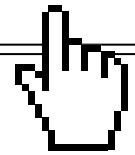
Prédiction S&P500 :

Projet Data

Par Félix Robotti, Yohann Le Couster, Margaux Carron, Fanny Lamothe et Alexandra Gomes



01 Mise en situation

Que va-t-on modéliser ?

02 Collecte et préparation des données

Quelles données a-t-on choisi et comment les a-t-on traité ?

03 Développement des modèles prédictifs

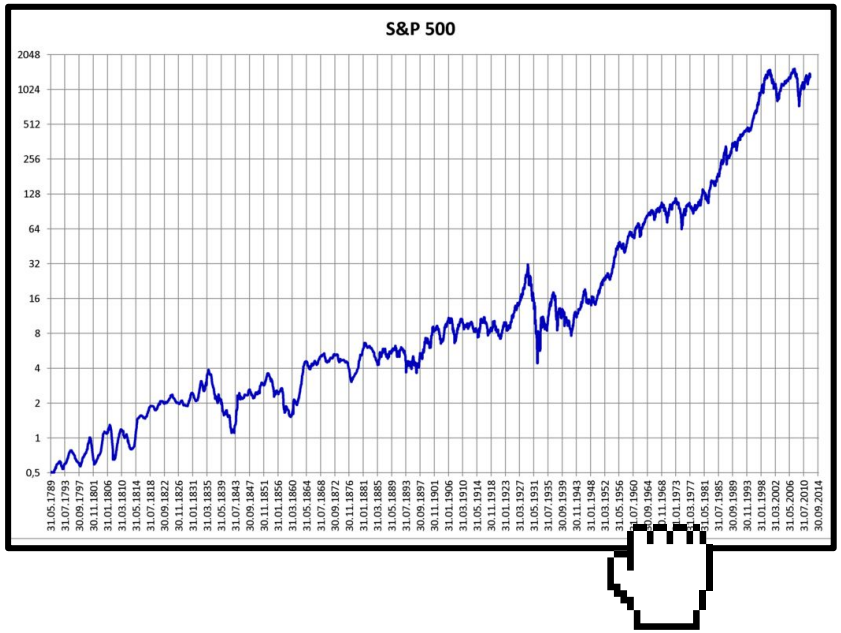
Quels modèles a-t-on développé pour modéliser ?

04 Résultats

Qu'obtient-on ?

01 : Mise en situation

Le cours du S&P500



Source : Wikipédia

Cours des 500 plus importantes entreprises américaines en terme de chiffre d'affaire

Dépend de nombreux paramètres macroéconomiques donc difficile à modéliser

→ Pertinence d'utiliser des modèles prédictifs

02 : Collecte et préparation des données

Cours du S&P500 et données macroéconomiques

Données choisies :

- S&P500_Close : valeur du S&P500 à la fermeture du marché
- S&P500_Volume : volume de transaction du S&P500 au cours d'une journée
- VIX_Close : indice de volatilité des marchés financiers à la fermeture
- xM et xY: taux d'intérêt à x mois (M : Months) et x années (Y : Years)
- S&P500_log_returns : log-returns du S&P500 au cours d'une journée
- EMAx : moyenne glissante exponentielle sur x jours
- MACD : différence entre un EMA12 et un EMA26, indique si la tendance de marché court-terme est meilleure que la tendance long-terme
- Signal_line : moyenne glissante exponentielle sur 9 jours de MACD
- RSI_x : Relative Strength Index sur les x derniers jours (indique une situation de survente ou de surachat)

02 : Collecte et préparation des données

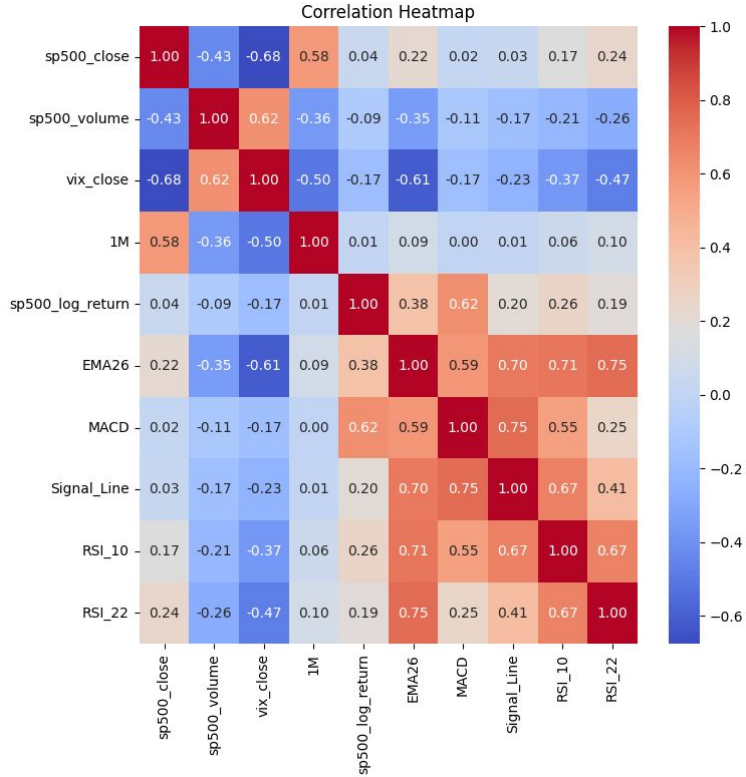
Cours du S&P500 et données macroéconomiques

