



## 2024 - 2025 PROJET DE FIN D'ÉTUDES

### DIPLÔME NATIONAL D'INGÉNIEUR

**SPÉCIALITÉ : INFORMATIQUE**  
**(Option Data Science & Actuariat)**

**Prévision des prix et analyse de la  
volatilité sur le marché des  
cryptomonnaies**

Réalisé par : PRINCE JEAN LUC De Nteppe

Encadré par :

Encadrant ESPRIT : Mme. HANFI Nardine

Encadrant Entreprise : M. MOHSEN Chebbi



Je valide le dépôt du rapport PFE relatif à l'étudiant nommé ci-dessous / I validate the submission of the student's report:

- Nom & Prénom /Name & Surname :

Prince Jean Luc de Népée

Encadrant Entreprise/ Business site Supervisor

- Nom & Prénom /Name & Surname :

Mohsen Chebbi

Cachet & Signature / Stamp & Signature

ESPRIT  
18. Rue de l'Usine, Cherguia 2 Antara  
Tél: 71 941 541 / 71 741 889



Encadrant Académique/Academic Supervisor

- Nom & Prénom /Name & Surname:

Signature / Signature

Ce formulaire doit être rempli, signé et scanné/This form must be completed, signed and scanned.

Ce formulaire doit être introduit après la page de garde/ This form must be inserted after the cover page.

---

## DÉDICACE

*À mes parents, Rodolphe Ebwea Djembele Thomas et Claire Epety Patricia Na'Ntingue, épouse Ebwea - socle de ma résilience - sans lesquels je serais perdu. Depuis mon arrivée en Tunisie, ils m'accompagnent et me soutiennent sans relâche.*

*À ma mère, qui, malgré l'âge, se lève très tôt et se couche très tard, passant la majeure partie de son temps devant la machine à coudre.*

*À mon père, qui, malgré la maladie, répond toujours présent.*

*Dans chacun de mes échecs, ils ont été là pour moi.*

*Je dédie également ce travail à mon petit frère, à toute ma famille et à mes amis qui me soutiennent depuis le Cameroun, malgré la distance ; ainsi qu'à mes amis en Tunisie et de par le monde, qui illuminent mon quotidien. Plus qu'une relation amicale, vous êtes ma famille.*

---

## REMERCIEMENTS

À l'issue de ce travail, je souhaite exprimer ma profonde gratitude à toutes celles et tous ceux qui ont rendu possible l'accomplissement de ce projet de fin d'études.

Je rends d'abord grâce au **Seigneur Tout-Puissant**, qui m'a accordé le souffle de vie, la force et la persévérance nécessaires pour mener ce projet à son terme.

Je tiens à exprimer ma reconnaissance la plus vive à **M. Mohsen Chebbi**, qui fut bien plus qu'un encadrant professionnel : un collaborateur attentif, disponible et toujours prêt à m'épauler. Ses conseils avisés et sa bienveillance ont constitué un appui décisif tout au long de cette aventure.

J'adresse également mes remerciements au **corps enseignant d'Esprit**, et plus particulièrement à **M. Mourad Hassini**, professeur de DevOps, à **M. Mohamed Anis BENLASMAR**, professeur dans plusieurs matières liées aux mathématiques appliquées, ainsi qu'à **M. Mohamed Hedi**, professeur de *Deep Learning*, qui m'a transmis l'essentiel de ce que je sais aujourd'hui en *data science*. Les connaissances ainsi transmises, conjuguées à l'exigence et à l'accompagnement dont j'ai bénéficié, m'ont permis de hisser ce projet à son plus haut niveau.

Ma gratitude va aussi à l'équipe **Esprit Tech** pour son accueil chaleureux et les conditions de travail propices à l'innovation qui m'ont été offertes.

Enfin, j'exprime ma profonde reconnaissance à **Mme Nardine Hanfi**, mon encadrante académique, qui m'accompagne depuis la quatrième année avec une constance, une patience et une disponibilité exemplaires. Sa confiance et ses orientations ont été déterminantes dans l'aboutissement de ce travail.

Que chacun trouve ici l'expression de ma sincère gratitude.

---

# PRÉVISION DES PRIX ET ANALYSE DE LA VOLATILITÉ SUR LE MARCHÉ DES CRYPTOMONNAIES

## Résumé Général

Le présent projet propose un travail de recherche et une solution intégrée d'aide à la décision et de gestion du risque sur le marché des cryptomonnaies, un univers caractérisé par une volatilité extrême et une spéculation élevée, qui freinent l'adoption et la création de valeur durable. L'objectif est double : mieux prévoir les prix et la volatilité et outiller les acteurs (traders, gestionnaires) pour mesurer, simuler et piloter les risques. Le projet répond à trois problèmes structurants : la prise de décision en environnement incertain, la difficulté à modéliser des dynamiques de marché instables, et l'optimisation des stratégies sous contrainte de risque.

Méthodologiquement, la démarche combine étude descriptive, modélisation économétrique et apprenante, gestion de risque et implémentation logicielle au sein d'une application web. L'architecture opère à l'échelle horaire et semi horaire : ingestion et alignement des flux (marché, actualités), module de prédiction qui publie les résultats via un serveur triton, et interface offrant indicateurs, prévisions et analyses pour la décision. Les fonctionnalités couvrent un tableau de bord marché, la simulation et la gestion de portefeuilles, des stress tests (historiques, facteurs, uniformes), ainsi qu'un assistant conversationnel capable de contextualiser ses recommandations et d'agir sur le système pour assister l'utilisateur.

En synthèse, la plateforme livre un chaînage complet des données aux décisions : collecte-préparation, prévisions horaires, mesures de risque et explicitation via l'application web.

**Mots-clés :** cryptomonnaies ; prévision ; volatilité ; GARCH/NGARCH ; XGBoost ; GRU ; gestion du risque ; simulation

---

---

# PRICE PREDICTION AND VOLATILITY ANALYSIS IN THE CRYPTOCURRENCY MARKET

## Abstract

This report presents a research work and an integrated solution for decision-making and risk management in cryptocurrency markets—an environment marked by extreme volatility and speculative behavior that hinders sustainable adoption. The project addresses three core challenges : decision-making under uncertainty, modeling unstable market dynamics, and strategy optimization under risk constraints. Our goal is to improve price and volatility forecasting while providing practitioners with reliable tools to measure, simulate, and manage risk.

Methodologically, we combine market analysis, econometric and machine learning models , and software engineering into a web application. The architecture runs on an hourly cadence : data ingestion and alignment, a prediction module exposing results through the backend API, and a front-end delivering indicators, forecasts, and analytics for decision support. Key features include a market dashboard, portfolio simulation and management, stress testing (historical, factor, uniform), and a conversational assistant that contextualizes recommendations.

Overall, the platform implements an end-to-end pipeline from data to decisions—hourly forecasts, risk metrics, and user-facing explanations—aiming for actionable insights while controlling computational costs via periodic processing and targeted user flows.

**Keywords** — cryptocurrencies, forecasting, volatility, GARCH/NGARCH, XGBoost, GRU, risk management, simulation, stress testing, web application.

**Keywords** : cryptocurrencies ; forecasting ; volatility ; GARCH/NGARCH ; XGBoost ; GRU ; risk management ; simulation ; stress testing ; web application.

# Table des matières

<b>Dédicace</b>	<b>I</b>
<b>Remerciements</b>	<b>II</b>
<b>Liste des sigles et abréviations</b>	<b>XIII</b>
<b>Introduction générale</b>	<b>1</b>
<b>Chapitre 1 Cadre du projet et méthodologie</b>	<b>3</b>
1.1 Introduction . . . . .	4
1.2 Contexte . . . . .	4
1.2.1 L'importance des cryptomonnaies aux 21 <sup>e</sup> siècle . . . . .	4
1.2.2 Les cryptomonnaies en Afrique . . . . .	5
1.2.3 Fonctionnement des cryptomonnaies . . . . .	5
1.2.4 Challenges . . . . .	7
1.3 Organisme d'accueil . . . . .	7
1.3.1 Présentation générale du groupe ESPRIT . . . . .	7
1.3.2 ESPRIT-Tech et l'équipe Graphics . . . . .	7
1.4 Présentation du projet . . . . .	8
1.4.1 Description . . . . .	8
1.4.2 Étude de l'existant, limites et solution proposée . . . . .	8
1.5 Choix méthodologique . . . . .	10
1.5.1 Panorama des méthodologies . . . . .	10
1.5.2 Synthèse et choix d'une approche hybride . . . . .	12
1.6 Choix des technologies . . . . .	12
1.6.1 Outils . . . . .	12
1.6.2 Framework . . . . .	14
<b>Chapitre 2 État de l'art</b>	<b>17</b>
2.1 Introduction . . . . .	18
2.2 Modèles économétriques . . . . .	18
2.2.1 Modèles économétriques : ARIMA . . . . .	18
2.2.2 ARCH/GARCH et extensions . . . . .	19
2.3 Modèles d'apprentissage (ML) et profonds (DL) . . . . .	21
2.3.1 Apprentissage automatique (RF, Gradient Boosting) . . . . .	21

---

## TABLE DES MATIÈRES

---

2.3.2 Random Forest (RF) . . . . .	21
2.3.3 XGBoost (XGB) . . . . .	21
2.4 Apprentissage profond . . . . .	22
2.4.1 Architecture du LSTM . . . . .	22
2.4.2 Architecture du GRU . . . . .	23
2.5 Conclusion . . . . .	23
<b>Chapitre 3 Compréhension et Préparation des données</b>	<b>24</b>
3.1 Introduction . . . . .	25
3.2 Analyse de données . . . . .	25
3.2.1 Analyse univariée . . . . .	25
3.2.2 Analyse bivariée . . . . .	34
3.3 Sources des données et architecture de collecte . . . . .	38
3.3.1 Organisation et provenance. . . . .	38
3.3.2 Distinction recherche / déploiement. . . . .	39
3.3.3 Préparation et alignement. . . . .	39
3.3.4 Jeux de sortie. . . . .	39
3.4 Jeu de données pour le déploiement : collecte et préparation . . . . .	40
3.4.1 Actualités et sentiment (CryptoPanic → FinBERT). . . . .	40
3.4.2 Prix et volumes (CoinGecko). . . . .	40
3.4.3 Alignement et fusion. . . . .	40
3.5 Conclusion . . . . .	41
<b>Chapitre 4 Modélisation</b>	<b>42</b>
4.1 Introduction . . . . .	43
4.2 Choix des modèles . . . . .	43
4.3 Pipeline de modélisation . . . . .	44
4.4 Modélisation . . . . .	45
4.4.1 ARIMA . . . . .	45
4.4.2 XGBoost . . . . .	56
4.4.3 GRU . . . . .	64
4.5 Evaluation des modèles . . . . .	73
4.6 Entraînement des modèles et suivi expérimental . . . . .	73
4.6.1 Analyse de la MFTR par crypto (modèles GRU) . . . . .	74
4.7 Foreward testing . . . . .	75
4.8 Conclusion . . . . .	76
<b>Chapitre 5 Tarification des options</b>	<b>77</b>
5.1 Introduction . . . . .	78
5.2 Gestion de risque . . . . .	78
5.3 Comparaison NGARCH et Black–Scholes (maturité une semaine) . . . . .	80
5.4 Simulation du payoff et du P&L d'un call européen . . . . .	82
5.5 Effet du strike sur la probabilité d'obtenir un P&L positif . . . . .	83

---

---

## TABLE DES MATIÈRES

---

5.6	Décroissance convexe de la prime $C(K)$ en fonction du strike . . . . .	83
5.7	Conclusion . . . . .	84
<b>Conclusion générale</b>		<b>85</b>
<b>Chapitre A Analyse des besoins</b>		<b>87</b>
<b>Chapitre B Architecture globale de la solution</b>		<b>90</b>
B.1	Introduction . . . . .	90
B.2	Vue globale de l'architecture . . . . .	90
B.3	Flux de données et processus de prédiction . . . . .	91
B.4	Conclusion . . . . .	91
<b>Chapitre C Parcours utilisateur et interfaces</b>		<b>92</b>
<b>Chapitre D Couches techniques détaillées</b>		<b>96</b>
D.1	Couche de données . . . . .	97
D.1.1	Description des entités . . . . .	97
D.2	Couche applicative (Backend) . . . . .	99
D.2.1	Endpoints principaux . . . . .	99

