

Sujet:

PRÉDICTION DU S&P500 :
COMPARAISON DES
MODÈLES LSTM ET ARIMA

Réalisé par :

Zakariae BOUYAKNIFEN

Filière :

Data Science

Table des matières

Résumé	3
1 Introduction	3
1.1 Contexte et Problématique	3
1.2 Objectifs	3
2 Méthodologie	4
2.1 Collecte des Données	4
2.2 Prétraitement	4
2.3 Modèle LSTM	4
2.4 Modèle ARIMA	4
2.5 Métriques d'Évaluation	5
3 Résultats	5
3.1 Performance LSTM (1 an)	5
3.2 Performance ARIMA	6
3.3 Comparaison Directe (1 an)	6
3.4 Visualisations	7
4 Discussion	9
4.1 Analyse Comparative	9
4.2 Recommandations	10
4.3 Limites et Perspectives	10
5 Conclusion	11
Workflow du Projet	11
Références	12

Résumé

Ce rapport présente une comparaison rigoureuse entre les réseaux LSTM (Long Short-Term Memory) et le modèle ARIMA pour la prédiction de l'indice S&P500. Les données ont été collectées via `yfinance`, traitées en Python, et visualisées dans Power BI.

Résultats principaux : Le LSTM surpasse ARIMA avec un RMSE de 151.30\$ (vs 605.84\$) sur une prédiction d'un an, soit une amélioration de 75%. Sur 31 jours, ARIMA reste compétitif (RMSE : 82.94\$, MAPE : 0.95%).

Conclusion : Le LSTM est recommandé pour les prédictions long terme, tandis qu'ARIMA conserve son utilité pour le court terme.

1 Introduction

1.1 Contexte et Problématique

Le S&P500 est l'indice boursier de référence regroupant les 500 plus grandes entreprises américaines. Sa prédiction présente un intérêt majeur pour les investisseurs et gestionnaires d'actifs.

Question de recherche : Les méthodes d'apprentissage profond (LSTM) offrent-elles de meilleures performances que les approches statistiques traditionnelles (ARIMA) pour prédire le S&P500 ?

1.2 Objectifs

1. Implémenter et comparer LSTM et ARIMA
2. Évaluer sur deux horizons : 31 jours et 1 an
3. Visualiser les résultats dans Power BI
4. Recommander le modèle optimal

2 Méthodologie

2.1 Collecte des Données

Les données ont été téléchargées via `yfinance` (Yahoo Finance API) :

- **Ticker** : ^GSPC (S&P500)
- **Période 1 an** : 2020-01-01 à 2024-12-31 (1257 observations)
- **Période 31 jours** : 2025-01-01 à 2025-10-28 (206 observations)
- **Variable** : Prix de clôture (Close)

2.2 Prétraitement

1. **Nettoyage** : Suppression des valeurs manquantes
2. **Division** : 80% training / 20% test (temporelle)
3. **Normalisation (LSTM)** : MinMaxScaler [0, 1]
4. **Différenciation (ARIMA)** : Stationnarisation de la série

2.3 Modèle LSTM

Architecture :

- 3 couches LSTM (50 unités chacune)
- Dropout 0.2 après chaque couche LSTM
- Couche dense (25 unités) + couche de sortie (1 unité)
- Séquence d'entrée : 60 jours
- Optimiseur : Adam, Loss : MSE
- Early stopping : patience 10 epochs

2.4 Modèle ARIMA

Sélection :

- Test ADF pour vérifier la stationnarité
- Grid search sur (p, d, q) avec $p \in [0,3]$, $d \in [0,3]$, $q \in [0,4]$
- Critère de sélection : AIC (Akaike Information Criterion)
- **Modèle optimal** : ARIMA(2, 2, 3)

2.5 Métriques d'Évaluation

- **RMSE** : $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ (erreur quadratique)
- **MAE** : $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$ (erreur absolue)
- **MAPE** : $\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$ (erreur en %)
- **MSE** : $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ (erreur quadratique moyenne)

