Alexandra Gomes Félix Robotti Yohann le Couster
Fanny Lamothe Margaux Carron

3A - DDEFI

PROJET DATA

Rapport_TD2_Prédiction_S&P

Table des matières

Alexandra Gomes Félix Robotti Yohann le Couster
Fanny Lamothe Margaux Carron

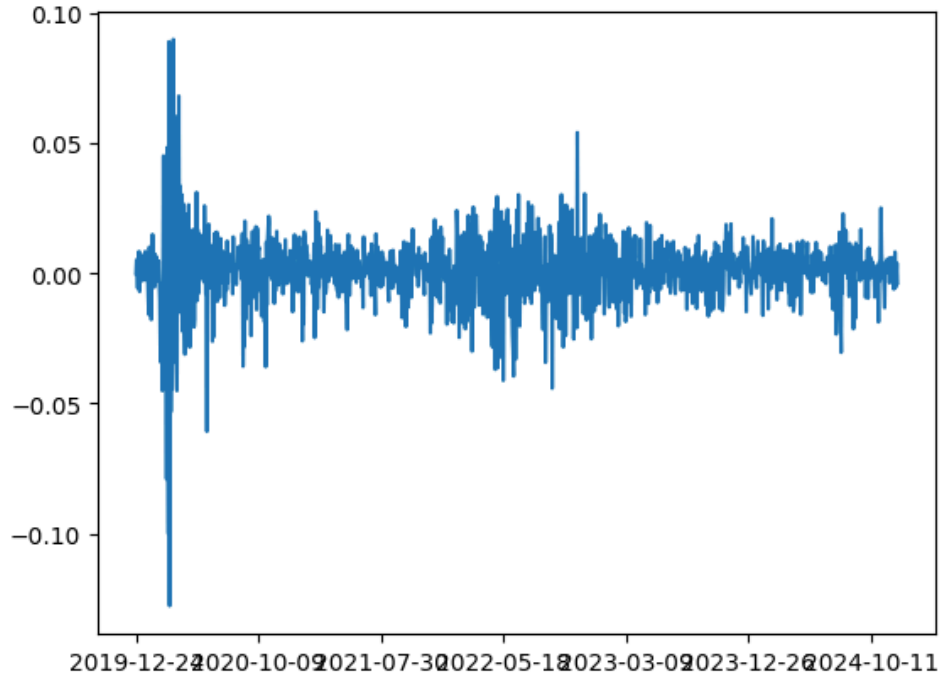
Partie 1 : Collecte et Préparation des Données

1. Collecte des données

SP55_Close : prix de clotûre du S&P 500
SP500_Volume : Volume de transaction
VIX_Close : Valeur de l'indice VIX (volatilité)

2. Prétraitement des données

Le programme nous rend cette figure pour le Log-Returns du S&P500 :



Le Log-return est une mesure utilisée pour calculer les rendements d'une série financière. On observe que le log-returns oscille autour de 0, ce qui est caractéristique des rendements financiers.
On observe aussi deux périodes de "cluster" de volatilité : début 2020 et milieu 2022.

```
ADF Statistic: -10.646993319983778
p-value: 4.767476993501795e-19
La série est stationnaire (p-value <= 0.05)
L'ordre optimal pour ARIMA est: (3, 2) avec AIC = -7263.139674467109
```

Pour le test ADF, nous obtenons les informations ci-dessus.

