

SYNTHÈSE DU PROJET

PRÉDICTION DES PRIX ET ACTIFS DU S&P 500

Groupe 02 :

Nolan Bouigue, Rémi Chabo, Maxime Falanga,
Joseph Rigaut, Sibylle Lehmann

SOMMAIRE

- **Partie 1 :**

Collecte et Préparation des Données

- **Partie 2 :**

Développement du Modèle Prédicatif

- **Partie 3 :**

Analyse des Résultats et Interprétation



Collecte des
Données

PARTIE 1

Feature
engineering

Préparation
des Données

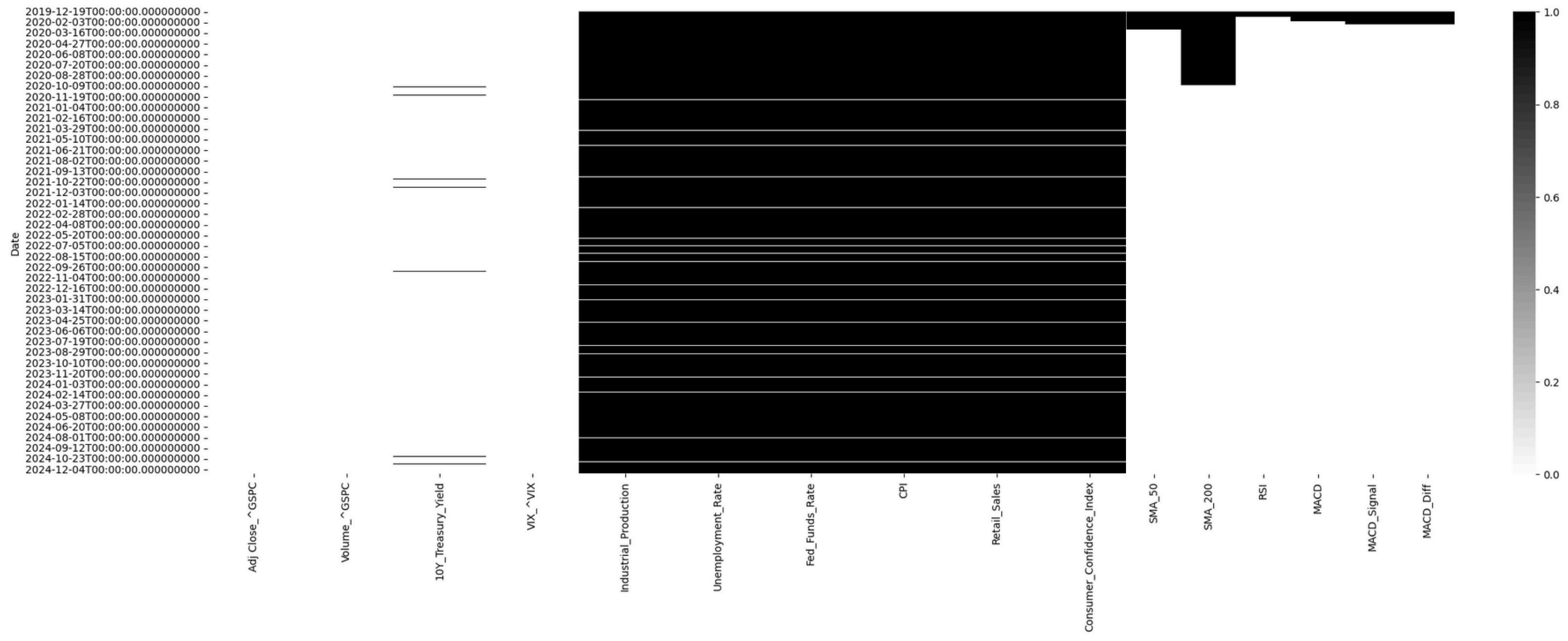
CHOIX DES DONNÉES

Indice économique	Variable
prix de clôture ajusté	Adj Close ^GSPC
volume	Volume ^GSPC
VIX	VIX ^VIX
Taux de chômage (UNRATE)	Unemployment Rate
Taux des fonds fédéraux (FEDFUNDS)	Fed Funds Rate
Indice des prix à la consommation (CPIAUCSL)	CPI
Ventes au détail (RSAFS)	Retail Sales
Indice de confiance des consommateurs (UMCSENT)	Consumer_Confidence_Index
Indice de production industrielle (INDPRO)	Industrial Production
Intérêts bons du Trésor à 10 ans (DGS10)	10Y_Treasury_Yield