3 Résultats

3.1 Performance LSTM (1 an)

Métrique	Valeur
RMSE	151.30 \$
MAE	130.88 \$
MAPE	2.51%
MSE	26,210

TABLE 3.1 – Performance du modèle LSTM sur 1 an

3.2 Performance ARIMA

Métrique	1 an	31 jours	Modèle
RMSE	605.84 \$	82.94 \$	ARIMA(2,2,3)
MAE	502.47 \$	63.92 \$	ARIMA(2,2,3)
MAPE	8.92%	0.95%	ARIMA(2,2,3)
MSE	289,510	3,290	ARIMA(2,2,3)
AIC	10,662.16	1,990.22	-

TABLE 3.2 – Performance des modèles ARIMA

3.3 Comparaison Directe (1 an)

Métrique	LSTM	ARIMA	Amélioration
RMSE (\$)	151.30	605.84	-75.0%
MAE (\$)	130.88	502.47	-73.9%
MAPE (%)	2.51	8.92	-71.9%
MSE	26,210	289,510	-90.9%

TABLE 3.3 – Comparaison LSTM vs ARIMA (1 an)

Le LSTM réduit l'erreur de prédiction de 75% par rapport à ARIMA sur un horizon d'un an.

3.4 Visualisations

4.i LSTM 1-Year (Image 3)

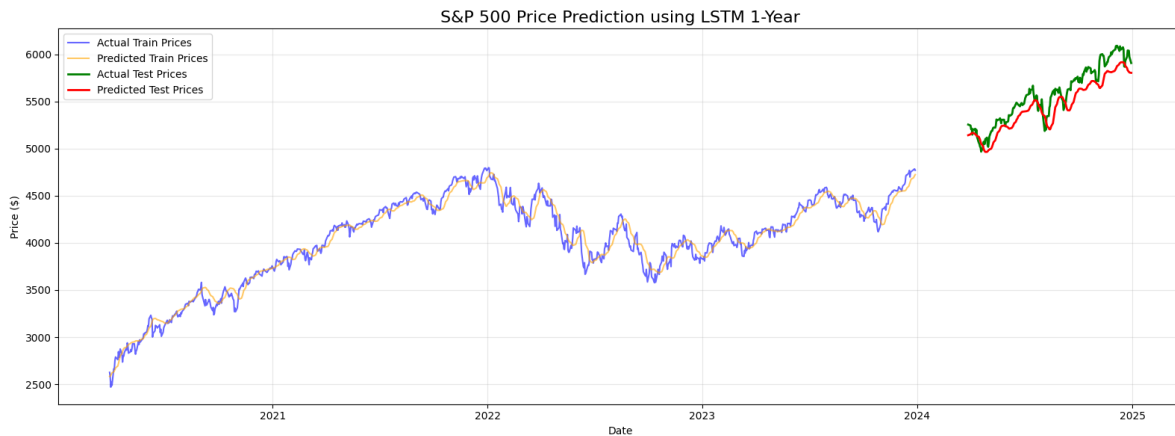


FIGURE 3.1 – Prédictions LSTM sur 1 an - Les prédictions (rouge) suivent étroitement les valeurs réelles (vert)

Observations :

- Excellente capture de la tendance générale
- Prédictions précises sur les données d'entraînement (bleu/orange quasi-confondus)
- Légère sous-estimation en fin de période test, mais erreur contrôlée

4.ii ARIMA 1-Year (Image 2)

