Prévision des mouvements du S&P 500

Projet DATA // Filière DDEFi
Décembre 2024

Jules Besson, Antoine Delcambre, Loïc Blocquel, Thomas Pham, Aubin Maillard

1

LES OBJECTIFS

LE BUT

Développer un modèle de machine learning pour prédire les prix et rendements des actifs du S&P 500.

LA METHODE

- Collecter des données historiques du S&P 500.
- Intégrer des variables exogènes
- Développer des modèles adaptés aux séries temporelles et aux relations non linéaires.
- Évaluer les performances des modèles.
- Fournir des recommandations d'investissement basées sur les prédictions

2

COLLECTER LES DONNEES

COLLECTER LES DONNEES

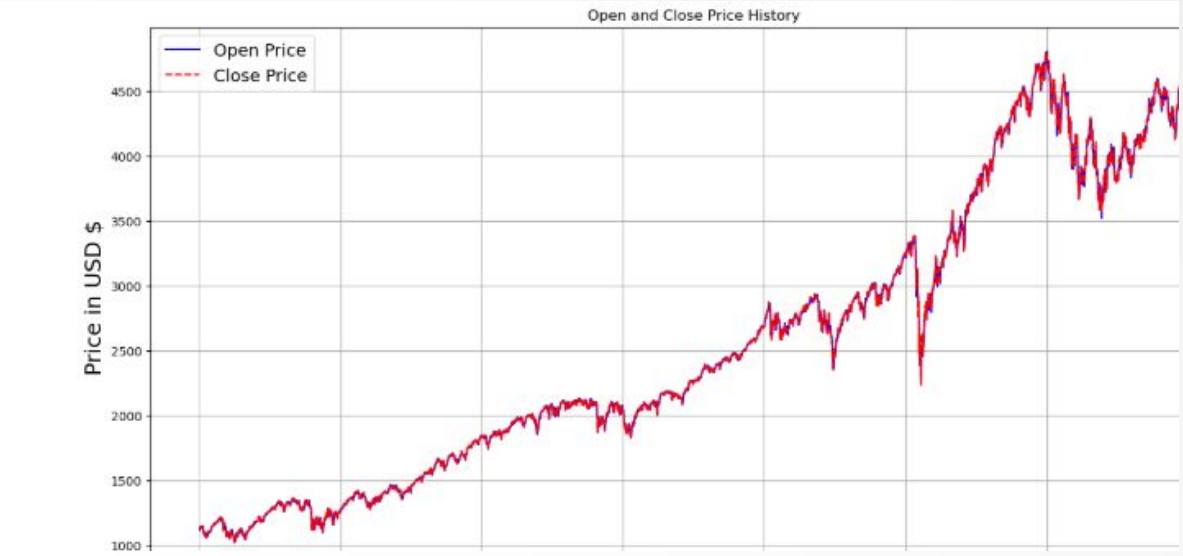
LA SOURCE

Développer un modèle de machine learning pour prédire les prix et rendements des actifs du S&P 500.

LES ETAPES

- Nettoyer les données (suppression des doublons, gestion des NaN)
- Transformer les prix en rendements logarithmiques.
- Vérifier la stationnarité (test ADF).

Après collecte des données, graphique des prix d'ouverture et de fermeture du S&P 500 sur 14 ans



On ajoute également d'autres données comme des variables exogènes pour permettre une évaluation complète. Ici nous avons rajouté :

- L'indice de volatilité (VIX)
- Taux d'intérêt à 10 ans (Reflète les conditions de marché à long terme)
- Taux de chômage (Représentent les données macroéconomiques)
- Indice sur l'inflation