Variables obtenues via Yahoo Finance et la base de données de la FED

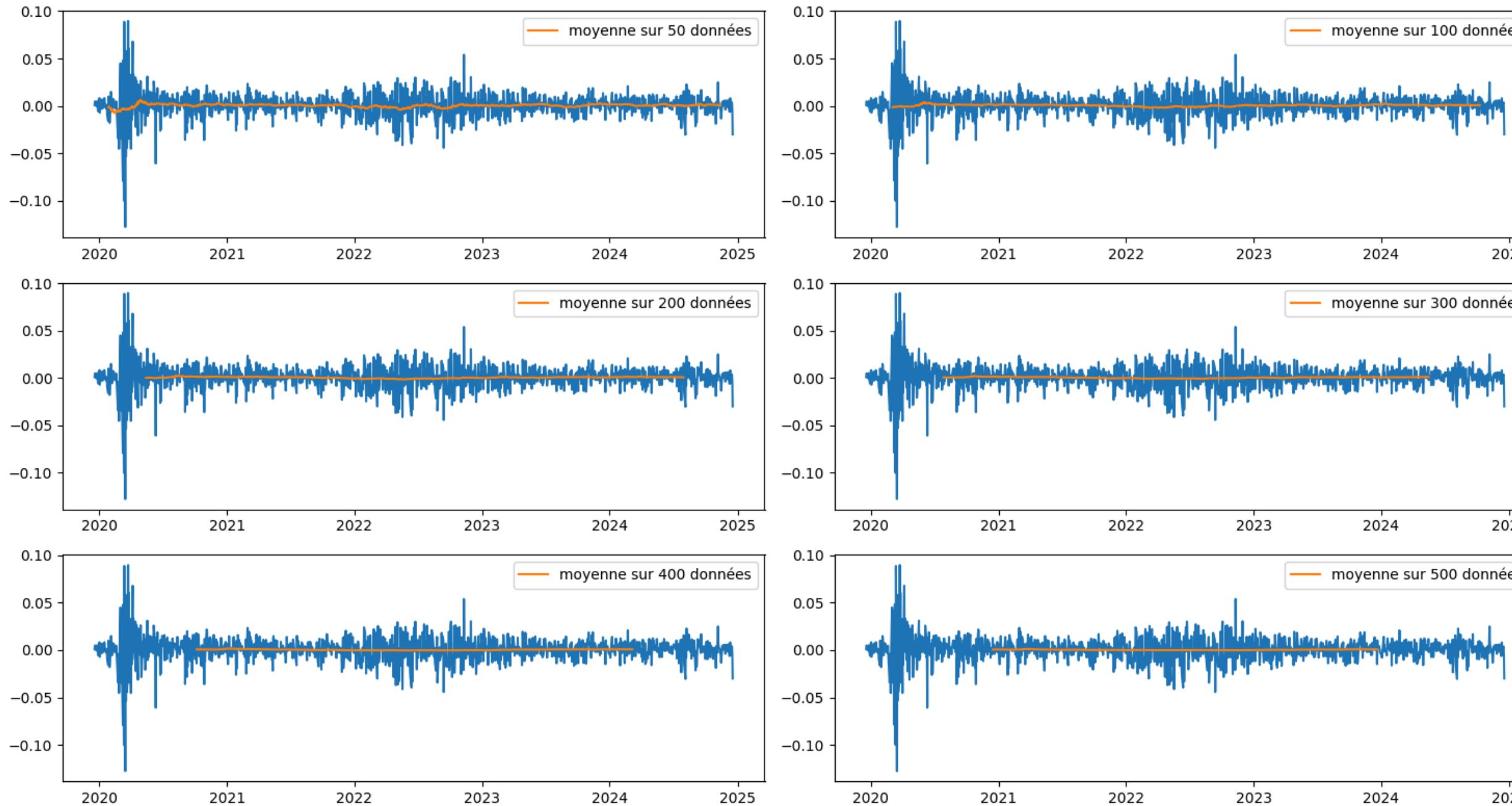
PRÉPARATION DES DONNÉES

- Transformation des données mensuelles en données journalières avec bfill et ffill
- Remplacement des valeurs manquantes par leur moyennes (notamment pour les n premières valeurs des données calculées)