# Liste des tableaux

3.1	Ruptures (moyenne) détectées par <i>Voisin segmenté</i> . . . . .	29
3.2	Ruptures (variance) détectées par <i>Voisin segmenté</i> ( <i>SegNeigh</i> ) . . . . .	30
4.1	ARIMA(1, 0, 1) sur les log-rendements : métainformations d'ajustement . . . . .	46
4.2	ARIMA(1, 0, 1) : coefficients estimés . . . . .	46
4.3	ARIMA(1, 0, 1) : diagnostics sur les résidus . . . . .	46
4.4	Comparaison de lois pour les résidus de l'ARIMA(1, 0, 1) . . . . .	48
4.5	Paramètres GPD par queue (résidus ARIMA) . . . . .	53
4.6	Poids du splice et seuils associés (résidus ARIMA) . . . . .	53
4.7	GARCH(1, 1) — paramètres estimés et métriques (résidus ARIMA) . . . . .	53
4.8	NGARCH(1, 1) — paramètres estimés et métriques (résidus ARIMA) . . . . .	55
4.9	GARCH-X — paramètres estimés et métriques (résidus ARIMA) . . . . .	56
4.10	XGBoost — espace de recherche (Grid) . . . . .	57
4.11	Comparaison de lois pour les résidus de XGBoost . . . . .	58
4.12	Poids du splice et seuils associés (résidus XGBoost) . . . . .	60
4.13	Test ADF sur les résidus (trend = constant) . . . . .	60
4.14	Test LM d'hétéroscédasticité ARCH (lags = 48) sur les résidus XGBoost . . . . .	61
4.15	GARCH(1, 1) — paramètres estimés et métriques (résidus XGBoost) . . . . .	61
4.16	NGARCH(1, 1) — paramètres estimés et métriques (résidus XGBoost) . . . . .	62
4.17	GARCH-X — paramètres estimés et métriques (résidus XGBoost) . . . . .	63
4.18	GRU — meilleur WINDOW_SIZE (sélection par score pondéré) . . . . .	64
4.19	GRU — meilleur batch_size (sélection par score pondéré) . . . . .	65
4.20	Comparaison de lois pour les résidus de GRU . . . . .	66
4.21	Test ADF sur les résidus (trend = constant) . . . . .	68
4.22	Test LM d'hétéroscédasticité ARCH (lags = 48) sur les résidus . . . . .	68
4.23	Poids du splice et seuils associés (résidus GRU) . . . . .	68
4.24	GARCH(1, 1) — paramètres estimés et métriques (résidus GRU) . . . . .	69
4.25	NGARCH(1, 1) — paramètres estimés et métriques (résidus GRU) . . . . .	70
4.26	GARCH-X — paramètres estimés et métriques (résidus GRU) . . . . .	71
4.27	Tableau comparatif (Modèles × Métriques) . . . . .	72
4.28	Jeu de données et variables utilisées pour l'entraînement . . . . .	73
4.29	Hyperparamètres GRU . . . . .	74
4.30	Hyperparamètres XGBoost . . . . .	74
4.31	Artefacts produits pour l'inférence . . . . .	74

## LISTE DES TABLEAUX

---

D.2 Définitions des métriques utilisées (FinTSB et conventions financières) . . . . .	103
---	-----

# Table des figures

1.1	Évolution globale de l'activité en cryptomonnaie - Chainalysis . . . . .	4
1.2	Motivation à l'acquisition des cryptomonnaies - ADAN . . . . .	5
1.3	mapping concurrentiel . . . . .	9
1.4	CRISP-DM - Source :IBM . . . . .	10
1.5	TDSP - Source : datascience-pm.com . . . . .	10
1.6	Financial Time Series Benchmark - Source :HU et al. 2025 . . . . .	11
1.7	outils d'analyse et de modélisation . . . . .	12
1.8	Outil de Graphisme et de Maquettage . . . . .	13
1.9	outils de développement et de déploiement . . . . .	14
1.10	Framework back-end-front-end . . . . .	14
1.11	entraînement et déploiement de modèle . . . . .	15
1.12	Frise chronologique des réalisations (Semaines 1–24). . . . .	16
2.1	Schéma Random Forest - Source : ibm.com . . . . .	21
2.2	Logo Documentation officielle XGBoost - Source : xgboost.readthedocs.io . . .	22
2.3	Schéma LSTM - Source : FISCHER et KRAUSS 2017 . . . . .	22
2.4	Schéma GRU - Source : NG, KNITTEL et UHLER 2020 . . . . .	23
3.1	Distribution du prix du bitcoin de 2010 à 2025 . . . . .	26
3.2	Évolution du prix du Bitcoin de 2010 à 2025 . . . . .	26
3.3	Tendances de la série temporelle du Bitcoin avec la moyenne mobile . . . . .	27
3.4	Test de changement de structure par la méthode du voisin segmenté(Moyenne) .	28
3.5	Test de changement de structure par la méthode de PELT(Moyenne) . . . . .	29
3.6	Test de changement de structure par la méthode du voisin segmenté (Variance) .	29
3.7	Test de changement de structure par la méthode PELT (Variance) . . . . .	30
3.8	Rendements $r_t$ du Bitcoin de 2010 à 2025. . . . .	31
3.9	Log-rendements du Bitcoin de 2010 à 2025 $\ell_t = \ln P_t - \ln P_{t-1}$ . . . . .	32
3.10	ACF rendements $r_t$ Bitcoin 2010-2025. . . . .	32
3.11	PACF rendements $r_t$ Bitcoin 2010-2025. . . . .	33
3.12	Test de McLeod-Li rendements Bitcoin 2010-2025 : $p$ -valeurs par retard pour $r_t^2$ .	33
3.13	QQ-plot des rendements $r_t$ . La courbure en S révèle des queues plus épaisses que la normale. . . . .	34
3.14	Evolution des actifs à faible capitalisation (Polkadot-Litecoin) . . . . .	35
3.15	Evolution des actifs à forte capitalisation Bitcoin-Ethéreum) . . . . .	35
3.16	Test de corrélation linéaire entre le Bitcoin et l'Ethéreum . . . . .	35

---

## TABLE DES FIGURES

---

3.17	Test de corrélation non linéaire entre le Bitcoin et l’Ethereum . . . . .	36
3.18	Test de causalité sur les cryptomonnaies . . . . .	36
3.19	Evolution du prix de l’Or . . . . .	37
3.20	Corrélation linéaire entre le bitcoin et l’Or . . . . .	37
3.21	Corrélation non linéaire entre le bitcoin et l’Or . . . . .	37
3.22	Test de causalité entre l’or et le bitcoin . . . . .	38
4.1	Modèles de prévision . . . . .	43
4.2	Cadre de modélisation : des variables d’entrée à la moyenne $\mu_t$ , puis des résidus $\hat{i}_t$ à la variance conditionnelle $\sigma_t^2$ . . . . .	44
4.3	Performance des modèles économétriques . . . . .	45
4.4	Split du dataset pour le modèle ARIMA . . . . .	45
4.5	Graphique de la fonction d’autocorrélation ARIMA . . . . .	47
4.6	Histogramme des résidus et distribution ajustés - ARIMA . . . . .	47
4.7	QQ-Plot entre la loi de student et les résidus - ARIMA . . . . .	48
4.8	Ajustement des queues par la loi de student - ARIMA . . . . .	49
4.9	Evolution des paramètres de la loi GPD en fonction du seuil - ARIMA . . . . .	49
4.10	Analyse de l’ajustement des lois extrêmes sur les queues des résidus - ARIMA . . . . .	50
4.11	loi des mélanges . . . . .	51
4.12	Paramètres pour les queues de distribution - ARIMA . . . . .	53
4.13	Résultats d’ajustement du modèle GARCH (Loi Mixte) - ARIMA . . . . .	54
4.14	Variances conditionnelles et rendements prédictifs loi mixte GARCH - ARIMA . . . . .	54
4.15	Variances conditionnelles et rendements prédictifs loi mixte NGARCH - ARIMA . . . . .	55
4.16	Paramètre ajusté GARCH-X- ARIMA . . . . .	56
4.17	Variances conditionnelles et rendements prédictives loi mixte GARCH-X - ARIMA . . . . .	57
4.18	Importance des variables XGBoost . . . . .	57
4.19	Graphique de la fonction d’autocorrélation XGBoost . . . . .	58
4.20	Histogramme des résidus et distribution ajustés - XGBoost . . . . .	59
4.21	QQ-Plot entre la loi de student et les résidus - XGBoost . . . . .	59
4.22	Ajustement des queues par la loi de student - XGBoost . . . . .	59
4.23	Evolution des paramètres de la loi GPD en fonction du seuil - XGBoost . . . . .	59
4.24	Analyse de l’ajustement des lois extrêmes sur les queues des résidus - XGBoost . . . . .	60
4.25	Résultats d’ajustement du modèle GARCH (Loi Mixte) - XGBoost . . . . .	61
4.26	Variances conditionnelles et rendements prédictifs loi mixte NGARCH - XGBoost . . . . .	62
4.27	Paramètre ajusté GARCH-X- XGBoost . . . . .	63
4.28	Recherche window size et batch size optimale-GRU . . . . .	64
4.29	Ajustement du modèle GRU . . . . .	65
4.30	Graphique de la fonction d’autocorrélation GRU . . . . .	66
4.31	Histogramme des résidus et distribution ajustés - GRU . . . . .	66
4.32	QQ-Plot entre la loi de student et les résidus - GRU . . . . .	67
4.33	Ajustement des queues par la loi de student - GRU . . . . .	67
4.34	Evolution des paramètres de la loi GPD en fonction du seuil - GRU . . . . .	67

---

---

## TABLE DES FIGURES

---

4.35	Analyse de l'ajustement des lois extrêmes sur les queues des résidus - GRU . . . . .	68
4.36	Résultats d'ajustement du modèle GARCH (Loi Mixte) - GRU . . . . .	69
4.37	Variances conditionnelles et rendements prédictifs loi mixte NGARCH - GRU . . . . .	70
4.38	Paramètre ajusté GARCH-X- GRU . . . . .	71
4.39	MFTR moyenne par crypto pour les modèles GRU. Les barres montrent –MFTR tel qu'enregistré dans MLflow ; plus la barre est à gauche, plus la MFTR vraie est élevée. . . . .	75
4.40	Comparatif du forward testing (9–25 août, coût 0,4 %/transaction) : <b>(a)</b> XGBoost en haut ; <b>(b)</b> GRU en bas. . . . .	76
 5.1	 Haut : smile implicite $\sigma_{\text{impl}}(K)$ obtenu en inversant les prix NGARCH et, en pointillé, Black–Scholes calibré à l'ATM ( $\sigma = \sigma_{\text{ATM}}$ ). Bas : erreur de prix $E_{\text{BS}}(K) = \pi_{\text{BS}}(K; \sigma_{\text{ATM}}) - \pi_{\text{NG}}(K)$ . . . . .	81
5.2	Payoff (bleu) et P&L (orange) d'un call européen à $T$ . . . . .	82
5.3	Probabilité $\mathbb{P}(\text{P&L} > 0)$ (haut) et prime $\pi(K)$ (bas), en cohérence avec la <i>même</i> simulation de $S_T$ . Les points A et B marquent $K = 6988$ et $K = 7588$ . . . . .	83
 A.1	 prise de décision dans un environnement incertains - diagramme cas d'utilisation . . . . .	87
A.2	modéliser les dynamiques de marché - diagramme cas d'utilisation . . . . .	88
A.3	gestion de risque et optimisation des stratégies - diagramme cas d'utilisation . . . . .	88
 B.1	 architecture globale - Vue A : Contexte . . . . .	90
B.2	architecture globale - Vue A : Flux de données et ML . . . . .	91
 C.1	 Page d'accueil : Capture d'écran . . . . .	92
C.2	Page Marché : Capture d'écran . . . . .	93
C.3	Parcours utilisateur . . . . .	93
C.4	Page Simulation : Capture d'écran . . . . .	94
C.5	Page Gestion de risque : Capture d'écran . . . . .	94
C.6	Page Assistant : Capture d'écran . . . . .	95
 D.1	 Relations entre entités (vue synthétique). . . . .	97
D.2	Entités et attributs (sans relations) — vue élargie. . . . .	98
D.3	Architecture de déploiement de la plateforme POSA. . . . .	102

---

# LISTE DES SIGLES ET ABRÉVIATIONS

**ACF** Autocorrelation Function (fonction d'autocorrélation)

**ADF** Augmented Dickey–Fuller (test de Dickey–Fuller augmenté)

**AIC** Akaike Information Criterion (critère d'information d'Akaike)

**API** Application Programming Interface (interface de programmation d'applications)

**ARCH** AutoRegressive Conditional Heteroskedasticity (hétérosclélasticité conditionnelle autorégressive)

**ARIMA** AutoRegressive Integrated Moving Average (moyenne mobile intégrée autorégressive)

**ARR** Annualized Rate of Return (taux de rendement annualisé)

**ASR** Adjusted Sharpe Ratio (ratio de Sharpe ajusté)

**BIC** Bayesian Information Criterion (critère d'information bayésien)

**CRISP-DM** Cross-Industry Standard Process for Data Mining (processus standard inter-industrie pour l'exploration de données)

**GARCH** Generalized ARCH (GARCH généralisé)

**GARCH-X** GARCH with eXogenous variables (GARCH avec variables exogènes)

**GJR-GARCH** Glosten–Jagannathan–Runkle GARCH (GARCH de Glosten–Jagannathan–Runkle)

**GPD** Generalized Pareto Distribution (loi de Pareto généralisée)

**GRU** Gated Recurrent Unit (unité récurrente à portes)

**IA** intelligence artificielle

**LSTM** Long Short-Term Memory (réseau à mémoire à long court terme)

**MAE** Mean Absolute Error (erreur absolue moyenne)

**MCFD** Mean Correct Forecast Direction (taux moyen de bonne direction des prévisions)

**MFTR** Mean Forecasting Trading Return (rendement moyen d'une stratégie de trading basée sur la prévision)

**MSE** Mean Squared Error (erreur quadratique moyenne)

**NGARCH** Nonlinear GARCH (GARCH non linéaire)

**ODD** Objectifs de Développement Durable

**PACF** Partial Autocorrelation Function (fonction d'autocorrélation partielle)

**PELT** Pruned Exact Linear Time (algorithme de détection de ruptures)

**PoS** Proof of Stake

**PoW** Proof of Work

**QQ** Quantile–Quantile (graphe quantile–quantile)

**TDSP** Team Data Science Process (processus Team Data Science)

---

# INTRODUCTION GÉNÉRALE

L'année 2021 a marqué un tournant majeur dans le domaine des technologies, avec une accélération sans précédent des avancées, notamment en IA. Ce domaine, en pleine effervescence, a profondément transformé de nombreux secteurs tels que le graphisme, la médecine, l'éducation et l'ingénierie. Aujourd'hui, le rôle de l'ingénieur informatique ne se limite plus à la résolution de problèmes métiers par des compétences techniques ; il implique une synergie entre *soft skills*, expertise approfondie et capacité à tirer parti des outils d'apprentissage automatique pour optimiser les processus.