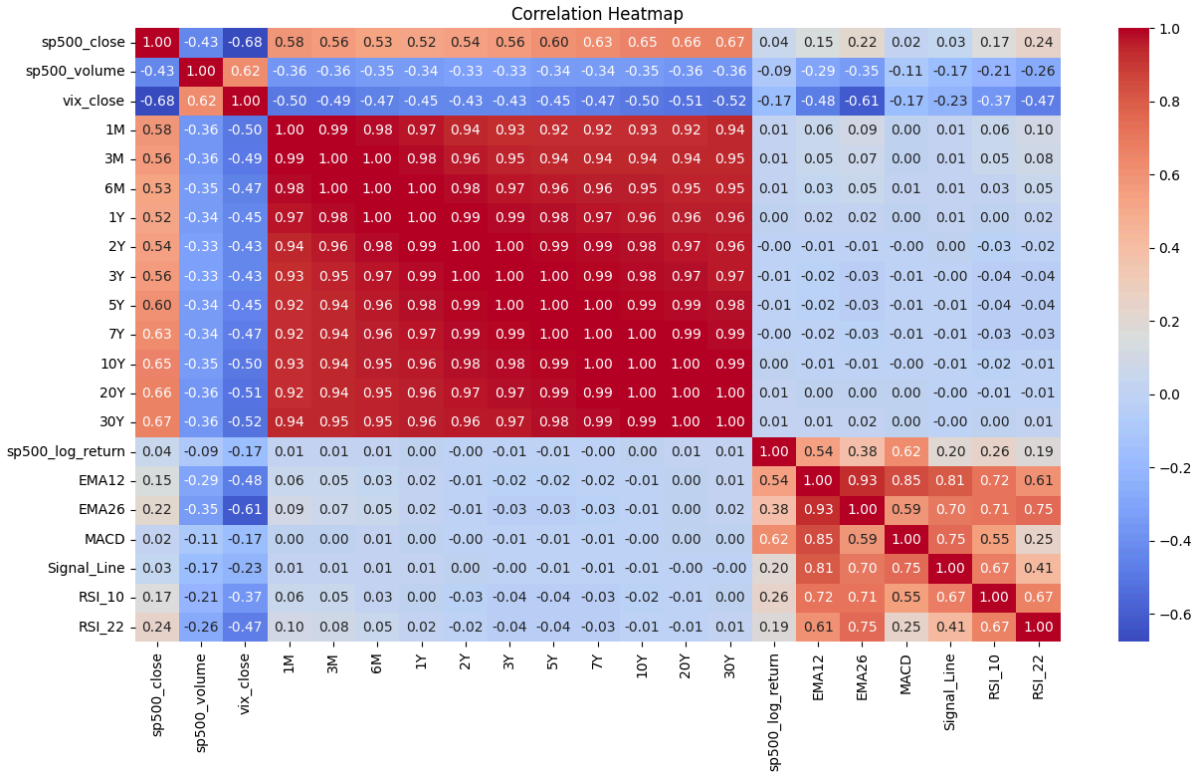
Le test ADF permet de vérifier si une série temporelle est stationnaire. En réalité, on voyait déjà que les valeurs oscillaient autour de 0 donc on avait présence d'une forme de stationnarité.

Le test ADF couplé à la p-value confirme que la série est stationnaire.

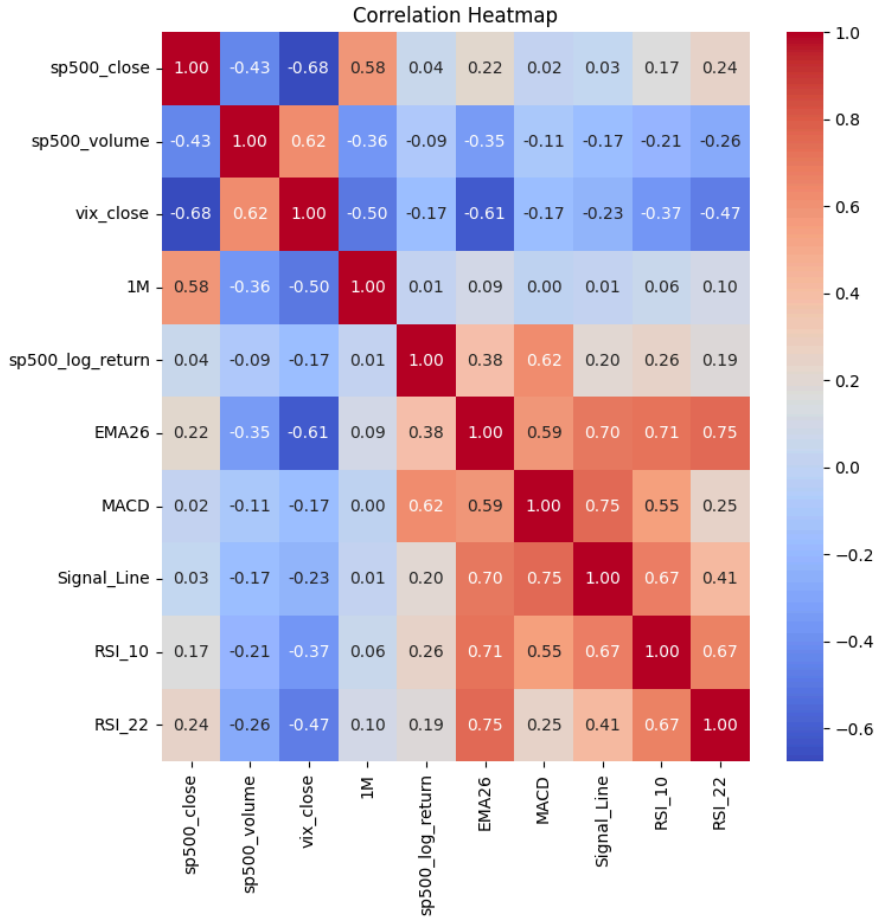
AIX : Akaike Information Criterion, est une mesure utilisée pour évaluer la qualité d'un modèle statistique, une valeur négative est un bon signe, nous avons AIC = -7263.14 donc probablement un bon modèle ARIMA(3,2).