Traitement des données manquantes - Python seaborn

ANALYSE DES LOG-RENDEMENTS



- Calcul log-rendements (valeur cible à prédire) à partir des prix de clôture du S&P500 sur 5 ans
- Stationnarité de la série temporelle (test ADF)
- Pas de saisonnalité de la série

SELECTION DES DONNÉES

Scores des features avec SelectKBest:

	Feature	Score
0	Adj Close_ ^GSPC	4.989837e+15
9	SMA_50	1.716521e+04
10	SMA_200	2.903410e+03
6	CPI	7.193717e+02
7	Retail_Sales	7.048655e+02
3	VIX_ ^VIX	5.454870e+02
5	Fed_Funds_Rate	3.604260e+02
2	10Y_Treasury_Yield	3.214506e+02
13	MACD_Signal	1.660940e+02
12	MACD	1.519688e+02
11	RSI	1.035453e+02
1	Volume_ ^GSPC	9.456049e+01
4	Unemployment_Rate	9.446997e+01
8	Consumer_Confidence_Index	3.851444e+00
14	MACD_Diff	1.469552e+00

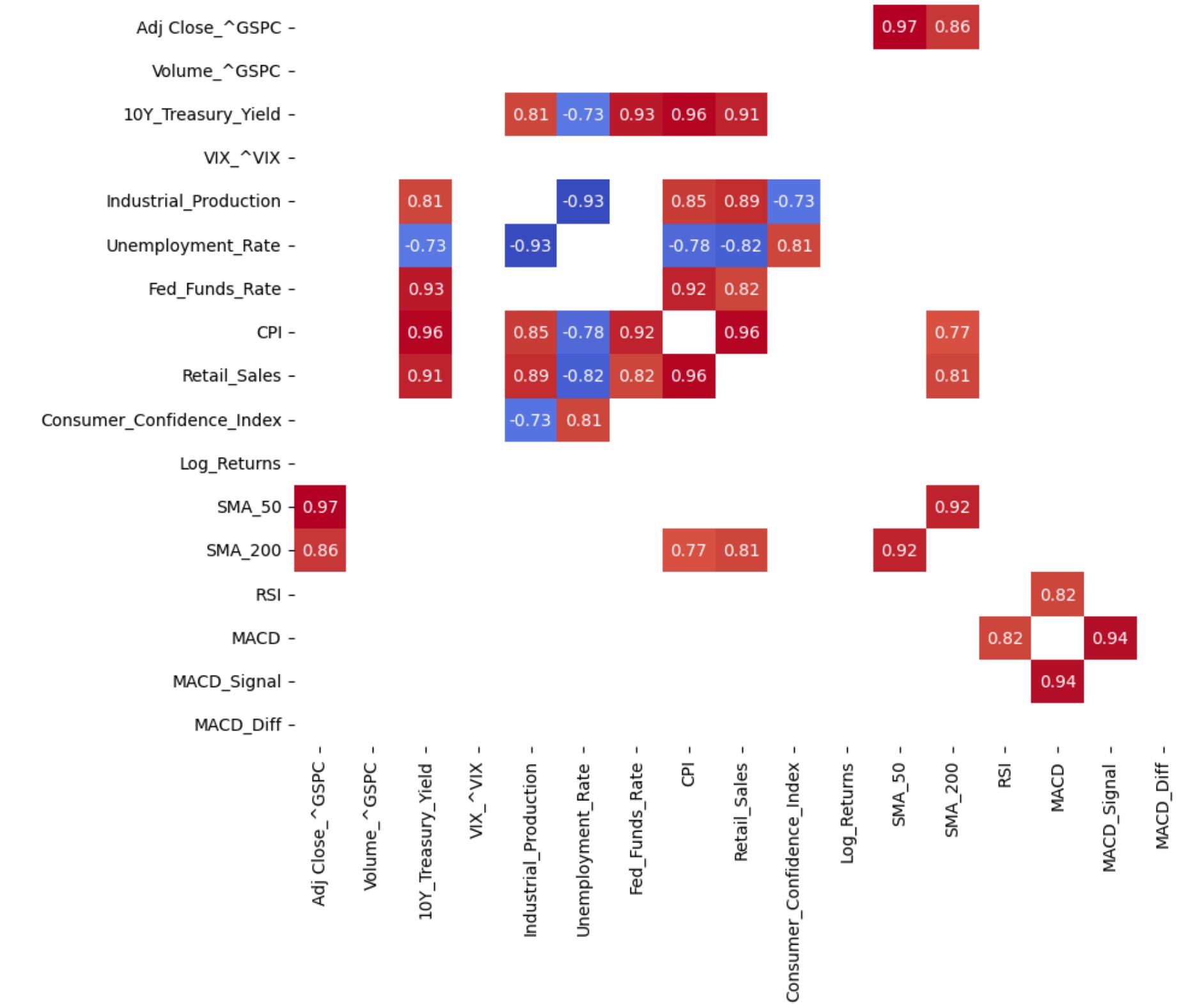
Score des variables avec la méthode SelectKBest - Python sklearn