C'est dans cette logique qu'a été conçu le cursus *Génie Informatique – spécialité Data Science* à l'école Esprit. Il vise, sur deux années de spécialisation, à doter les étudiants de compétences avancées en communication, en mathématiques appliquées, en *machine learning*, et d'une réelle capacité d'adaptation aux enjeux métier. Dans cette même dynamique, j'ai eu l'honneur de suivre le parcours *Data Science et Actuariat*, mis en place pour adapter la gestion des risques financiers à l'ère de l'IA. Ce programme cible en priorité les domaines de l'assurance et du trading, dans lesquels les investisseurs ont besoin d'outils sophistiqués – tels que *AI for Alpha* – pour mieux évaluer les risques et prendre des décisions éclairées.

Cependant, malgré l'enthousiasme que suscitent ces avancées, certains marchés à fort potentiel restent peu exploités en raison de l'absence d'outils adaptés. Le marché des cryptomonnaies, en particulier, incarne une opportunité unique : il repose sur une architecture décentralisée, sans intermédiaire, offrant une certaine liberté financière. Cette caractéristique en fait un levier stratégique pour les pays africains, confrontés à des défis monétaires majeurs. Les cryptomonnaies peuvent ainsi représenter une alternative aux monnaies controversées comme le Franc CFA, tout en proposant un cadre d'échange plus simple et sécurisé.

Néanmoins, ce marché souffre d'une volatilité extrême et d'une spéculation excessive, ce qui freine son adoption comme source de revenus durable. Une gestion renforcée des risques permettrait non seulement de stabiliser ce marché, mais également de contribuer à la création d'emplois via les *fintechs* (ODD 9), de stimuler l'innovation à travers le développement d'infrastructures blockchain, et de promouvoir une approche plus responsable sur le marché qui à chaque agitation entraîne des coûts élevés en termes de calcul (ayant un impact négatif sur l'environnement).

Face à cette situation il était cruciale de comprendre comment :

*Dans un contexte de forte volatilité et d'incertitude propre au marché des cryptomonnaies, concevoir et mettre en œuvre une solution intégrée alliant étude qualitative et modélisation quantitative, afin d'améliorer la prévision des prix et l'analyse de la volatilité, et fournir aux acteurs du marché des outils fiables pour la gestion des risques et l'optimisation de leurs stratégies ?*

Fort de cette problématique se dérivent trois problèmes. Premièrement, la prise de décision dans un environnement incertain. La difficulté à modéliser les dynamiques de marché. Et enfin, la gestion des risques et l'optimisation des stratégies. Pour ce faire, il sera essentiel de premièrement étudier la dynamique de marché, puis réussir à modéliser les variations historiques des prix en utilisant des modèles avancés, qu'ils soient économétriques ou bien encore d'apprentissage automatique, afin de développer des modèles prédictifs de volatilité et de prix. Pour la gestion de risque nous nous concentrerons sur la tarification des options. Le tout sera donc embarqué dans une solution technique, notamment une application web de gestion des risques et de prise de décision.

---

---

# CHAPITRE 1

---

## CADRE DU PROJET ET MÉTHODOLOGIE

### Sommaire

---

1.1	Introduction . . . . .	4
1.2	Contexte . . . . .	4
1.2.1	L'importance des cryptomonnaies aux 21 <sup>e</sup> siècle . . . . .	4
1.2.2	Les cryptomonnaies en Afrique . . . . .	5
1.2.3	Fonctionnement des cryptomonnaies . . . . .	5
1.2.4	Challenges . . . . .	7
1.3	Organisme d'accueil . . . . .	7
1.3.1	Présentation générale du groupe ESPRIT . . . . .	7
1.3.2	ESPRIT-Tech et l'équipe Graphics . . . . .	7
1.4	Présentation du projet . . . . .	8
1.4.1	Description . . . . .	8
1.4.2	Étude de l'existant, limites et solution proposée . . . . .	8
1.5	Choix méthodologique . . . . .	10
1.5.1	Panorama des méthodologies . . . . .	10
1.5.2	Synthèse et choix d'une approche hybride . . . . .	12
1.6	Choix des technologies . . . . .	12
1.6.1	Outils . . . . .	12
1.6.2	Framework . . . . .	14

---

## 1.1 Introduction

Ce chapitre établit le cadre général du projet et précise la logique qui guide l'ensemble de la démarche. Nous clarifions d'abord le contexte et les motifs qui rendent le sujet pertinent, puis nous présentons l'organisme d'accueil et la place effective du projet en son sein. Sur cette base, nous définissons les contours du travail à mener ainsi que les livrables attendus, avant d'expliciter la démarche méthodologique retenue et les choix technologiques qui soutiennent l'implémentation.

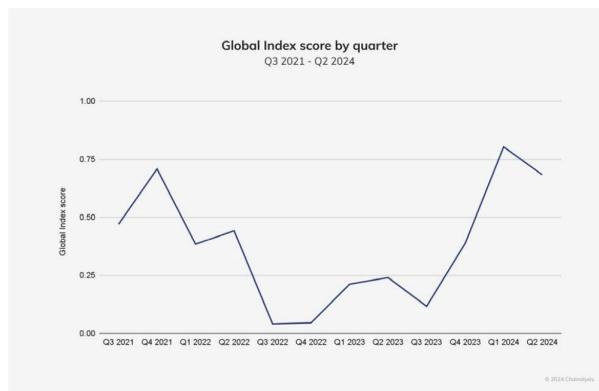
La suite est structurée de manière progressive : la section 1.2 revient sur le contexte et les besoins auxquels le projet répond ; la section 1.3 décrit l'organisme d'accueil et les contraintes réelles du terrain ; la section 1.4 expose le périmètre opérationnel du projet ; la section 1.5 précise l'orientation méthodologique ; enfin, la section 1.6 synthétise les principaux choix techniques. L'objectif est de disposer d'un fil conducteur clair, cohérent avec les exigences du problème et les ressources effectivement mobilisables.

## 1.2 Contexte

La première chose à laquelle on pense lorsqu'on entend cryptomonnaie, c'est de l'arnaque. Ce constat aussi amer cache un problème sous-jacent. Mais, avant de rentrer dans le vif du sujet, il est important de comprendre l'importance qu'ont pris les cryptomonnaies durant ces dernières années, quelles sont leur fonctionnement, mais aussi quels sont les problèmes qui nous ont poussés à mener à bien ce projet.

### 1.2.1 L'importance des cryptomonnaies aux 21<sup>e</sup> siècle

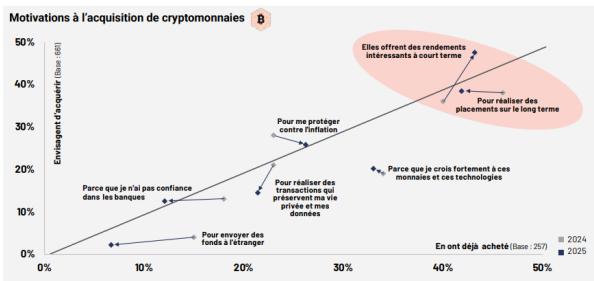
D'après le rapport de CHAINALYSIS 2024 paru en octobre 2024, intitulé "The 2024 Geography of Crypto Report", Le nombre d'utilisateurs de crypto-actifs a fortement progressé au cours des dernières années, avec une adoption de plus en plus marquée sur tous les continents. Les transferts de valeur en cryptomonnaies ne se limitent plus aux échanges spéculatifs, mais s'étendent aux paiements, aux transferts transfrontaliers et à l'épargne numérique.



**FIGURE 1.1 – Évolution globale de l'activité en cryptomonnaie - Chainalysis**

Cette adoption massive figure 1.1 peut être expliquée d'une part par la diversification et

l'innovation dans le Web3. On parle donc de finances décentralisées et d'intégration dans des tendances majeures telles que l'IA. D'autre part, nous pouvons aussi compter les facteurs économiques et politiques qui encouragent les investisseurs à sauvegarder leurs avoirs dans les monnaies numériques. Face à l'inflation et ou à la dévaluation de monnaie locale, les particuliers peuvent se refugier dans des "stables coins"<sup>1</sup>.



**FIGURE 1.2 – Motivation à l'acquisition des cryptomonnaies - ADAN**

Mise à part la fonction bouclier, les cryptomonnaies sont la nouvelle route de l'or. D'après le rapport DÉVELOPPEMENT DES ACTIFS NUMÉRIQUES (ADAN) 2025, c'est une augmentation de 15% figure 1.2 de personnes envisageant acheter des cryptomonnaies car elles offrent un rendement plus fructueux à court terme.

## 1.2.2 Les cryptomonnaies en Afrique

De 2023 à 2024, l'Afrique, plus précisément l'Afrique subsaharienne, représente une part assez modeste dans l'économie des cryptomonnaies, soit 2,7% des transactions mondiales (CHAINALYSIS 2024). Cette part aussi minime peut s'expliquer par une quantité d'infrastructures techniques insuffisante et une absence de vulgarisation. Néanmoins, la fatalité n'est pas au rendez-vous. L'Afrique présente des pays leaders tels que le Nigeria(deuxième mondiale, l'Éthiopie, le Kenya et l'Afrique du Sud, qui sont des figures phares de la cryptomonnaie en Afrique et dans le monde.

## 1.2.3 Fonctionnement des cryptomonnaies

Le terme *cryptomonnaie*, ou plus précisément *crypto-actif*, désigne un actif numérique reposant sur une technologie de registre distribué appelé *blockchain*. Contrairement à une monnaie traditionnelle, un crypto-actif ne possède pas de cours légal, et sa valeur n'est pas garantie par une autorité centrale. Elle résulte uniquement du jeu de l'offre et de la demande sur les plateformes d'échange (AUTORITÉ DES MARCHÉS FINANCIERS 2024).

La technologie blockchain fait partie de la couche d'implémentation des systèmes distribués. Son objectif principal est de garantir l'intégrité des données sans recourir à un tiers de confiance. Elle y parvient en assurant sept fonctions essentielles : la description et la protection de la propriété (via des clés publiques et privées ANCRYPTO 2024), l'enregistrement immuable des transactions, la diffusion des registres dans un environnement non fiable (grâce aux mécanismes

1. Cryptomonnaies ayant une parité fixe avec des monnaies plus ou moins fortes comme le Dollar ou l'Euro

de consensus comme la PoW CRYPTOAST 2024 ou la PoS COINHOUSE 2024), l'ajout de nouveaux blocs, et enfin la résolution des divergences (ou forks) COINBASE 2024.

Afin de bien comprendre le rôle et la spécificité de la *blockchain* dans les systèmes distribués modernes, il est nécessaire de la replacer dans l'architecture logicielle globale d'un système. Cette architecture peut être schématiquement divisée en deux couches complémentaires : la couche applicative et la couche d'implémentation.

La **couche applicative** se concentre sur les besoins de l'utilisateur. Elle concerne ce que le système est censé faire : c'est le *quoi*. À l'inverse, la **couche d'implémentation** traite de la manière dont ces fonctionnalités sont réalisées : c'est le *comment*. Elle englobe les aspects techniques, tels que les protocoles, la sécurité, la gestion des erreurs ou encore la performance. Dans cette logique, les aspects *fonctionnels* d'un système informatique décrivent ses capacités visibles (envoi de transactions, affichage de solde, exécution de contrats...), tandis que les aspects *non fonctionnels* décrivent des exigences sous-jacentes, souvent invisibles pour l'utilisateur final mais cruciales pour le bon fonctionnement du système. L'un des aspects non fonctionnels les plus importants ici est l'**intégrité**. Celle-ci regroupe trois éléments majeurs (DRESCHER 2017) :

- *L'intégrité des données* : garantir que les données ne sont ni altérées, ni perdues.
- *L'intégrité comportementale* : assurer un fonctionnement logique et prévisible du système.
- *La sécurité* : prévenir les accès non autorisés ou malveillants.

La plupart des défaillances logicielles sont la conséquence d'une atteinte à l'un de ces piliers : perte de données, comportements incohérents ou failles de sécurité.

Un système peut être centralisé ou distribué. Dans un système centralisé, une entité unique concentre le contrôle, ce qui constitue un point de défaillance critique. En revanche, un **système distribué** repose sur la coopération entre plusieurs machines autonomes, sans autorité centrale.

C'est dans ce contexte que s'insère la blockchain : elle fait partie de la couche d'implémentation d'un système logiciel distribué, et son but fondamental est d'assurer l'intégrité globale de ce système, sans dépendre d'un acteur central. Elle y parvient par un mécanisme d'écriture immuable et partagé, combiné à un consensus entre pairs.

Autrement dit, la **blockchain est une réponse logicielle à un besoin non fonctionnel critique : maintenir la confiance et l'intégrité dans un environnement décentralisé et potentiellement non fiable**.

Après avoir parlé du cadre global de la cryptomonnaie, de la technologie qui la supporte, nous parlerons maintenant des acteurs qui font tourner le marché et comprendrons quels sont les différents challenges qui s'y posent.