3. Feature engineering

Nous obtenons une figure de corrélation comme celle-ci :



Toutes les colonnes de taux d'intérêts sont très corrélées entre elles, nous décidons de ne garder que les taux à un mois. Les deux colonnes EMA sont également très corrélées donc nous ne gardons que EMA26.



Tout d'abord, nous pouvons observer qu'il y a une corrélation importante (0,67) entre le RSI10 et le RSI22, ce qui est attendu et expliqué par le fait que le RSI10 correspond au Relative Strength Index (indice compris entre 0 et 100 qui évalue si une action est en situation de surachat, de survente ou non) sur les 10 derniers jours d'ouverture du marché, et le RSI22 à ce même indice mais sur les 22 derniers jours d'ouverture du marché. Il est donc évident que le RSI22 dépend en partie du RSI10, d'où le fait qu'une corrélation élevée est attendue.

Ensuite, nous pouvons observer que le VIX_Close est inversement corrélé au S&P500_Close. Ceci s'explique par le fait qu'une plus haute volatilité permet de plus grands mouvements de marché ; les moments où les mouvements ont été les plus grands dans la période observée correspond à la crise du COVID-19, où les marchés financiers ont plongé. Il est donc normal qu'ici, une volatilité qui augmente corresponde à un S&P500 plus bas.

Enfin, nous pouvons observer une corrélation négative entre le VIX_Close et l'EMA26 (Exponential Moving Average 26, qui correspond à la moyenne glissante du cours du S&P500 sur les 26 dernières périodes). Ceci est dû à des

causes similaires à ce que nous venons d'expliquer, car en effet, sur la période étudiée, les moments de croissance du marché correspondent à des périodes de stabilité, et donc, de faible volatilité.

Partie 2 : Développement du Modèle Prédictif

1. Choix des modèles

Nous choisissons comme modèle la Régression Linéaire et Random Forest. La régression linéaire est un modèle simple mais efficace pour établir des relations linéaires, ce modèle est un bon point de référence pour évaluer les performances d'un modèle plus complexe. Random Forest est un algorithme non linéaire basé sur des arbres de décision.

2. Entraînement et validation

Nous créons un backtest pour mesurer les performances des deux modèles grâce à deux indicateurs : **MAE** et **RMSE**.

- Mean Absolute Error (**MAE**) évalue l'erreur moyenne absolue des prédictions par rapport aux valeurs réelles
- Root Mean Squared Error (**RMSE**) mesure la racine carrée de l'erreur quadratique moyenne, donnant plus de poids aux grandes erreurs

Nous nous baserons donc sur ces deux indicateurs pour valider les modèles.