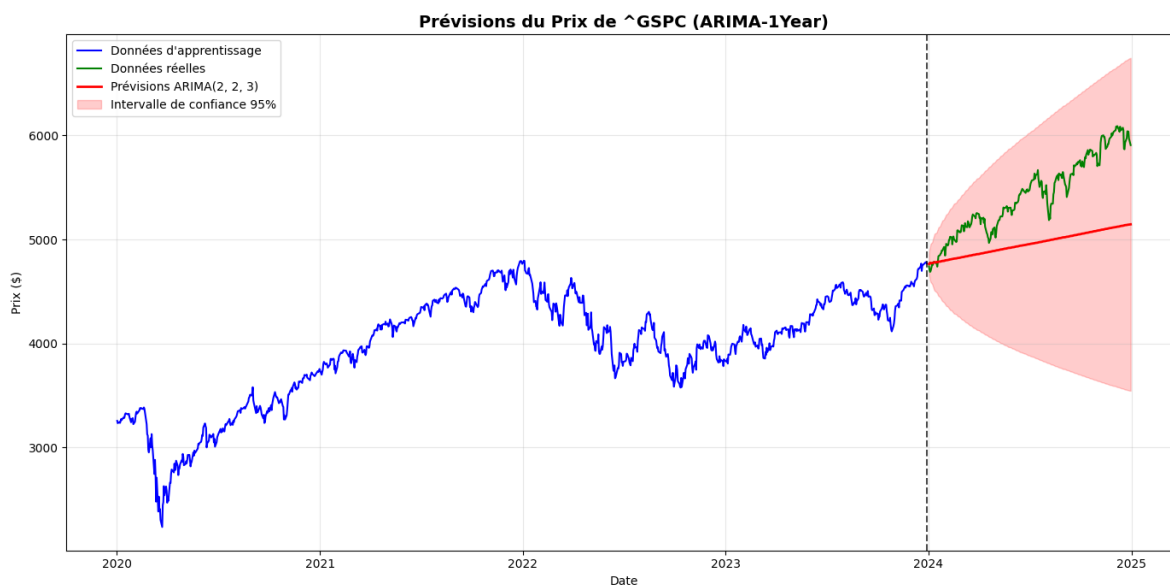


FIGURE 3.2 – Prédictions ARIMA sur 1 an - Large intervalle de confiance et prédiction linéaire

Observations :

- Prédiction linéaire (rouge) ne capture pas la volatilité réelle
- Sous-estimation systématique de la hausse de 2024
- Intervalle de confiance très large (3000\$ - 7000\$)

4.iii ARIMA 31-Day (Image 1)

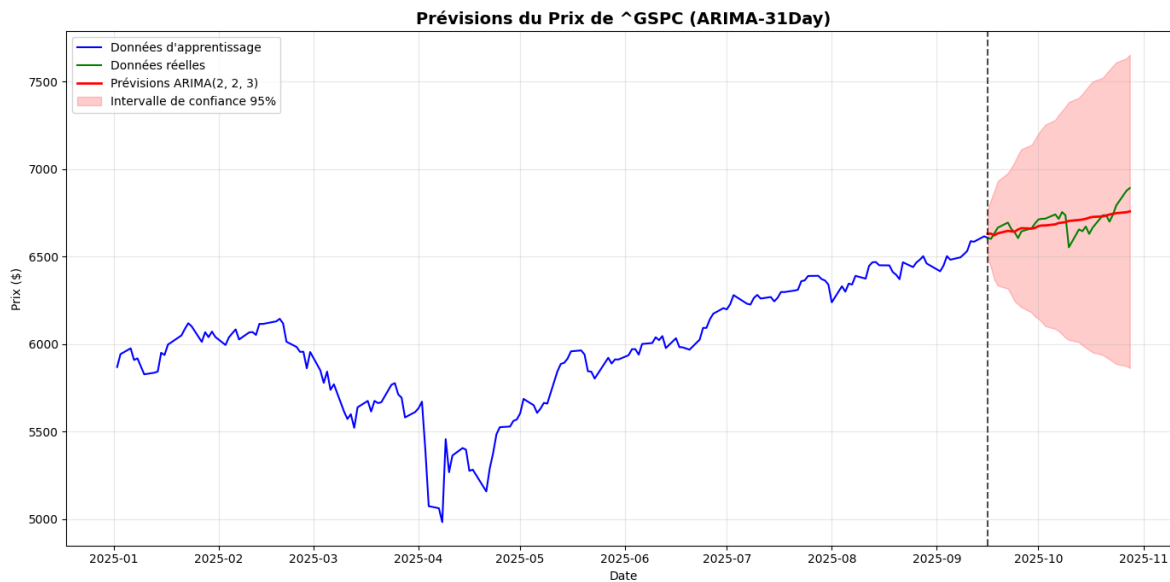


FIGURE 3.3 – Prédvisions ARIMA sur 31 jours - Bonne performance court terme

Observations :