Dans le marché de la cryptomonnaie, on retrouve les développeurs qui sont au cœur de l'innovation. Ensuite, les investisseurs, qu'ils soient particuliers ou institutionnels, misent sur le potentiel de valorisation à long terme des crypto-actifs. Les traders, quant à eux, cherchent à exploiter la forte volatilité du marché à travers des stratégies souvent sophistiquées, allant du day-trading à l'algorithme. Les régulateurs, bien que peu comparé au marché traditionnel, jouent un rôle essentiel dans la définition des cadres juridiques. En parallèle, se présentent des plateformes d'échange, qu'elles soient centralisées comme Binance ou Coinbase, ou décentralisées telles que Uniswap et PancakeSwap, qui facilitent l'accès aux crypto-actifs pour des millions d'utilisateurs.

### 1.2.4 Challenges

Les entreprises ou encore les investisseurs institutionnels se lancent de plus en plus sur le marché de la cryptomonnaie. Bien vrai que la volatilité sous-jacente reste très souvent enchaînée sur ce marché, les activités telles que les négociations sur les contrats à terme ou encore les produits dérivés qui sont contenus dans les portefeuilles de ces institutions sont fortement touchées par les fluctuations (FINANCIAL STABILITY BOARD 2022). Et le problème ne fera qu'empirer qu'au fur et à mesure que l'adoption des cryptomonnaies montera en flèche. A noter que cette augmentation est souvent due à l'émergence de nouvelles technologies basées sur la blockchain. En Afrique, le contexte économique pousse les acteurs financiers à se projeter vers le marché des monnaies numériques. Cependant, ceci n'est pas sans conséquences.

Semblable aux addictions au jeu, la dopamine dès la première victoire incite les gens à revenir. Mais souvent, comme pour le jeu, il ne s'agit même pas de gagner de l'argent en soi - il s'agit de l'excitation que procure le risque (WELLER 2025). Enfin, certains problèmes empêchent l'adoption généralisée des cryptomonnaies, leur volatilité.

## 1.3 Organisme d'accueil

### 1.3.1 Présentation générale du groupe ESPRIT

Fondé avec la volonté d'allier excellence académique, innovation et ouverture sur le monde, le groupe **ESPRIT** s'impose aujourd'hui comme un acteur majeur de l'enseignement supérieur privé en Tunisie. Fort d'une vision résolument tournée vers les enjeux contemporains, il place la recherche appliquée, le développement technologique et la formation d'ingénieurs et de managers compétents au cœur de sa mission. L'institution se distingue par sa capacité à conjuguer rigueur scientifique et adaptation aux réalités socio-économiques, en cultivant un lien fort avec les secteurs industriels, les communautés scientifiques et les initiatives entrepreneuriales.

Au fil des années, **ESPRIT** a su bâtir un environnement académique dynamique où l'apprentissage ne se limite pas aux salles de classe, mais s'étend à des projets concrets, des partenariats stratégiques et des événements d'envergure internationale. L'accueil de la 20<sup>e</sup> *Conférence Internationale CDIO* en 2024, réunissant des experts venus de 26 pays, illustre parfaitement cette ouverture et cet engagement envers l'innovation pédagogique. Plus qu'un établissement d'enseignement, **ESPRIT** se positionne comme un catalyseur de progrès, contribuant activement au développement durable et à la compétitivité de la Tunisie à l'échelle mondiale.

### 1.3.2 **ESPRIT-Tech et l'équipe Graphics**

Créé en 2010, **ESPRIT-Tech** est le pôle Recherche, Développement et Innovation (RDI) du groupe. Sa mission est de définir et mettre en œuvre les politiques et stratégies de recherche de l'institution, en favorisant la collaboration entre enseignants-chercheurs, étudiants, entreprises et acteurs publics. **ESPRIT-Tech** couvre un large spectre thématique allant des technologies de l'information à l'ingénierie industrielle, en passant par les sciences appliquées et la modélisation

avancée. En plaçant l'impact socio-économique au centre de ses priorités, la structure encourage des projets qui allient pertinence scientifique, innovation et utilité concrète pour la société.

Au sein d'ESPRIT-Tech, l'équipe **GRAPHICS**, intégrée à la sous-structure, constitue un pôle d'expertise reconnu dans le domaine de la gestion de risque. L'approche pluridisciplinaire de l'équipe permet de répondre à des problématiques complexes en combinant rigueur analytique et simulations numériques avancées. C'est au sein de cette unité que s'inscrit mon présent stage, offrant un cadre de recherche stimulant, en prise directe avec des enjeux scientifiques et techniques d'actualité.

## 1.4 Présentation du projet

Le contexte étant établi et les challenges mis en lumière, nous présenterons dans cette section les différents problèmes, les objectifs et la solution proposée.

### 1.4.1 Description

Les marchés des cryptomonnaies ont su faire leurs places durant cette décennie. Grâce au nombre quasi inexistant d'intermédiaire financier, et d'une transparence à tous les niveaux, ce secteur est devenu la route de la soi des traders à la recherche de profit rapide. Cependant, en carence de régulateurs, le marché est purement spéculatoire marqué par une volatilité importante. Les particuliers désireux de se lancer avec des outils d'analyse conventionnels finissent par regretter leurs choix. Dans un contexte de forte volatilité et d'incertitude inhérente au marché des cryptomonnaies.

Trois problèmes sont donc à résoudre.

- La prise de décision dans un environnement incertain
- La difficulté à modéliser les dynamiques de marché
- la gestion des risques et optimisation de stratégies

Afin de résoudre ces problèmes, nous étudierons les dynamiques du marché, modéliseront les variations historiques des prix en utilisant des techniques avancées (économétriques et d'apprentissage automatique), développerons des modèles prédictifs des prix. Enfin, nous mettrons sur pied une application de gestion de risque pour permettre aux acteurs du marché de la cryptomonnaie d'agir de manière plus responsable, tout en se protégeant des risques sous-jacent.

### 1.4.2 Étude de l'existant, limites et solution proposée

En prenant en compte l'actualité de ce cadre de travail, il est important, premièrement, de mettre une étude de l'existence sur le cadre de recherche. Deuxièmement, de prendre aussi l'aspect technique, notamment avec les différentes solutions existantes sur le marché.

**Modèles et recherches.** Les approches *ARCH/GARCH* ENGLE 1982 ; BOLLERSLEV 1986 constituent la base de la modélisation de la volatilité conditionnelle. Pour la gestion du risque de queue, l'*Extreme Value Theory* (POT/GPD) est largement mobilisée MCNEIL 1999 ; MCNEIL

et FREY 2000, tandis que la validation des modèles de risque s'appuie sur HU et al. 2025. Plus récemment, des architectures apprenantes et hybrides (*LSTM/GRU* couplées à *GARCH*) montrent des gains sur des indices actions et, dans certains cas, sur des séries financières complexes ROSZYK et SLEPACZUK 2024; KAKADE et al. 2022. Cependant, appliquées aux cryptomonnaies, ces méthodes affrontent des faits stylisés marqués : *queues épaisse*s, *sauts*, *ruptures de régime* et forte instationnarité CUNHA et al. 2020; SHEN et al. 2020; SAEF et al. 2024, ce qui dégrade les hypothèses des modèles linéaires et la stabilité hors-échantillon.

**Solutions techniques.** Côté outils, AI FOR ALPHA 2025 cible des clients professionnels avec des solutions d'IA intégrant gestion du risque et prévision marché . Des plateformes *open source* comme *OpenBB* proposent des tableaux de bord de risque, des données crypto et une API extensible OPENBB 2025b ; OPENBB 2025a qui sont adapté aux plus aguéri . D'autres acteurs *data/analytics* — *Messari*, *Glassnode*, *Kaiko* — offrent recherche, on-chain et APIs, souvent avec des plans *Pro/Enterprise* MESSARI 2025 ; GLASSNODE 2025 ; KAIKO 2025. En DeFi<sup>2</sup>, *Gauntlet* fournit une gestion du risque institutionnelle axée protocoles GAUNTLET 2025. Toutes ces solutions précédentes sont sujettes à des limites : l'accessibilité ou le prix (plans entreprises), orientation très technique sans prise en compte des spécificités particulières du marché de la cryptomonnaie, faible ouverture du code (sauf OpenBB).

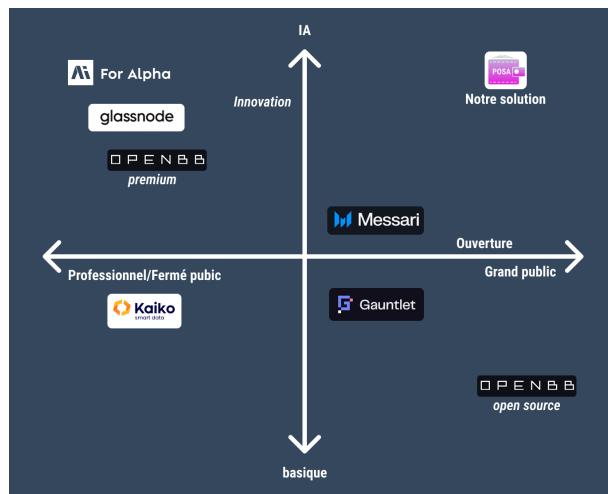


FIGURE 1.3 – mapping concurrentiel

**Positionnement du présent travail.** Notre contribution répond à ces lacunes par une méthodologie hybride adaptée aux marché de la crypto et, une solution grand public User-friendly orienté IA figure 1.3.

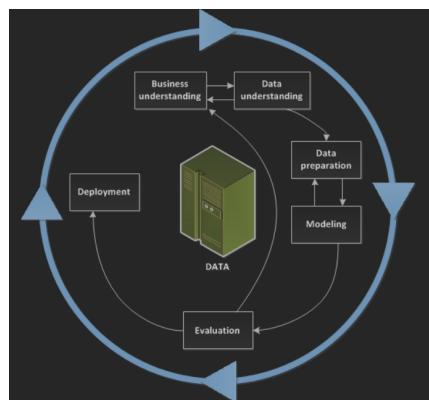
---

2. Finance décentralisée

## 1.5 Choix méthodologique

### 1.5.1 Panorama des méthodologies

#### CRISP-DM

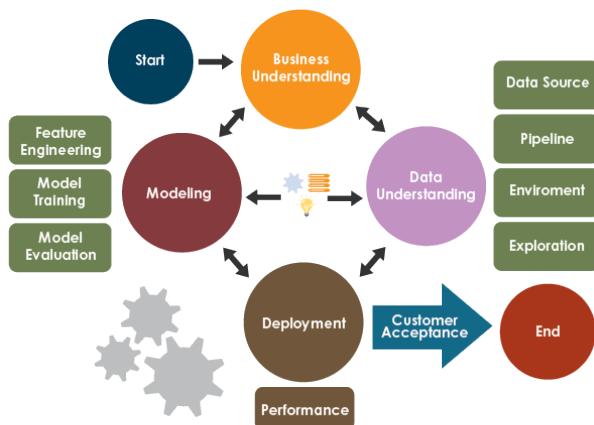


**FIGURE 1.4 – CRISP-DM - Source :IBM**

Le *Cross-Industry Standard Process for Data Mining (CRISP-DM)* est une démarche en six phases qui structure un projet d’analyse prédictive de l’expression du besoin jusqu’au déploiement figure 1.4.

**Limites dans notre contexte.** CRISP-DM reste volontairement généraliste. Dans un cadre marchés financiers/crypto et *risk management*, il donne peu de guidance opérationnelle sur l’évaluation des modèles.

#### TDSP



**FIGURE 1.5 – TDSP - Source : datascience-pm.com**

La *Team Data Science Process (TDSP)* figure 1.5, proposée par Microsoft, est une démarche itérative pensée pour livrer des modèles en production. Elle s'articule autour de quatres phases : La compréhension métier, L' acquisition & la compréhension des données, la modélisation (incluant feature engineering et validation), et le déploiement.

**Limites dans notre contexte.** TDSP excelle pour l'ingénierie et la mise en prod (avec un accent collectif), mais ce n'est pas un cadre scientifique d'évaluation de séries temporelles financières. Il n'impose pas un protocole de backtesting/métriques finance.

### Financial Time Series Benchmark (FinTSB)

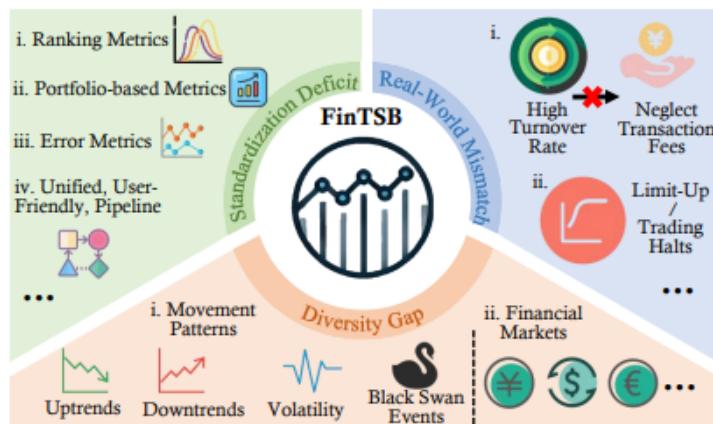


FIGURE 1.6 – Financial Time Series Benchmark - Source :HU et al. 2025

FinTSB est un *benchmark* récent dédié à la prévision de séries financières. Il part de trois lacunes observées dans la littérature : (1) **Diversity Gap** — les jeux de données couvrent mal la diversité des régimes de marché ; (2) **Standardization Deficit** — absence de protocole d'évaluation unifié ; (3) **Real-World Mismatch** — des hypothèses irréalistes (short selling, omission des frais, etc.). Côté données, FinTSB construit 20 jeux couvrant 4 *movement patterns* (hausse, baisse, volatilité, événements extrêmes) à partir de 15 ans d'historiques, avec tokenisation/anonymisation, normalisation par titre et splits 7 : 1 : 2. Côté évaluation, il unifie trois familles de métriques : *ranking* (IC, RankIC, etc.), *portefeuille* (ARR, AVol, Sharpe, MDD, IR) et *erreur* (MSE, MAE), en rappelant que de faibles erreurs ne garantissent pas la profitabilité. Enfin, FinTSB propose un pipeline unifié en quatre couches : **données** (prétraitement, dataloader par régimes), **entraînement** (backbones classiques à profonds/RL/LLM), **backtesting** (stratégies avec coûts et contraintes de marché), **feedback** (boucle de rétroaction pour affiner modèles et paramètres).