Comparaison du **MAE** : La régression linéaire a un MAE légèrement plus bas 0.00805 que le Random Forest 0.00851, donc en moyenne, les prédictions de la régression linéaire sont légèrement plus proches des valeurs réelles que celles du Random Forest.

Comparaison du **RMSE** : Le RMSE est légèrement plus élevé pour le Random Forest 0.01079 par rapport à la régression linéaire 0.01018, donc le Random Forest commet des erreurs plus importantes sur certains points.

Les écarts de valeurs sont faibles, les deux modèles se valent, même si nos résultats montrent une légère performance pour la régression linéaire.

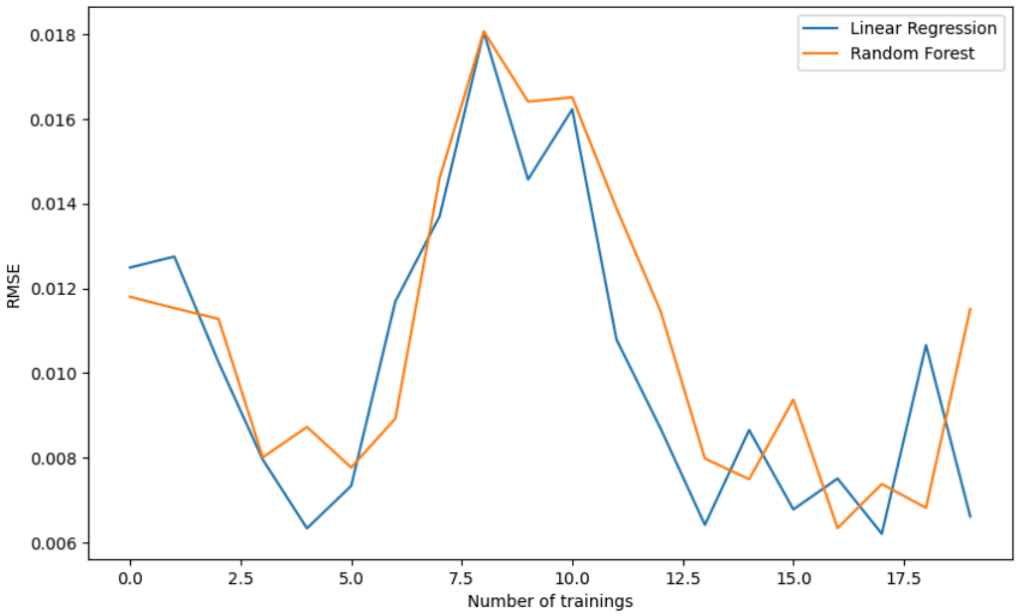
Partie 3 : Analyse des Résultats et Interprétation

1. Analyse des performances

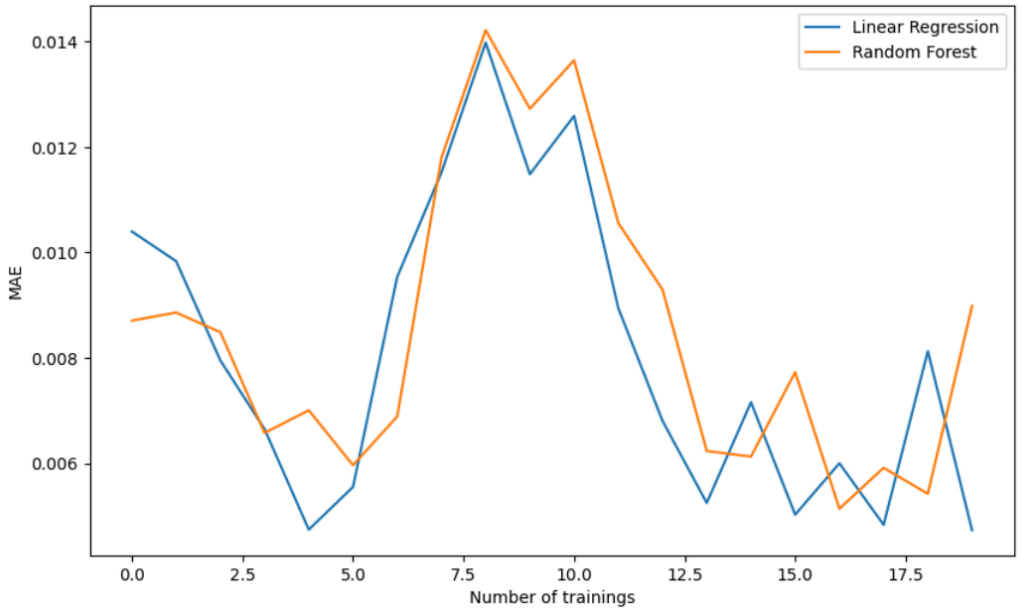
Afin de comparer les performances des modèles, nous avons tracé les graphiques d'évolution des erreurs des deux modèles et les comparaisons des prédictions avec les vraies valeurs.