- Excellente précision (MAPE 0.95%, RMSE 82.94\$)
- Les prédictions (rouge) suivent bien les valeurs réelles (vert)
- Intervalle de confiance raisonnable
- Confirme la pertinence d'ARIMA pour le court terme

4 Discussion

4.1 Analyse Comparative

Forces du LSTM :

- Précision supérieure de 75% sur le long terme
- Capture des patterns non-linéaires complexes
- Robuste aux changements de régime du marché (COVID-19, inflation, etc.)
- Mémoire contextuelle de 60 jours

Forces d'ARIMA :

- Excellent sur court terme (< 1 mois)
- Rapidité d'entraînement (secondes vs minutes)
- Interprétabilité mathématique
- Intervalles de confiance natifs

4.2 Recommandations

Cas d'usage	Modèle	Justification
Investissement > 6 mois	LSTM	Précision 75% supérieure
Trading < 1 mois	ARIMA	MAPE 0.95%, rapide
Budget calcul limité	ARIMA	Entraînement rapide
Applications critiques	LSTM	Risque financier réduit

TABLE 4.1 – Guide de sélection du modèle

4.3 Limites et Perspectives

Limites :

- Données univariées (prix uniquement)
- Absence de variables exogènes (taux, inflation, VIX)
- Période incluant COVID-19 (atypique)
- Hyperparamètres LSTM non optimisés exhaustivement

Améliorations futures :

- Intégrer indicateurs techniques (RSI, MACD, Bollinger)
- Ajouter données macroéconomiques
- Tester architectures Transformer
- Développer modèle hybride LSTM-ARIMA
- Application à d'autres indices (CAC40, DAX)

5 Conclusion

Ce projet démontre la supériorité des réseaux LSTM pour la prédiction du S&P500 sur des horizons moyens à longs termes, avec une réduction de 75% de l'erreur par rapport à ARIMA. Les visualisations confirment que le LSTM capture mieux la complexité des marchés financiers.

Cependant, ARIMA conserve sa pertinence pour les prédictions ultra court terme (< 1 mois) avec un MAPE exceptionnel de 0.95% sur 31 jours, tout en offrant rapidité et interprétabilité.

Recommandation finale : Utiliser LSTM pour l'allocation stratégique d'actifs et la gestion de portefeuille long terme, et ARIMA pour les décisions tactiques court terme. Une approche hybride combinant les deux méthodes pourrait offrir le meilleur compromis.

Les dashboards Power BI développés facilitent la communication des résultats aux parties prenantes et permettent un suivi en temps réel des performances prédictives.

Workflow du Projet

1. **Collecte** : yfinance \rightarrow données S&P500
2. **Nettoyage** : Python (pandas) \rightarrow suppression NaN, extraction Close
3. **Modélisation** :
 - LSTM : TensorFlow/Keras avec normalisation MinMaxScaler
 - ARIMA : statsmodels avec grid search
4. **Exportation** : Résultats vers CSV
5. **Visualisation** : Import CSV dans Power BI
6. **Dashboards** : 4 pages interactives (Accueil, LSTM, ARIMA 1Y, ARIMA 31D)

Références

1. Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735-1780.
2. Box, G. E., et al. (2015). *Time series analysis : forecasting and control*. Wiley.
3. Fischer, T., & Krauss, C. (2018). *Deep learning with LSTM for financial market predictions*. European Journal of Operational Research, 270(2), 654-669.
4. Siami-Namini, S., et al. (2018). *ARIMA vs LSTM in forecasting time series*. ICMLA.
5. Documentation yfinance : <https://pypi.org/project/yfinance/>

— *Fin du Rapport* —

BOUYAKNIFEN ZAKARIAE

Année 2024-2025