**Limites dans notre contexte.** FinTSB est un *benchmark* d'évaluation, pas une méthodologie de gestion de projet.

## 1.5.2 Synthèse et choix d'une approche hybride

Au vu de nos besoins (évaluation standardisée, forte composante financière, prise en compte de l'aspect déploiement), nous retenons une approche **hybride** :

- **TDSP** pour piloter le *lifecycle* (organisation en dépôts Git, structure de projet, déploiement, suivis).
- **FinTSB** comme *spécification d'évaluation* pour la modélisation.

Ce couplage préserve la rigueur d'ingénierie (TDSP) tout en garantissant une évaluation pertinente pour le risque financier (FinTSB).

## 1.6 Choix des technologies

### 1.6.1 Outils

#### Analyse et modélisation

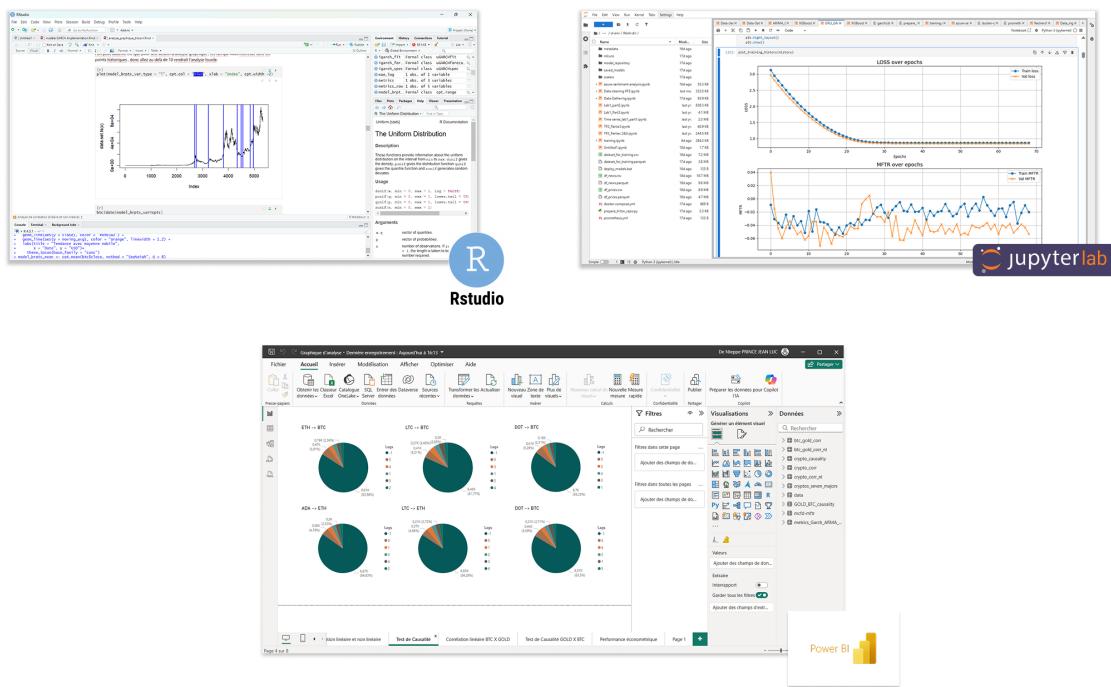


FIGURE 1.7 – outils d'analyse et de modélisation

l'analyse et la modélisation est au cœur de notre travail, raison pour laquelle il fallait des outils reconnus pour leur utilité.

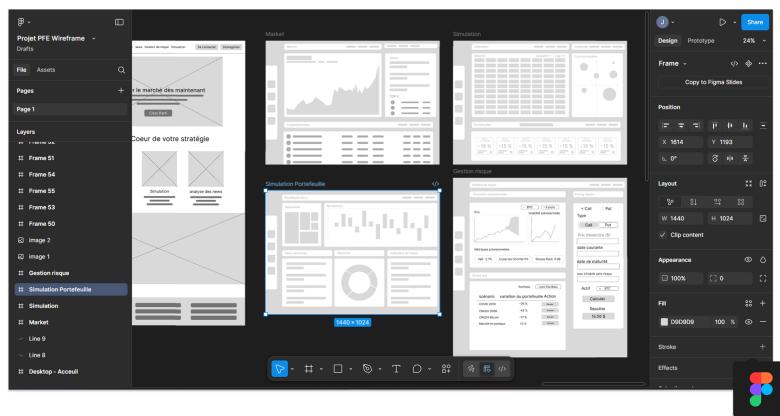
- **JupyterLab** D'une interface riche, l'outil m'a accompagné en grande partie dans tout mon projet : du traitement des données à la modélisation. Comparé à Jupyter notebook il était plus complet<sup>3</sup>. Bien évidemment des solutions en ligne existent comme Google Colab et Kaggle. Mais ils sont instables (déconnexion multiples) bien qu'ils fournissent des unités de calculs bien plus performant. A noter aussi la difficulté à pouvoir dupliquer

3. gestion multiple de fichier, opération extensible comme l'ouverture de terminale, etc .

l'environnement d'entraînement sur le server VPS pour le déploiement des modèles avec des environnement non contrôlé tel que Kaggle et Google Colab il est difficile de gérer cela.

- **Rstudio** Adosser au langage R, il est bien plus rapide en ce qui concerne les tests statistiques.
- **POWER BI** Simple, puissant. C'est le parfait outils de visualisation, complémentaire avec Jupyter Lab. Il est utilisé directement sur les données nettoyées et transformées.

## Maquettage



**FIGURE 1.8 – Outil de Graphisme et de Maquettage**

**Figma** : simple, facile et accessible fut mon outil de prédilection pour tout ce qui était graphisme. Il se démarque par efficacité. Pas besoin d'installation, disponible à travers internet il était facile de travailler sur la maquette de l'application et de retrouver mon travail n'importe où.

## développement et déploiement

Faire des recherches c'est bien, mais rendre la solution utile et utilisable c'est bien mieux. Les outils qui suivent ont contribuer fortement à la réussite de ce projet.

- **Visual Studio Code** : Open Source, fiable et extensible. C'était l'outil idéal pour le développement du back-end et du front-end de l'application.
- **Terminal VPS Contabo** : l'outil majeur dans la gestion des fichiers et des conteneurs docker sur le serveur virtuel.
- **Portail Azure** : Points de contrôle de tous les services azure utilisés notamment les Azure functions et Azure AI Services. Le choix d'utiliser une solution cloud est principalement dû au prérequis professionnels demandé aux ingénieurs demandé dans le milieu de l'IA. Aujourd'hui l'IA/apprentissage automatique n'est plus à la portée des ordinateurs traditionnelles il est nécessaire de savoir utiliser ou construire des services clouds orienté IA.

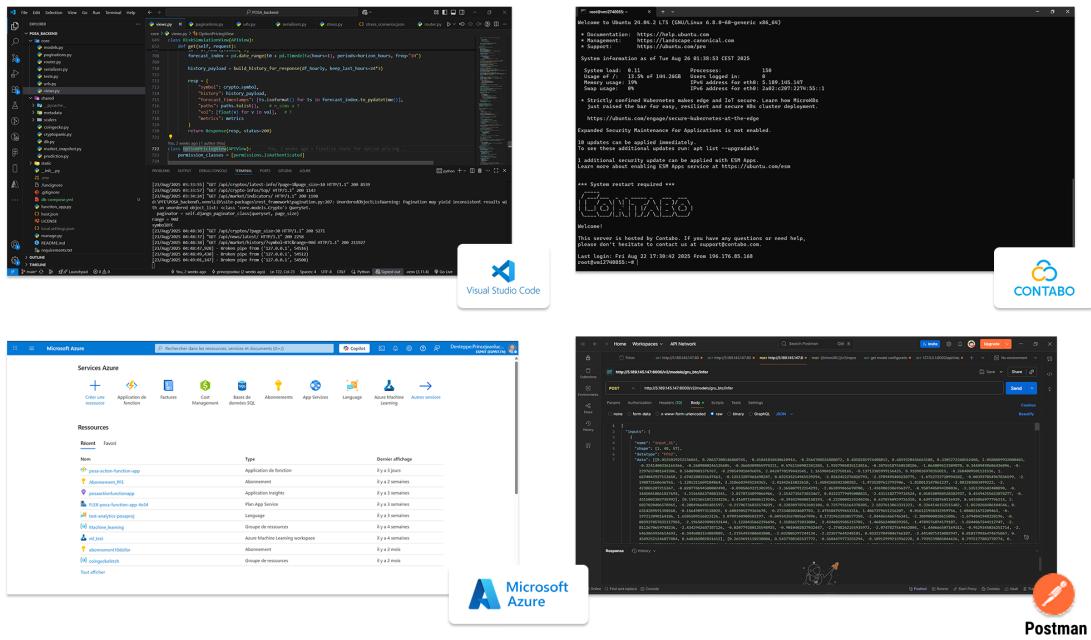


FIGURE 1.9 – outils de développement et de déploiement

- **Postman** : l'outil de test incontournable, il m'a permis de tester les endpoints de mon API back-end et du serveur Triton.

## 1.6.2 Framework

### Back-end & Front-end

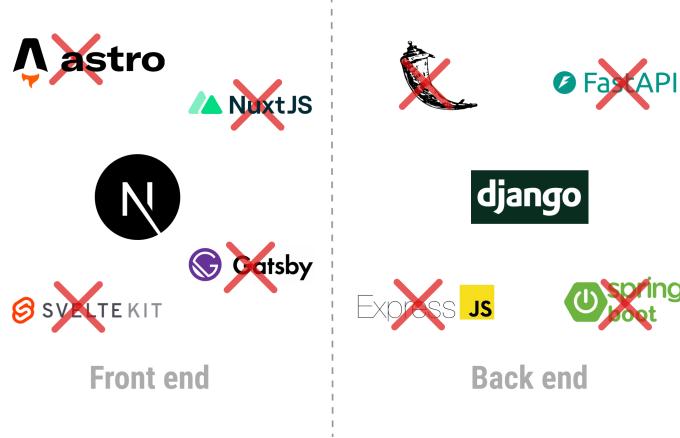


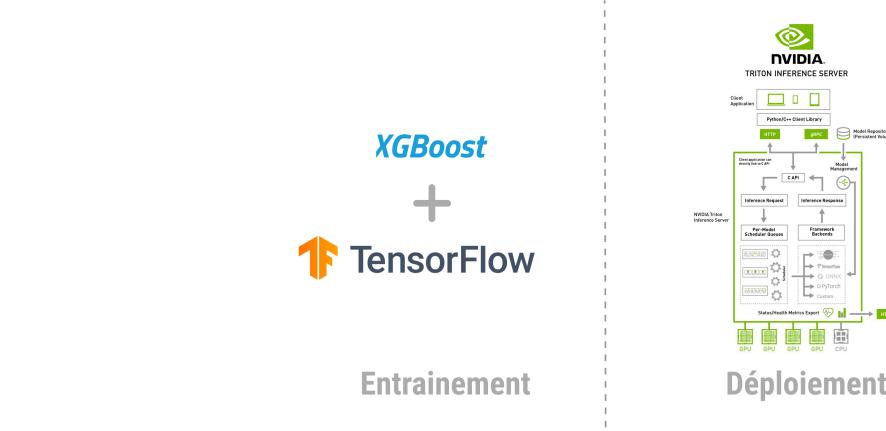
FIGURE 1.10 – Framework back-end-front-end

Du côté back-end des solutions telles que Flask et FastAPI sont connus pour leur simplicité et leur légèreté. Cependant, ils ne sont pas assez robuste pour accueillir le projet. Une solution comme Django est le candidat parfait : robuste (supporte bien les montées en charge) sécurisé (protection CSRF, injection SQL, XSS), une couche ORM (plus besoin de coder directement les requêtes SQL). Une alternative à django serait Spring boot : puissant et fonctionnant sur JAVA.

Mais profitant déjà d'un environnement python et en prenant en compte la courbe d'apprentissage (j'ai déjà eu à travailler avec) Django reste la meilleure option.

Côté frontend des solutions simples tels que Astro, Gatsby, SvelteKit et Nuxt existent . Cependant Next.js est une solution robuste, optimisé.