En observant d'abord les graphes de MAE et RMSE selon le nombre d'entraînement des deux modèles, on constate que les deux courbes suivent à chaque fois la même tendance globale avec des valeurs légèrement plus élevées pour la *Random Forest* qui cumule des prédictions moins proches du réel que le modèle linéaire (MAE) avec des erreurs sur des points spécifiques (RMSE). Les deux modèles se valent d'après ces deux métriques.

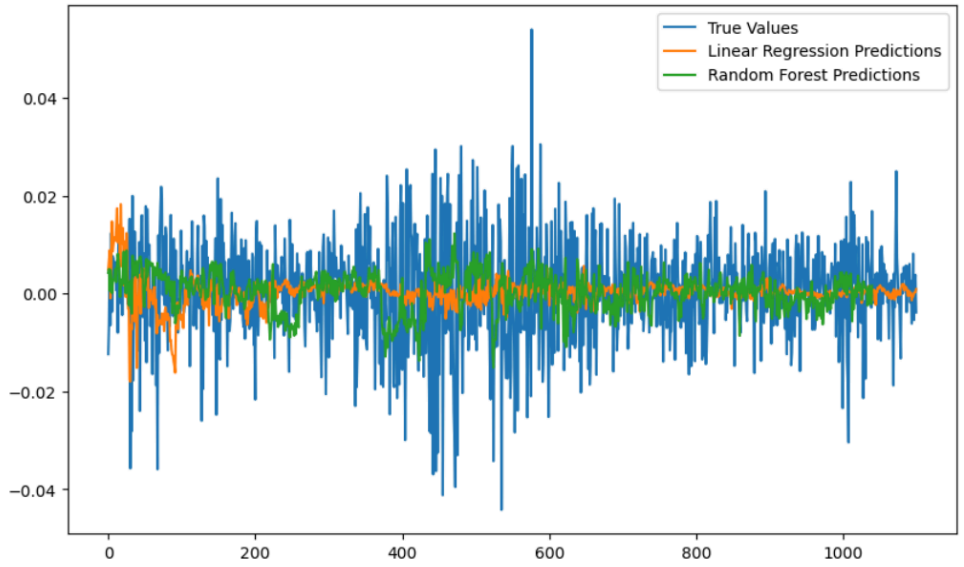
Courbes des RMSE



Courbes des MAE

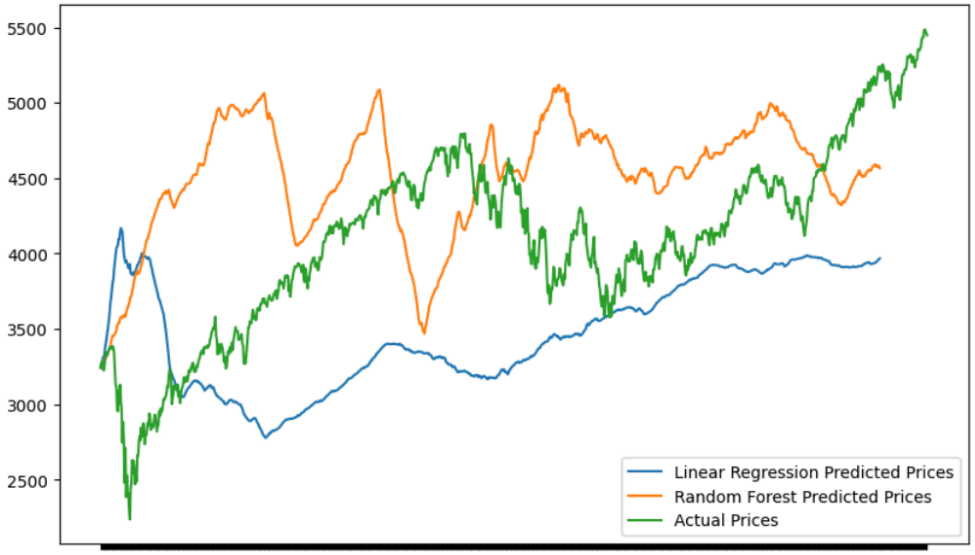


En comparant ensuite les erreurs de prédiction, on observe des écarts relatifs plutôt moyens pour les deux modèles. En regardant plus précisément, on voit que le modèle de régression linéaire a une erreur un peu plus faible et des valeurs moins extrêmes que le *Random Forest*, ce qui corrobore les résultats précédents sur la MAE et le RMSE.



Enfin nous comparons nos résultats prédits avec les vraies données. Les prédictions du modèle de régression linéaire sont très lissées, ce qui limite ce modèle dans les périodes de forte volatilité. Le modèle *Random Forest* est quant à lui moins lisse, et prévoit mieux les fluctuations à court terme au détriment des

variations globales qui ne sont pas très bien prédites. En quelque sorte, nous pouvons en déduire que ces deux modèles peuvent être utilisés à des desseins différents : si la précision globale est prioritaire, la régression linéaire est plus pertinente tandis que si l'essentiel est de capturer des variations rapides, alors le Random Forest pourrait être privilégié.



2. Discussion sur les erreurs courantes

Parmi les potentielles sources d'erreur se trouve le surajustement du modèle Random Forest qui est susceptible de capturer non seulement les relations pertinentes mais aussi le bruit.