Features sélectionnées par RFE:

```
Index(['Adj Close_ ^GSPC', '10Y_Treasury_Yield', 'Unemployment_Rate',
       'Fed_Funds_Rate', 'Consumer_Confidence_Index'],
      dtype='object')
```

Sélection de 5 variables avec l'algorithme RFE - Python sklearn

Matrice de Corrélation abs(R) > 0.7



Matrice de corrélation |R| > 0,7 - Python pandas

Choix des
modèles

PARTIE 2

Entraînement
et validation

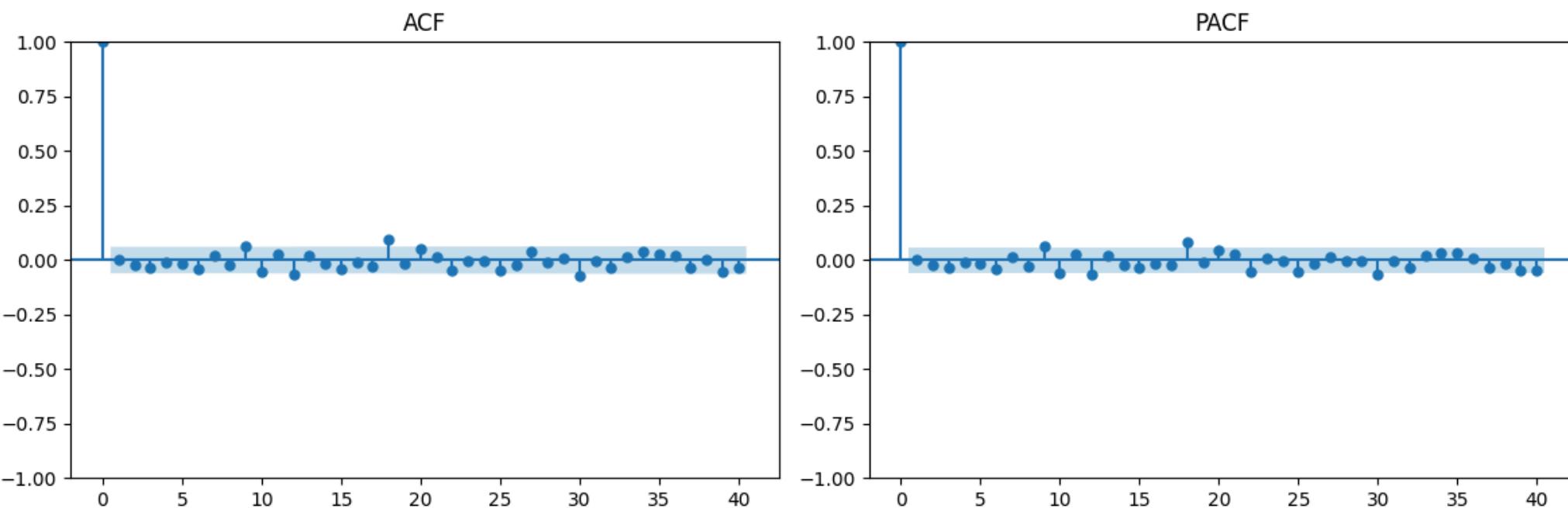
CHOIX DES MODÈLES

Régression linéaire: simple à implémenter, rapide à entraîner et fournit une base de comparaison pour évaluer d'autres modèles peut être plus complexes.

ARIMA (AutoRegressive Integrated Moving Average): simple à paramétrier, nécessite assez peu de données pour produire des résultats fiables, solution directe lorsqu'il s'agit de série univariée et sans saisonnalité.