### Entraînement des modèles et déploiement



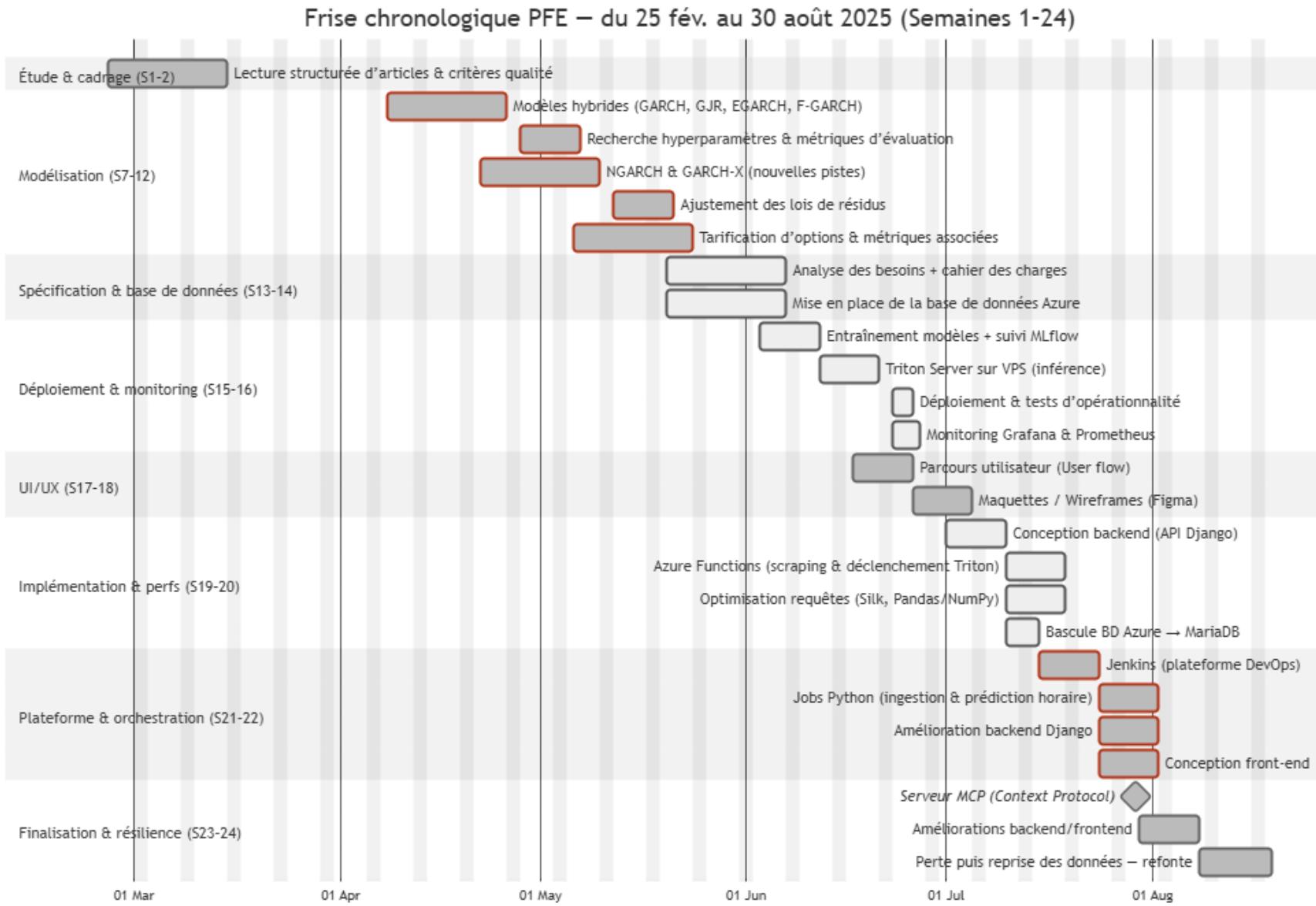
**FIGURE 1.11 – entraînement et déploiement de modèle**

L'entraînement des modèles est fait avec TensorFlow et des librairies Tierces (XGboost). Certes une solution comme pytorch était possible : simple, intuitif et offrant de la liberté.Cependant pytorch est orienté recherche et n'offre pas les mêmes compris de mise en production de TensorFlow qui est adapté pour ce contexte.

Concernant le déploiement, la solution adoptée est NVIDIA Triton Inference Server, Bien évidement des solutions comme TensorFlow serving (robuste et adapté modèle TensorFlow) et TorchServe (adapté modèle Pytorch) existent. Néanmoins TensorFlow serving ne prend pas en charge des modèles xgboost tandis que TorchServe est orienté pytorch. NVIDIA Triton Inference Server est un bon compromis pour le déploiement fiable, supportant le chargement dynamique pour réduire la charge système, et prenant en compte des modèles XGBoost avec la couche FIL.

## Conclusion

Ce chapitre a posé le cadre et les repères de travail : nous avons rappelé le contexte et les besoins auxquels le projet répond, précisé la réalité de l'organisme d'accueil, défini le périmètre opérationnel et explicité la démarche retenue ainsi que les choix techniques. L'ensemble fournit une base cohérente pour la suite des travaux, tant pour l'implémentation que pour l'évaluation, en gardant le cap sur les objectifs concrets et les contraintes de réalisation.



**FIGURE 1.12 – Frise chronologique des réalisations (Semaines 1-24).**

---

---

# CHAPITRE 2

---

## ÉTAT DE L'ART

### Sommaire

---

2.1	Introduction . . . . .	<b>18</b>
2.2	Modèles économétriques . . . . .	<b>18</b>
2.2.1	Modèles économétriques : ARIMA . . . . .	18
2.2.2	ARCH/GARCH et extensions . . . . .	19
2.3	Modèles d'apprentissage (ML) et profonds (DL) . . . . .	<b>21</b>
2.3.1	Apprentissage automatique (RF, Gradient Boosting) . . . . .	21
2.3.2	Random Forest (RF) . . . . .	21
2.3.3	XGBoost (XGB) . . . . .	21
2.4	Apprentissage profond . . . . .	<b>22</b>
2.4.1	Architecture du LSTM . . . . .	22
2.4.2	Architecture du GRU . . . . .	23
2.5	Conclusion . . . . .	<b>23</b>

---

## 2.1 Introduction

Ce chapitre dresse un état de l'art ciblé afin de situer précisément le projet dans la littérature existante et d'éclairer les choix méthodologiques et techniques retenus. L'objectif n'est pas d'être exhaustif, mais d'aller à l'essentiel : les notions, approches et résultats directement mobilisables pour la suite du travail.

La suite est structurée en sections complémentaires : section 2.2, section 2.3 et section 2.5, qui constituent le fil conducteur de ce chapitre.

## 2.2 Modèles économétriques

### 2.2.1 Modèles économétriques : ARIMA

L'idée des modèles ARIMA (*AutoRegressive Integrated Moving Average*) est d'expliquer une série temporelle par (i) ses propres retards (composante autorégressive), (ii) une combinaison linéaire de chocs passés (composante moyenne mobile), et (iii) une ou plusieurs différenciations destinées à rendre le processus stationnaire (composante intégrée).

On note  $(Y_t)$  la série d'intérêt,  $(e_t)$  un bruit blanc (i.i.d. de moyenne nulle et variance  $\sigma_e^2$ ),  $B$  l'opérateur de retard (*backshift*),  $BY_t = Y_{t-1}$ , et  $\nabla = 1 - B$  l'opérateur de différence. On introduit les polynômes

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

**Composantes AR et MA.** Un processus autorégressif d'ordre  $p$ , AR( $p$ ), vérifie

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t \iff \phi(B)Y_t = e_t.$$

La *stationnarité* d'un AR( $p$ ) exige que toutes les racines de l'équation caractéristique  $\phi(z) = 0$  soient à l'*extérieur du cercle unité* ( $|z| > 1$ ).

Un processus moyenne mobile d'ordre  $q$ , MA( $q$ ), vérifie

$$Y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} \iff Y_t = \theta(B)e_t,$$

et l'*invertibilité* requiert les racines de  $\theta(z) = 0$  à l'*extérieur du cercle unité*.

**ARMA et ARIMA.** Un processus ARMA( $p, q$ ) combine les deux composantes :

$$\phi(B)Y_t = \theta(B)e_t.$$

Si la série n'est pas stationnaire en niveau, on applique  $d$  différenciations : un ARIMA( $p, d, q$ ) est défini par

$$\phi(B)\nabla^d Y_t = \theta(B)e_t, \quad \text{où } \nabla^d = (1 - B)^d.$$

La stationnarité est requise pour la série différenciée  $\nabla^d Y_t$ , et l'invertibilité pour la partie MA.

## 2.2.2 ARCH/GARCH et extensions

Les modèles ARCH/GARCH introduisent une variance conditionnelle dynamiquement déterminée, capturant l'*agrégation de la volatilité* et le *volatility clustering*.

Soit  $r_t$  un rendement (éventuellement centré par une moyenne conditionnelle  $\mu_t$ ). On écrit

$$r_t = \mu_t + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t,$$

où  $(z_t)$  est une suite i.i.d. de moyenne nulle et variance unitaire (souvent normale ou Student- $t$ ), et  $\sigma_t^2 = \text{Var}(\varepsilon_t | \mathcal{F}_{t-1})$  la variance conditionnelle. Les diagnostics portent sur les résidus standardisés et des tests de normalité. Des variantes comme EGARCH (asymétrie logarithmique) et GJR-GARCH (effet de levier) améliorent la modélisation des chocs négatifs. Points de vigilance : sensibilité au choix de la loi d'innovation, instabilité des paramètres en cas de ruptures, et performance parfois inférieure à des méthodes ML/DL lorsque de nombreuses covariables exogènes sont disponibles.

**Modèle GARCH( $p, q$ ).** Le modèle *Generalized ARCH* introduit des retards de variance :

$$\sigma_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2,$$

avec contraintes usuelles  $\omega > 0$ ,  $\alpha_i \geq 0$ ,  $\beta_j \geq 0$ . Une condition suffisante de stationnarité faible est

$$\sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j < 1,$$

qui donne la variance inconditionnelle

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^q \alpha_i - \sum_{j=1}^p \beta_j}.$$

Le cas usuel GARCH(1,1) s'écrit  $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ .

Les paramètres se déduisent par maximum de vraisemblance (ou quasi-vraisemblance, QML) en posant une loi pour  $z_t$ . Les modèles se comparent via des critères d'information (AIC, BIC). En pratique on initialise  $\sigma_1^2$  par la variance stationnaire (si applicable) ou la variance empirique, puis on calcule récursivement  $\sigma_t^2$ .

**IGARCH.** Quand  $\alpha + \beta = 1$  dans un GARCH(1,1), on obtient un *Integrated GARCH* (IGARCH) non stationnaire au sens faible. Une forme classique (avec  $\omega = 0$ ) revient à une moyenne mobile exponentielle des carrés des résidus :

$$\sigma_t^2 = (1 - \beta) \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \quad 0 < \beta < 1.$$

**Effet de levier (asymétrie).** Pour capturer des réponses asymétriques de la volatilité aux chocs positifs/négatifs :

- **GJR-GARCH** GLOSTEN, JAGANNATHAN et RUNKLE 1993 :

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \gamma \mathbf{1}_{\{\varepsilon_{t-1} < 0\}} \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2.$$

- **TGARCH** (*Threshold GARCH*) ZAKOIAN 1994 :

$$\sigma_t = \omega + \sum_{i=1}^q (\alpha_i + \gamma_i \mathbf{1}_{\{\varepsilon_{t-i} < 0\}}) |\varepsilon_{t-i}| + \sum_{j=1}^p \beta_j \sigma_{t-j}.$$

- **EGARCH**<sup>1</sup> (*Exponential GARCH*) NELSON 1991 :

$$\varepsilon_t = \sigma_t z_t, \quad z_t \text{ i.i.d., } \mathbb{E}[z_t] = 0, \mathbb{V}[z_t] = 1,$$

$$\ln \sigma_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \ln \sigma_{t-j}^2.$$

**APARCH.** Le modèle **APARCH**<sup>2</sup> (*Asymmetric Power ARCH*) DING, GRANGER et ENGLE 1993 applique une transformation de puissance sur l'écart-type conditionnel et les résidus :

$$\sigma_t^\delta = \omega + \sum_{i=1}^q \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^\delta + \sum_{j=1}^p \beta_j \sigma_{t-j}^\delta, \quad \delta > 0.$$

Il englobe plusieurs spécifications comme cas particuliers (dans la paramétrisation de DING, GRANGER et ENGLE 1993) :

- **TS-GARCH** (Taylor–Schwert) lorsque  $\delta = 1$  et  $\gamma_i = 0$  ;
- **GJR-GARCH** lorsque  $\delta = 2$  ;
- **T-ARCH / TGARCH** de Zakoïan lorsque  $\delta = 1$  ;
- **Log-ARCH** lorsque  $\delta = 0$  (cas limite) ;
- **N-ARCH** de Higgins et Bera lorsque  $\delta = 0$  et  $\beta_i = 0$ .

**GARCH-X (variance avec régresseurs exogènes).** Dans de nombreux cas, on enrichit l'équation de variance par des variables exogènes  $x_{t-1}$  (volatilité réalisée, volume, mesures de liquidité, facteurs macro) :

$$\sigma_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 + \delta^\top x_{t-1},$$

où  $\delta$  est un vecteur de sensibilités. Cette extension, souvent appelée *GARCH-X*, permet d'intégrer des informations de marché supplémentaires tout en conservant la structure récursive.