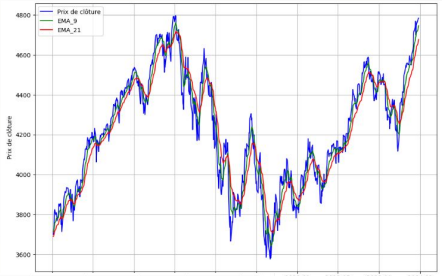
3

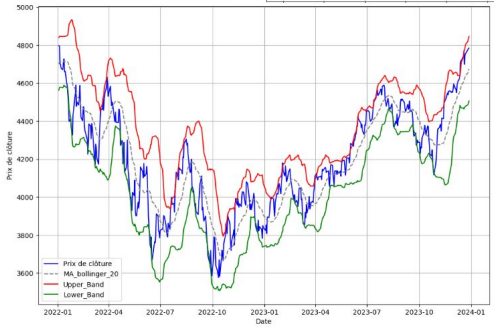
FEATURE ENGINEERING

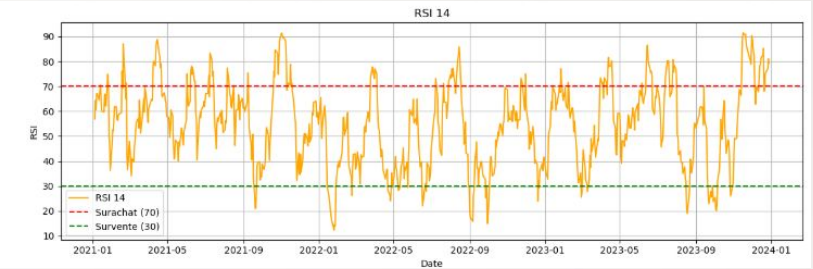
FEATURE ENGINEERING

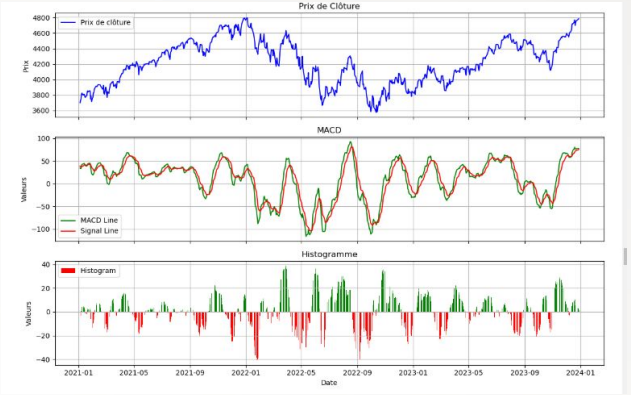
CREATION D'INDICATEURS TECHNIQUES

- **RSI** : identification des zones de surachat/survente
- **Moyennes mobiles** (MA, EMA) : suivi des tendances
- **MACD** : détection des signaux d'achat/vente
- **Bandes de Bollinger** : analyse de la volatilité





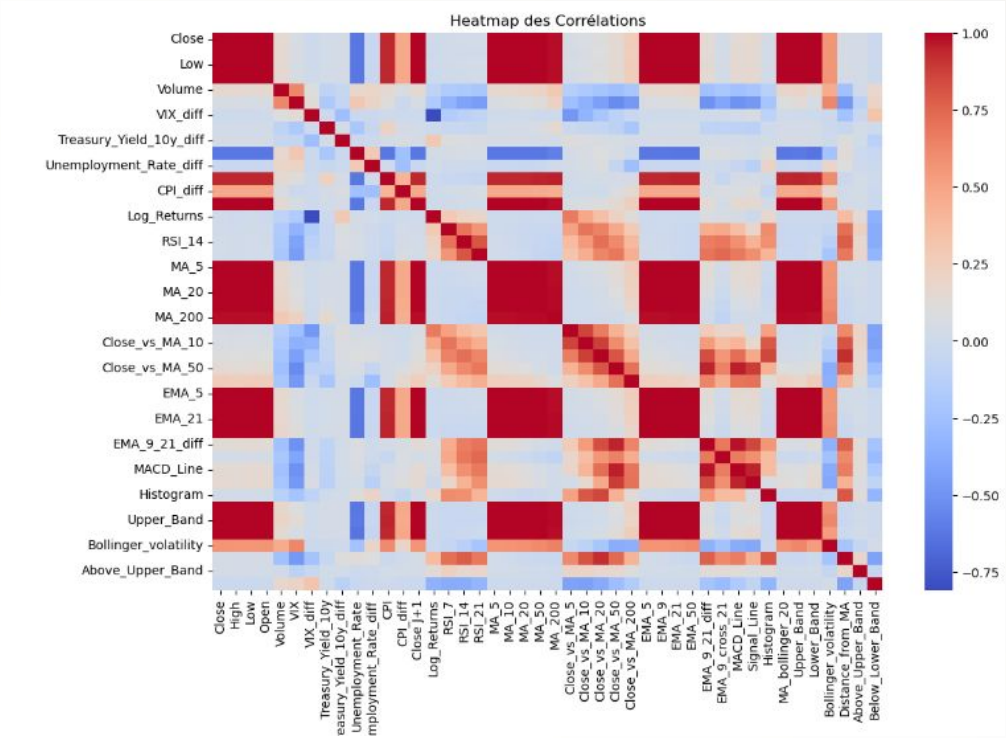




FEATURE ENGINEERING

Sélection de variables pertinentes

afin de sélectionner les variables les plus pertinente, il est intéressant de faire une analyse de corrélation et des test de multicolinéarité (VIF) par exemple avec une Heatmap des corrélation



4

MODELES DE ML TESTEES

Nous avons testé les modèles suivants

REGRESSION LINEAIRE

Baseline simple mais limitée.

RANDOM FOREST REGRESSOR

Capture les relations non linéaires.

XGBOOST

Optimisation des performances sur données complexes.

VOTING REGRESSOR

Agrégation pour combiner les forces des modèles.

5

RESULTATS

RESULTATS

METRIQUES D'EVALUATION

MAE : erreur absolue moyenne.

RMSE : erreur quadratique moyenne.

RESULTATS

Model	MAE	RMSE
Régression Linéaire	56.39	79.12
Random Forest	56.39	79.12
XGBoost	62.85	87.14
Voting Regressor	48.58	64.90

Les modèles avancés réduisent les erreurs par rapport à la baseline.

Mais il reste des **difficultés à capturer la complexité des données financières**.

DEFIS

1. Multicolinéarité, à cause des corrélations entre variables explicatives.

Solution : réduire le nombre de variables

2. **Non-linéarité des données** (nécessite des modèles sophistiqués.)

3. **Performance des prédictions** : des erreurs élevées malgré l'optimisation des modèles.

6

CONCLUSION

CONCLUSION

APPROCHES EVENTUELLES A L'AVENIR

- Exploration de modèles profonds (LSTM, réseaux neuronaux)
- Optimisation des hyperparamètres pour XGBoost et Random Forest
- Ajout de données contextuelles (news sentiment, données sectorielles)

PROCHAINES ETAPES

- Affiner les modèles pour mieux capturer la complexité des marchés.
- Créer une nouvelles variables appelé "mouvement" en calculant la différence entre le prix de clôture du jour suivant et le prix de clôture actuel et l'utiliser pour les modéles futures.