SARIMAX Results						
Dep. Variable:	Log_Retruns	No. Observations:	1059			
Model:	ARIMA(2, 0, 3)	Log Likelihood	3333.377			
Date:	Fri, 03 Jan 2025	AIC	-6652.754			
Time:		BIC	-6617.998			
Sample:	0 - 1059	HQIC	-6639.581			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
const	0.0005	0.000	1.686	0.092	-8.56e-05	0.001
ar.L1	-0.2378	0.548	-0.434	0.664	-1.311	0.836
ar.L2	0.5560	0.368	1.510	0.131	-0.166	1.278
ma.L1	0.2107	0.554	0.381	0.704	-0.875	1.296
ma.L2	-0.6042	0.351	-1.721	0.085	-1.292	0.084
ma.L3	-0.0031	0.046	-0.067	0.946	-0.093	0.087
sigma2	0.0001	3.51e-06	30.786	0.000	0.000	0.000
Ljung-Box (L1) (Q):		0.83	Jarque-Bera (JB):		178.28	
Prob(Q):		0.36	Prob(JB):		0.00	
Heteroskedasticity (H):		0.72	Skew:		-0.34	
Prob(H) (two-sided):		0.00	Kurtosis:		4.89	

ARMA(2,3)

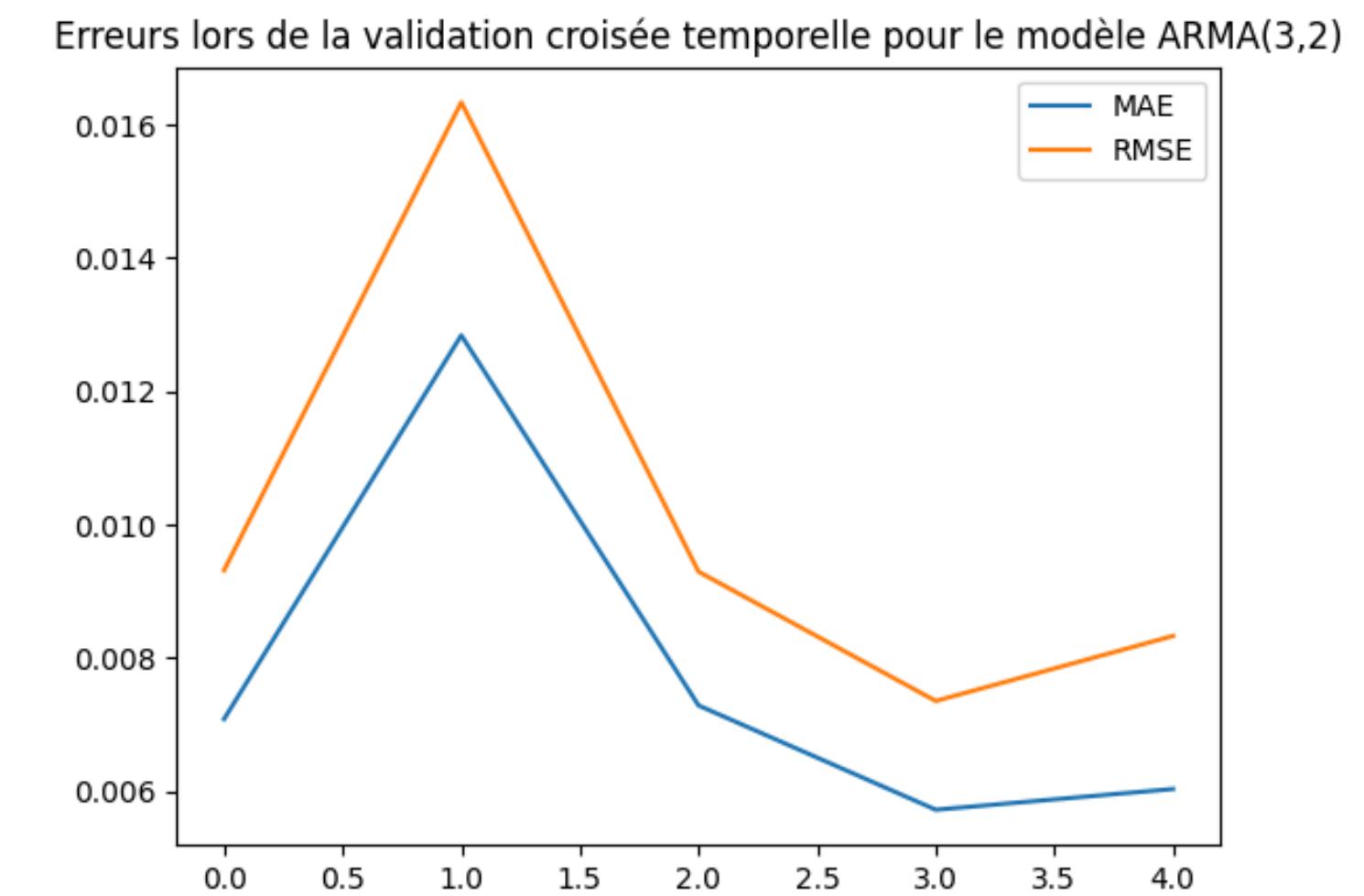
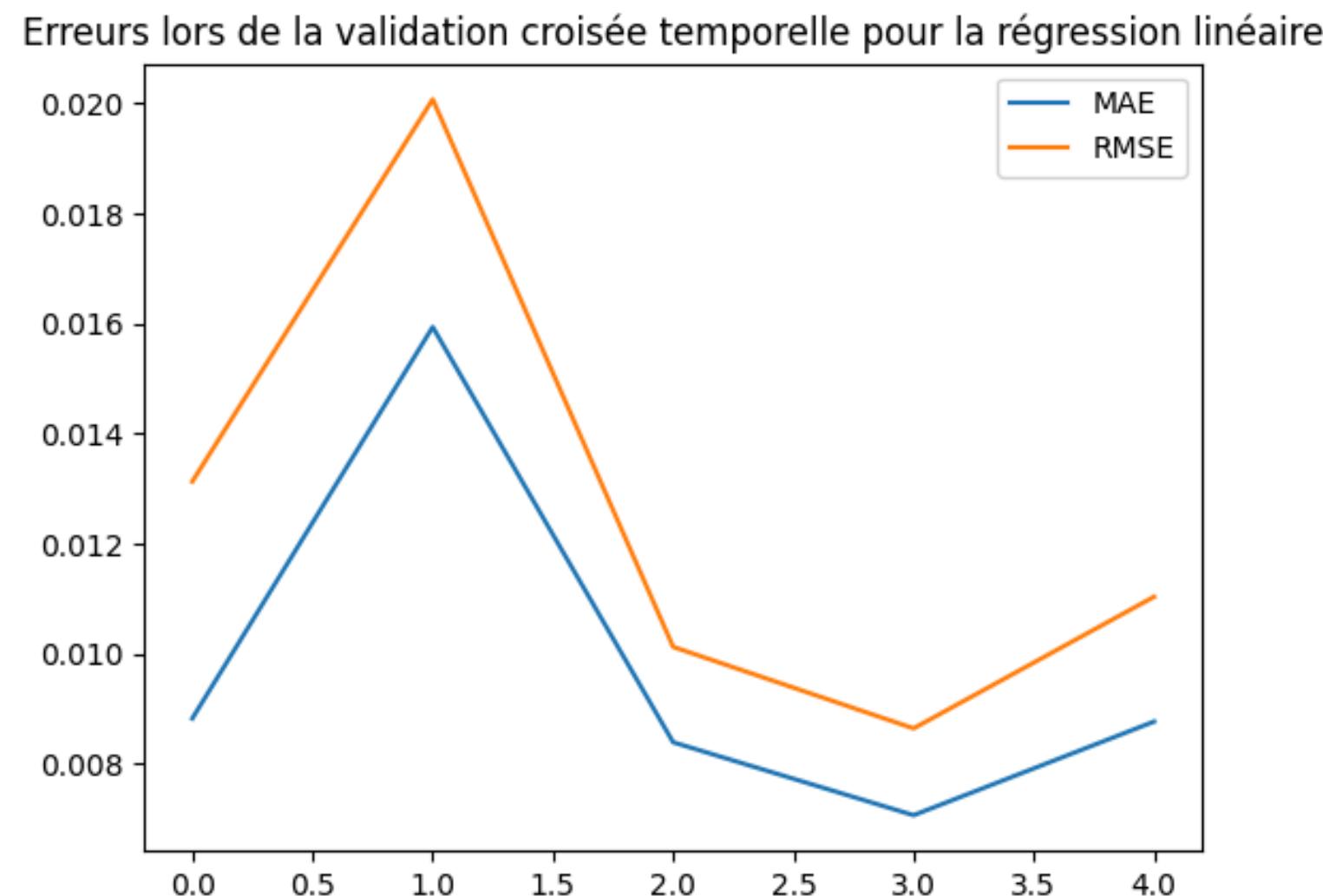


fonction d'autocorrélation (ACF) et fonction d'autocorrélation partielle (PACF) des log-rendements.