Le cours du S&P500 est **stationnaire** et semble suivre un modèle **ARIMA(3,2)** :

- Stationarité du cours du S&P500
  - Test augmenté de Dickey-Fuller avec p-value = $4{,}76 \times 10^{-9}$
- Modèle ARIMA(3,2)
  - AIC = -7263

Matrice de **corrélation** des facteurs macroéconomiques choisis :

- Seuls 1M, EMA26, RSI_10, RSI_22 gardés car seuls pertinents



03 : Développement des modèles prédictifs

Régression Linéaire et Random Forest

Régression Linéaire :

- Choisi car simple mais efficace
- MAE = 0,00805
- RMSE = 0,0108

Random Forest :

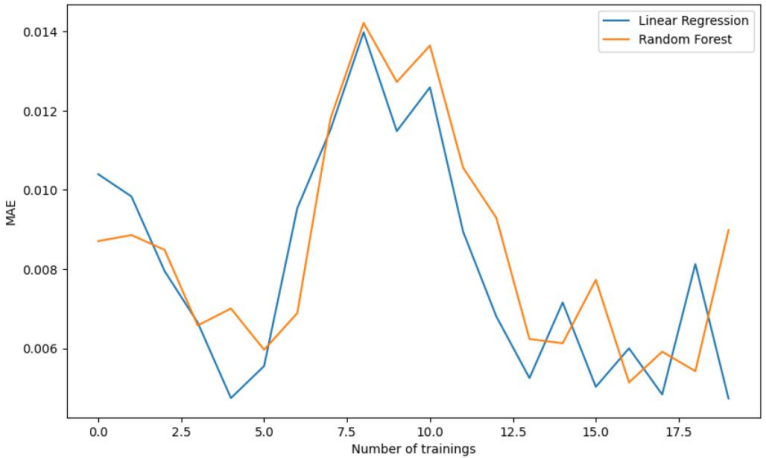
- Choisi car non linéaire et donc représente une approche totalement différente du modèle de Régression Linéaire
- MAE = 0,00851
- RMSE = 0,01079

→ Performances similaires des modèles

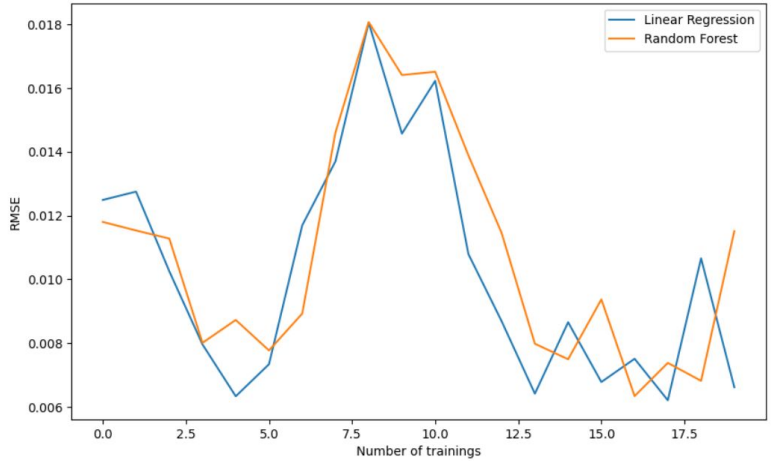
04 : Résultats

Analyse des performances

Courbe MAE



Courbe RMSE



→ Performances similaires des modèles

04 : Résultats

Analyse des performances

La régression linéaire "lisse" les résultats, Random Forest présente plus de valeurs extrêmes

Si la précision globale est prioritaire, mieux vaut utiliser un modèle de régression linéaire. Si l'objectif est de capturer des variations rapides, Random Forest est à privilegier.

Les sources d'erreurs sont un possible surajustement du modèle Random Forest, l'endogénéité et la multicolinéarité des variables et les évènements extrêmes du marché sur la période utilisée pour l'entraînement des modèles.

Cours prédit et cours réel du S&P500