---

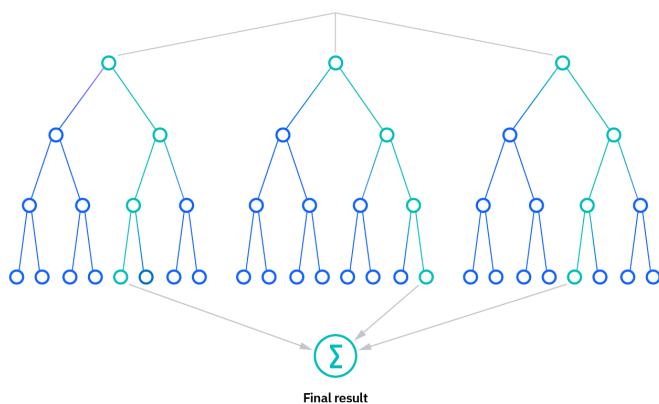
1. Nous utilisons ici la paramétrisation log-variance couramment appelée EGARCH (ou log-GARCH).  
 2. Le modèle APARCH est comme une généralisation qui classe plusieurs modèles

## 2.3 Modèles d'apprentissage (ML) et profonds (DL)

### 2.3.1 Apprentissage automatique (RF, Gradient Boosting)

Avant l'ère des transformers “Attention is All you need”, les algorithmes à arbres était fortement sollicité. Car ils sont capable de performer sur des données tabulaire avec moins de restriction sur la qualité et la quantité des données, à cela s'ajoute qu'ils sont “IA explicable”<sup>3</sup>. Même aujourd’hui ils sont des alternatives fiables lorsque le coût computationnel est en jeu. Des articles tels que GHOSH, NEUFELD et SAHOO 2022 présentent de bon résultat sur le marché boursier.

### 2.3.2 Random Forest (RF)



**FIGURE 2.1 – Schéma Random Forest - Source :** ibm.com

**Fonctionnement** Une forêt est un ensemble d’arbre qui sont entraîné en bagging. Autrement dit pour chaque arbre on tire aléatoirement un jeu de donnée et on entraîne un sous ensemble de caractéristique sur chaque noeud d’arbre<sup>4</sup>, ainsi l’agrégation moyenne les erreurs et réduit la variance. S’il s’agit d’une régression on fait la moyenne des prédictions. S’il s’agit plutôt d’une classification la majorité des votes détermine le vote finale.

**Paramètres usuels.** `n_estimators` (nb d’arbres), `max_depth` (profondeur), `min_samples_leaf`, `max_features` (= $m_{try}$ ), `bootstrap` (oui/non).

### 2.3.3 XGBoost (XGB)

**Fonctionnement** XGBoost pour “Extrême Gradient Boosting” XGBOOST DEVELOPERS s. d. est un algorithme à arbre qui vient rajouter une “couche”<sup>5</sup> sur l’algorithme random forest .Concrètement lorsqu’une première prévision est faite par un arbre alors un nouvelle arbre est

- 3. les règles de régression ou de classification peuvent directement être lu et documenté
- 4. pour ajouter de la non colinéarité
- 5. pas réellement une couche mais une amélioration



**FIGURE 2.2 – Logo Documentation officielle XGBoost - Source :** [xgboost.readthedocs.io](https://xgboost.readthedocs.io)

construit sur cette erreur résiduel et ainsi de suite. Nativement l'algorithme gère les valeurs manquantes et le early stopping stabilise l'entraînement.

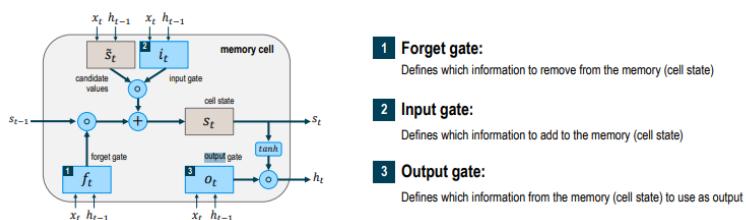
**Paramètres usuels.** n\_estimators (itérations), learning\_rate  $\eta$ , max\_depth (La profondeur), subsample (lignes), colsample\_bytree/by\_level (colonnes), régularisation lambda (L2), alpha (L1), min\_child\_weight, gamma (pénalisation de split).

## 2.4 Apprentissage profond

### 2.4.1 Architecture du LSTM

Le LSTM a toujours été connu pour sa capacité à résoudre le problème de mémoire à long terme<sup>6</sup>. Aujourd'hui il performe dans des applications combiné à des autoencoder (BAO, YUE et RAO 2017) et des mécanismes d'attention (QIU, WANG et ZHOU 2020). Un nœud LSTM contient une cellule d'état  $c_t$  et trois portes : forget ( $f_t$ ), input ( $i_t$ ), output ( $o_t$ ). Les équations suivantes décrivent son fonctionnement :

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$



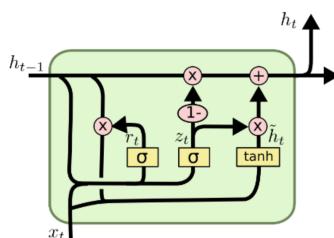
**FIGURE 2.3 – Schéma LSTM - Source :** FISCHER et KRAUSS 2017

6. le fameux vanishing gradient causé par la composé de plusieurs fonction tangente

### 2.4.2 Architecture du GRU

Le GRU, introduit par Cho et al. (2014) a montré de bonne performance dans des travaux de prévision de séries financière notamment DUTTA, KUMAR et BASU 2020<sup>7</sup>(où il performe mieux que le LSTM avec moins de coût computationnel). Il simplifie l'architecture LSTM avec deux portes : update ( $z_t$ ) et reset ( $r_t$ ). Les équations sont :

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \hat{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \end{aligned}$$



**FIGURE 2.4 – Schéma GRU - Source : NG, KNITTEL et UHLER 2020**

## 2.5 Conclusion

En définitive, ces modèles sont de parfait candidats pour la tache à réaliser. Les modèles LSTM et GRU modélise bien la dépendance à long termes, les algorithmes à arbres sont parfait pour tous ce qui est tabulaire et dépendance non linéaire. les modèles GARCH quant à eux sont parfait pour modéliser une volatilité évoluant au cour du temps. De plus, dans la mesure où on a besoin de modèles parcimonieux, interprétables et facilement disponibles en production sur serveur à basse capacité, ils s'inscrivent comme de parfaits candidats.

7. A Gated Recurrent Unit Approach to Bitcoin Price Prediction

---

---

# CHAPITRE 3

---

## COMPRÉHENSION ET PRÉPARATION DES DONNÉES

### Sommaire

---

3.1	Introduction . . . . .	25
3.2	Analyse de données . . . . .	25
3.2.1	Analyse univariée . . . . .	25
3.2.2	Analyse bivariée . . . . .	34
3.3	Sources des données et architecture de collecte . . . . .	38
3.3.1	Organisation et provenance. . . . .	38
3.3.2	Distinction recherche / déploiement. . . . .	39
3.3.3	Préparation et alignement. . . . .	39
3.3.4	Jeux de sortie. . . . .	39
3.4	Jeu de données pour le déploiement : collecte et préparation . . . . .	40
3.4.1	Actualités et sentiment (CryptoPanic → FinBERT). . . . .	40
3.4.2	Prix et volumes (CoinGecko). . . . .	40
3.4.3	Alignement et fusion. . . . .	40
3.5	Conclusion . . . . .	41

---

## 3.1 Introduction

Ce chapitre présente les données utilisées et les caractéristiques pertinentes pour l’analyse. L’ambition est de poser un socle clair : provenance des jeux de données, règles de construction, variables retenues et hypothèses opérationnelles qui guideront la suite du travail.

La progression suit un ordre pratique : une analyse à fin de comprendre les éléments influents dans la variations des prix, une présentation des sources de données et du processus de collecte .Enfin Nous présenterons la construction du jeu de donnée pour l’entraînement à des fins de déploiement.

## 3.2 Analyse de données

### 3.2.1 Analyse univariée

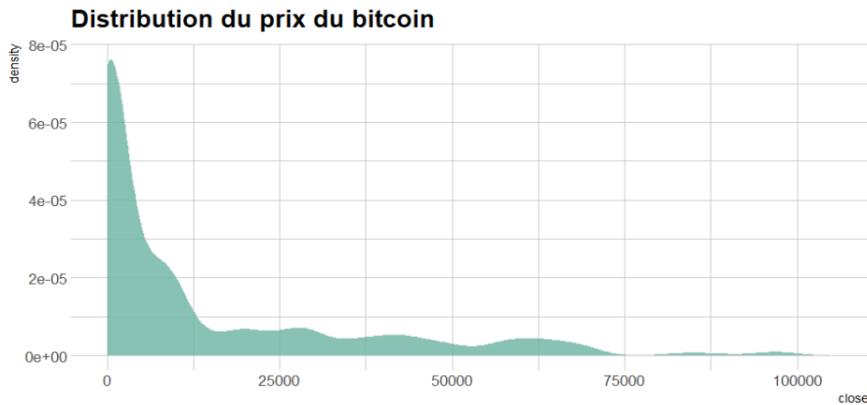
**Importation des données** Pour cette analyse, nous utilisons un jeu de données issu de Kaggle dans l’emplacement ci présent Data/Kaggle/. Ce jeu de données<sup>1</sup> contient les informations sur les prix des cryptomonnaies les plus connus sur le marché. Le présent DataFrame contient des données allant du 17 juillet 2010 au 6 avril 2025. La fréquence de capture est journalière et concerne les informations telles que le “ticker”, qui est une abréviation de la cryptomonnaie, “date” La date de capture des informations, “open” Le prix d’ouverture, “close” Le prix de fermeture, “high ”La plus haute valeur prise sur la cryptomonnaie lors de la journée, et enfin “low”, le prix le plus bas.Avant de débuter l’analyse, le jeu de données subit une étape cruciale, notamment une conversion de la colonne Date. Cette étape est jugée utile car les données de la colonne sont initialement considérées comme des chaînes de caractère.

**Densité** Nous débutons cette analyse par la densité. Elle nous donnera l’occasion de comprendre la répartition des prix. Et pour ce faire, nous utilisons le prix de fermeture. Car l’expérience démontre que le prix de clôture est fortement corrélé de manière linéaire avec les prix journaliers le plus haut et le plus bas. De plus, le marché de la cryptomonnaie a une particularité. Il est disponible H24. Ceci implique que le prix de fermeture est égal au prix d’ouverture du lendemain.

Ci présent figure 3.1 le graphique de la densité du prix de clôture du Bitcoin durant ces 15 dernières années. Il présente une forme asymétrique droite positif avec un pic assez marqué compris entre 0 et 1000\$, signe d’une fonteuse concentration des valeurs dans cet intervalle. La queue de distribution paraît assez longue et épaisse. Grâce à cette première observation, on peut remarquer que la queue de distribution présente une structure assez irrégulière. Une partie épaisse entre 20 000 et 75 000\$ et une partie très fine entre 75 000 et 100 000\$. Ce comportement nous indique donc qu’il n’est pas rare de voir le prix du Bitcoin à plus de 25 000\$, mais qu’il est de l’ordre du miracle de le voir passer la barre des 100 000\$.

---

1. Data/Kaggle/Set1/decompressed/BTC.csv



**FIGURE 3.1 – Distribution du prix du bitcoin de 2010 à 2025**



**FIGURE 3.2 – Évolution du prix du Bitcoin de 2010 à 2025**

**Décomposition de la série temporelle** Une première tentative figure 3.2, notamment sur l’outil RStudio, est d’utiliser la commande **decompose** afin de décomposer la série temporelle. Bien évidemment, cette commande échoue car elle considère que la série temporelle du Bitcoin ne possède pas plusieurs périodes, notamment de composantes répétitives<sup>2</sup>, ce qui atteste du degré d’incertitude lié à cet actif. Pour pallier à ce problème nous utilisons une alternative qu’est la moyenne mobile.

Pour afficher la tendance, nous réalisons une moyenne mobile sur un intervalle de 252 jours, soit le nombre de jours ouvrables du marché boursier traditionnel. Bien évidemment, 365 jours<sup>3</sup> auraient pu être utilisés, étant donné que ce paramètre n’affecte pas en grande partie notre analyse de tendance. Ce nombre est retenu. Car seule la tendance nous intéresse. Le graphique figure 3.3 présente l’évolution du Bitcoin entre 2010 et 2025. La coupe en jaune représente la tendance générale à l’aide d’une moyenne mobile. On observe une croissance globale marquée par deux périodes de hausse significative formant deux collines distinctes. La première colline de moyenne ampleur apparaît entre 2017 et 2018. Le second, plus important, s’étend de 2021

2. la composante de saisonnalité

3. Le marché de la cryptomonnaie est ouvert H24



**FIGURE 3.3 – Tendances de la série temporelle du Bitcoin avec la moyenne mobile**

à 2023. Selon une fouille approfondie, ces périodes de fortes croissances correspondent à des événements clés. La première hausse est liée à la levée de fonds en cryptomonnaie, encore appelée **ICO BOOM**, mobilisant des centaines de milliards de dollars pour le financement d'applications décentralisées. La seconde est attribuée à l'adoption du Bitcoin comme monnaie légale par certains États, notamment **le Salvador et la Centrafrique**. Ces deux faits marquants illustrent parfaitement le rôle déterminant des événements économiques et réglementaires dans la dynamique des prix du Bitcoin.

**Changement de structure** Ayant analyser la tendance, nous passons à la détection des changement de ruptures, cette analyse est cruciale, car les ruptures peuvent indiquer des événements ou facteurs externes qui impactent les prix du Bitcoin. On distingue plusieurs types de méthode de ruptures et plusieurs façon de les catégoriser<sup>4</sup> TRUONG, OUDRE et VAYATIS 2020 .La volatilité et la moyenne sont deux éléments prisés par les acteurs du marché .La volatilité mesure l'ampleur des variations de prix sur un intervalle de temps, et la moyenne indique le comportement moyen. Ayant connaissance de cette information, nous nous concentrerons sur ces points.

