons pour lesquelles les erreurs sont légèrement plus élevées.

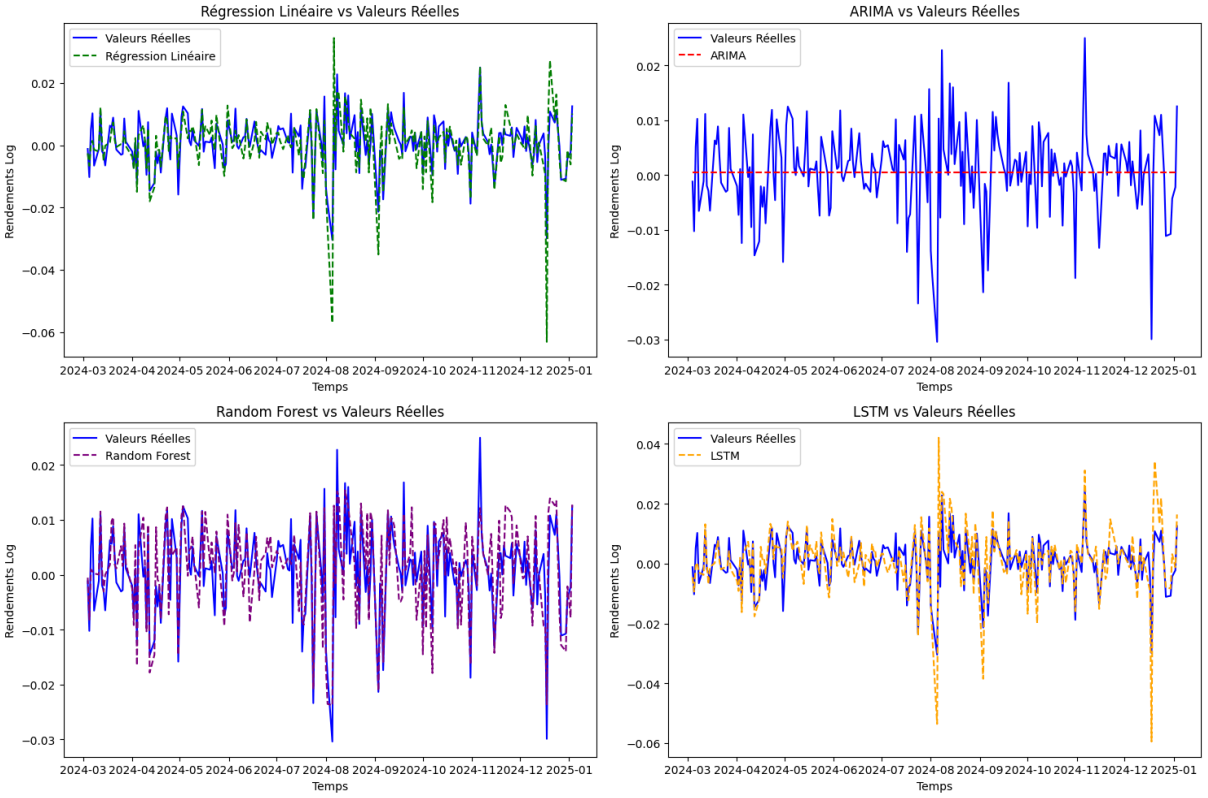
4. Analyse des Résultats

Les résultats montrent que la régression linéaire est simple mais fiable, avec une MAE et une RMSE modérées et une précision de 76,88 %, tandis que l'ARIMA a les performances les plus médiocres avec la MAE et la RMSE les plus élevées et la précision la plus faible de seulement 56,13 %, ce qui n'est pas adapté à ces données. Random Forest est en effet le modèle le plus performant avec des MAE et RMSE très faibles, tout en obtenant une précision de 77,83 %, ce qui permet de capturer efficacement les relations non linéaires. Le modèle LSTM l'emporte avec une bonne capacité à capturer les dépendances temporelles. Ses MAE et RMSE sont légèrement plus élevés, et donc sa précision de 77,35 % légèrement inférieure. Random Forest et LSTM s'imposent comme les meilleurs choix pour cette analyse.

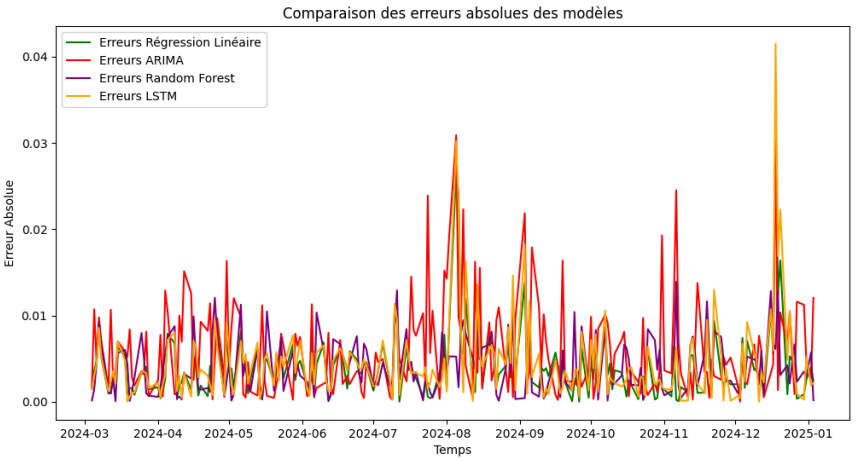
Table 2: Comparaison entre les modèles

```
Comparaison des Modèles:
                Modèle       MAE      RMSE  Accuracy
0  Régression Linéaire  0.004050  0.005749  0.768868
1                ARIMA  0.005988  0.008109  0.561321
2        Random Forest  0.004111  0.005183  0.754717
3                 LSTM  0.004458  0.006424  0.778302
```

Graphique 4: Modèles vs valeurs réelles



Graphique 5: Comparaison des erreurs absolues des modèles



Ce graphique compare les erreurs absolues des prédictions de différents modèles (Régression Linéaire, ARIMA, Random Forest, et LSTM) par rapport aux rendements réels du S&P 500. Chaque courbe représente l'écart entre les prédictions et les valeurs réelles pour un modèle spécifique. Les fluctuations montrent comment les modèles se comportent dans des conditions de volatilité variable, avec des pics indiquant des périodes où les erreurs sont les plus importantes. La Régression Linéaire et Random Forest montrent des erreurs plus faibles et stables, tandis que LSTM et ARIMA présentent des variations plus marquées à certains moments.

Conclusion

Le projet a démontré l'efficacité de divers modèles de prédiction appliqués aux rendements log du S&P 500. La régression linéaire, bien qu'elle soit un modèle simple, a offert une précision modérée avec une MAE faible, mais reste limitée face à des relations non linéaires. ARIMA, conçu pour les séries temporelles stationnaires, a montré des performances médiocres en raison de la volatilité des données. En revanche, Random Forest et LSTM se sont distingués par leur capacité à capturer des relations complexes et à offrir des prédictions précises, avec des MAE et RMSE compétitifs. Random Forest s'est avéré idéal pour les dépendances non linéaires, tandis que LSTM excelle dans l'analyse des dépendances temporelles.

Cette analyse montre que le choix du modèle dépend des caractéristiques des données et des objectifs spécifiques. Pour les marchés volatiles comme le S&P 500, des modèles capables de capturer des dynamiques complexes, comme Random Forest et LSTM, sont essentiels. Enfin, ce travail met en lumière l'importance d'une préparation minutieuse des données et de l'interprétation des résultats pour formuler des stratégies d'investissement robustes et informées.