SARIMAX Results						
Dep. Variable:	y	No. Observations:	1058			
Model:	SARIMAX(3, 0, 2)	Log Likelihood	3319.227			
Date:	Fri, 03 Jan 2025	AIC	-6626.454			
Time:	14:31:52	BIC	-6596.669			
Sample:	0 - 1058	HQIC	-6615.165			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	-0.7617	0.217	-3.509	0.000	-1.187	-0.336
ar.L2	-0.0363	0.033	-1.105	0.269	-0.101	0.028
ar.L3	-0.0624	0.027	-2.282	0.022	-0.116	-0.009
ma.L1	-0.2234	0.211	-1.061	0.289	-0.636	0.189
ma.L2	-0.7479	0.209	-3.583	0.000	-1.157	-0.339
sigma2	0.0001	3.66e-06	30.118	0.000	0.000	0.000
Ljung-Box (L1) (Q):		0.07	Jarque-Bera (JB):		157.70	
Prob(Q):		0.79	Prob(JB):		0.00	
Heteroskedasticity (H):		0.71	Skew:		-0.13	
Prob(H) (two-sided):		0.00	Kurtosis:		4.87	

ARMA(3,2)

Ratio 80/20: 80% des données sont utilisées pour entraîner le modèle et lui permettre d'apprendre les relations entre les variables et 20% des données sont réservées pour évaluer la performance du modèle sur des données jamais vues pendant l'entraînement.



La RMSE (Root Mean Squared Error) mesure l'erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles.

La MAE (Mean Absolute Error) mesure l'erreur moyenne absolue entre les valeurs prédites et les valeurs réelles

Discussion sur les erreurs courantes



Performances et Limites

- Erreurs quadratiques moyennes (validation croisée 5 blocs) : 0,8% à 1,6%.
- Précision limitée par la corrélation élevée des variables explicatives (cf matrice de corrélation) conduisant à une sous-performance due à la méthode des moindres carrés ordinaires.
- Non-linéarité des séries temporelles : La régression linéaire est peu adapté pour modéliser des phénomènes non linéaires

Améliorations possibles

- Utiliser des méthodes de régression LASSO ou RIDGE pour réduire l'impact des corrélations.
- Explorer des modèles non linéaires pour capturer des dynamiques complexes des séries temporelles (le papier Chongda Liu, University of Illinois (2016), utilise le modèle SVM avec le noyau linéaire, le noyau polynomial et la fonction de base radiale (RBF).)

Performances et Avantages

- Erreurs quadratiques moyennes (validation croisée 5 blocs) : 0,6% à 1,3%.
- Meilleur ajustement que la régression linéaire pour la modélisation du S&P500.
- Résidus homoscédastiques et non-correlés (Test de Ljung-Box : p-value = 0,79)

Points d'amélioration et contexte

- Utilisation de la fonction auto-ARIMA : Meilleures p-values pour les coefficients, seuls 2 au-dessus du seuil de 5%.
- Contexte particulier des données (période Covid-19) : Impact significatif sur les tendances économiques, réduisant la prédictibilité.

Conclusion : Le modèle ARMA(3,2) offre une meilleure précision et est plus adapté pour modéliser des séries temporelles complexes

CONCLUSION & RECOMMANDATIONS

Stratégies basées sur les prévisions

- Positions longues : Lorsque des rendements positifs sont prédits.
- Positions courtes : Lorsque des rendements négatifs sont prédits.
- Complément : Utiliser les indicateurs techniques RSI et MACD pour confirmer les signaux.

Données et horizons temporels

- Réévaluer le modèle avec des données post-Covid-19 pour mieux refléter les tendances actuelles.
- Étendre les horizons temporels pour capturer des phases de marché variées et améliorer la généralisation.

Intégration de nouveaux facteurs

- Ajouter des variables macroéconomiques validées par des tests de causalité (ex : test de Granger).
- Inclure des variables exogènes (chocs géopolitiques, climatiques, etc.) pour affiner les prédictions.
- Interaction mondiale : Intégrer des indicateurs d'autres places boursières (Nikkei, NASDAQ, USDJPY) et des matières premières (Or, pétrole).



MERCI