En outre, la présence de variables fortement corrélées peut biaiser les coefficients de la régression linéaire, rendant les prédictions moins fiables. En dépit de la sélection des variables que nous avons opérée (par exemple, en éliminant des colonnes très corrélées comme EMA), il reste possible que certaines corrélations trop importantes subsistent, c'est la multicolinéarité.

Il peut également y avoir de l'endogénéité si des variables explicatives sont influencées par une variable dépendante (ou si des variables manquantes influencent à la fois les prédicteurs et la cible). Cela peut fausser les résultats des deux modèles.

Enfin, on peut également noter que des événements de marché extrêmes ou des changements structurels peuvent toujours affecter la pertinence des prédictions et il peut en résulter des données non stationnaires.

3. Conclusion et recommandations

Tout d'abord, nous avons pu observer que le cours du S&P500 remplit des critères qui permettent son analyse statistique et sa modélisation. Nous avons ensuite entraîné deux modèles sur les données disponibles : un modèle basé sur la Régression Linéaire, et un modèle Random Forest.

Leur performance est similaire ; le RMSE et le MAE indiquent que nos modèles ne sont pas très bons lorsqu'il s'agit de prédire l'évolution du cours du S&P500, même si le modèle de régression linéaire y parvient un peu mieux que le Random Forest.

La nature même des marchés, très volatiles, nous amène à penser que de tels résultats étaient attendus ; en effet, un algorithme ne peut prévoir des crises économiques dûes à une pandémie, par exemple.

Une piste qu'il aurait pû être intéressant d'exploiter est celle de l'application d'un test de causalité statistique, comme le test de causalité de Granger, qui a d'ailleurs été développé en vue d'une utilisation en économétrie. Ce test aurait pû être appliqué entre les facteurs et le cours du S&P500. Les résultats de ce test auraient peut-être permis des améliorations sur les modèles prédictifs.

Nous aurions aussi pû examiner si d'autres indices boursiers (DowJones, CAC40, Nasdaq...) sont impactés de la même manière par les facteurs économiques choisis.