La première, la rupture de niveau<sup>5</sup>, ensuite, la rupture de variance, (lorsque la volatilité de la série change). Pour pouvoir effectuer notre test de changement de structure, diverses méthodes s'offre à nous. Les méthodes statistiques, telles que le test de Shaw, qui va vérifier si les coefficients de deux séries linéaires sont égaux, et les méthodes basées sur les changements de points, qui vont identifier où une fonction de coût change de manière significative.Nous nous concentrerons sur la dernière méthode (la recherche par point offre une dimension historique).

**Changement de structure par la moyenne (Rupture de niveaux)** Il existe plusieurs méthodes de changement de structure par la moyenne. La première est la méthode par **segmentation**

4. Par fonction de coût, méthode de recherche et par contrainte

5. Un changement brusque dans la moyenne

**binaire**<sup>6</sup>, où on cherche le meilleur point de rupture sur toute la série, ensuite, on coupe la série en deux, puis on recommence. Elle est simple, mais ne garantit pas une solution optimale, et peut rater des fois des ruptures subtiles, si elles sont masquées par des ruptures plus fortes. La méthode du **voisin segmenté**<sup>7</sup>, elle, est une méthode qui cherche toutes les segmentations possibles pour un nombre donné de changements et choisit la segmentation qui minimise un critère de coût global, comme l'erreur quadratique ou encore la log vraisemblance négative. La troisième méthode est une méthode d'élagage, encore appelée la méthode *PELT*, qui est une méthode exacte et optimisée par programmation dynamique, qui évite aussi de faire des calculs inutiles. Elle a l'avantage de trouver une solution optimale globale, avec une complexité plutôt linéaire en moyenne. Grâce à elle, pas besoin de connaître le nombre de ruptures à l'avance. Cependant, le choix du paramètre de pénalisation peut influencer les résultats. Pour notre analyse, les deux méthodes par **voisins segmentés** et **PELT** sont choisies. Le voisin segmenté est effectué en prenant en compte une pénalité de type *BIC*<sup>8</sup>, qui se trouve aux alentours de 17, 16 pour un nombre maximal de 8 points de rupture, afin de ne capturer que les grands changements. Le modèle a détecté 7 points. Ce même critère d'information Bayesien est utilisé avec la méthode *PELT*, qui elle a détecté plus de 3565 points de rupture. Ce résultat paraît assez surprenant, cependant, si l'on regarde la nature du marché très changeante, ce résultat peut être réaliste des tendances du marché.

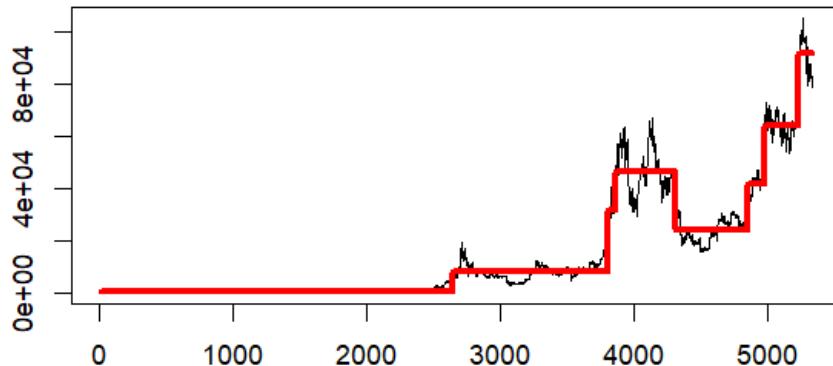


FIGURE 3.4 – Test de changement de structure par la méthode du voisin segmenté(Moyenne)

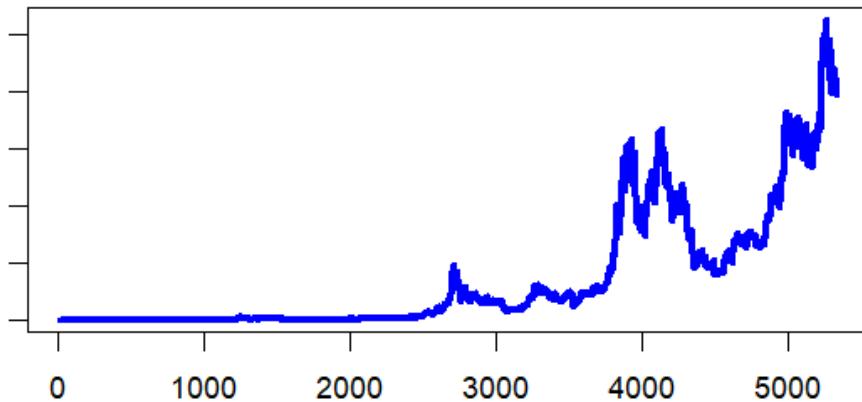
Curieux sur les causes de ces changements, nous effectuons une recherche ordonnée sur Google, avec le template suivant. La “*date Bitcoin News*”.

**Changement de structure par la variance** Les deux dernières méthodes sont toujours réutilisées dans le cadre de la variance. Pour la méthode du voisin segmenté, pour un maximum de 10 segments, la méthode a détecté 9 segments. Bien sûr, ce nombre est arbitraire. Et plus nous augmentons ce nombre, plus les segments augmentent. Ceci montre à quel point la méthode du voisin segmenté pour la rupture de variance n'est pas très adaptée en général. Mais ce que l'on

6. qui est une méthode assez gourmande et récursive

7. Elle a l'avantage de donner une solution optimale pour un nombre fixé de ruptures. a un coût computationnel élevé de l'ordre de  $n^2 * k$ , où  $n$  est la taille de la série et  $k$  est le nombre de ruptures. Elle nécessite de connaître le nombre de changements à détecter

8. le critère d'information Bayesien



**FIGURE 3.5 – Test de changement de structure par la méthode de PELT(Moyenne)**

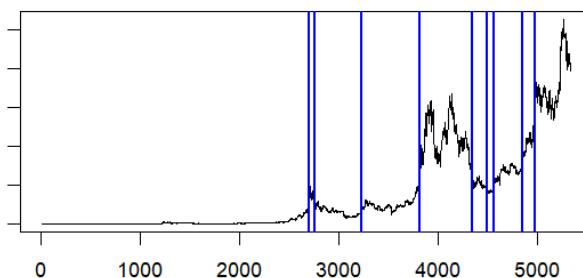
**TABLE 3.1 – Ruptures (moyenne) détectées par Voisin segmenté**

Date	Articles liés
2017-10-07	Le Bitcoin dans la tourmente : la fraude est de plus en plus évidente
2017-11-30	Bitcoin recovers from sudden selloff as large swings persist
2018-01-29	After the Biggest Cryptocurrency Hack Ever, Bitcoin, Ethereum and Ripple Are All — Why ?
2020-12-25	Bitcoin Tops \$24.6K on Christmas Day, Sets New All-Time High
2020-12-18	Bitcoin highs, retail lows and 48 creatures saved from extinction
2021-05-14	Bitcoin passes the ransom test
2021-05-18	China bans financial, payment institutions from cryptocurrency business
2022-05-05	Bitcoin drops by most in almost a month as Fed optimism fades
2023-10-23	Bitcoin Hits \$35,000 for First Time Since 2022 on ETF Optimism
2024-11-09	Bitcoin Wraps Up A Big Week After Trump Victory Spurs Fresh High
2025-02-25	Bitcoin falls below \$90,000 for first time in a month, ether tumbles

Paramètres : penalty=“BIC” (17.16), method=“SegNeigh” (7 points détectés).

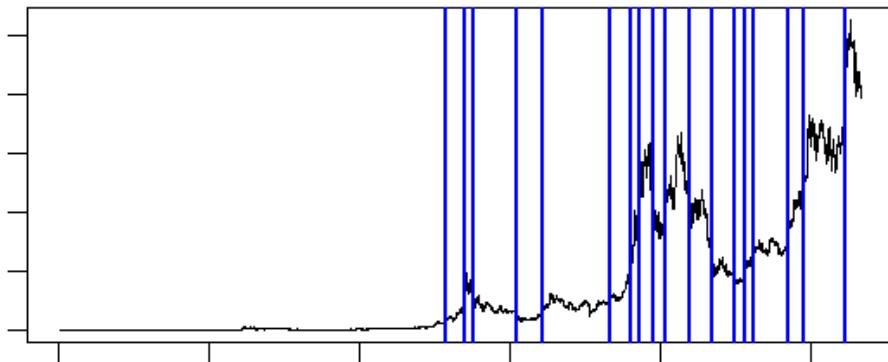
Remarque : 11 articles au alentours de ces dates trouvés.

peut observer est que pour un besoin d’analyse graphique, les faits qui nous intéressent le plus sont les points historiques. Donc, aller au-delà de 10 rendrait l’analyse lourde.



**FIGURE 3.6 – Test de changement de structure par la méthode du voisin segmenté (Variance)**

En faisant une synthèse, les différents chocs ou changements de structure sont fortement influencés soit par des événements endogènes tels que des adoptions massives, des changements technologiques, ou encore des événements exogènes tels que des crises financières, des régulations, ou encore des annonces réglementaires.



**FIGURE 3.7 – Test de changement de structure par la méthode PELT (Variance)**

**TABLE 3.2 – Ruptures (variance) détectées par Voisin segmenté (SegNeigh)**

Date de rupture	Articles liés
2017-11-30	Bitcoin plunges 20 percent from its high
2018-01-29	2018 might be the year of the great crypto rotation (Tom Lee)
2020-12-25	Bitcoin tops \$24.6K on Christmas Day, sets new all-time high
2019-05-18	Bitcoin Drops \$1,000 In Value Amid Market Sell-Off (CoinDesk).
2020-12-25	Bitcoin Price Hits Fresh Record Above \$24K on Christmas Eve (CoinDesk).
2022-06-12	Bitcoin Drops to One-Month Low as US CPI Report Hurts Sentiment (Bloomberg).
2022-11-08	Crypto exchange FTX halts client withdrawals amid liquidity crunch (Reuters).
2023-01-11	El Salvador passes digital-asset issuance law as it doubles down on bitcoin (Reuters).
2023-10-23	Bitcoin hits highest in nearly 18 months on ETF hopes (Reuters).
2024-02-25	MicroStrategy's X account hacked, leads to \$440K crypto being stolen (CoinDesk).
2025-04-06	Bitcoin last down 5% at \$78,892.92 (Reuters).

Paramètres : penalty=“BIC”(16.2), method=“SegNeigh”(10 points détectés), Q=10.

Remarque : 11 articles au alentours de ces dates trouvés.

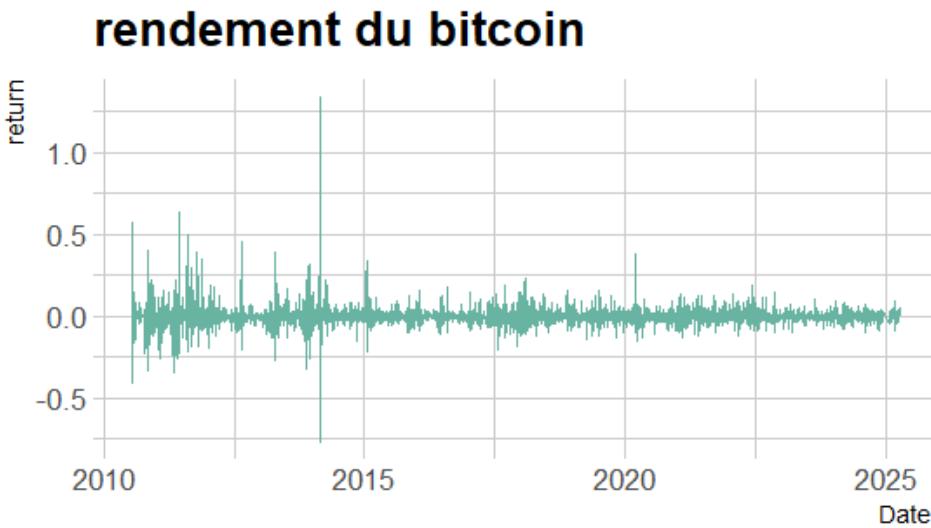
**Analyse des rendements** Les rendements offrent une perspective complémentaire à l'analyse des prix : au lieu d'observer l'évolution d'un cours, on mesure les *variations* période par période.

**Rendements simples** Soit  $P_t$  le prix (ou la clôture) à la date  $t$ . Le *rendement simple* s'écrit

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1. \quad (3.1)$$

Sur  $h$  périodes consécutives, le rendement simple cumulé vaut

$$R_{t \rightarrow t+h} = \prod_{s=t+1}^{t+h} (1 + r_s) - 1. \quad (3.2)$$



**FIGURE 3.8** – Rendements  $r_t$  du Bitcoin de 2010 à 2025.

**Log-rendements** Une autre écriture, très utilisée, consiste à prendre le logarithme népérien de la variation relative :

$$\ell_t = \ln P_t - \ln P_{t-1} = \ln(1 + r_t). \quad (3.3)$$

Les *log-rendements* possèdent des propriétés mathématiques appréciées : **additivité** dans le temps,

$$\sum_{s=t+1}^{t+h} \ell_s = \ln P_{t+h} - \ln P_t = \ln \left( \prod_{s=t+1}^{t+h} (1 + r_s) \right), \quad (3.4)$$

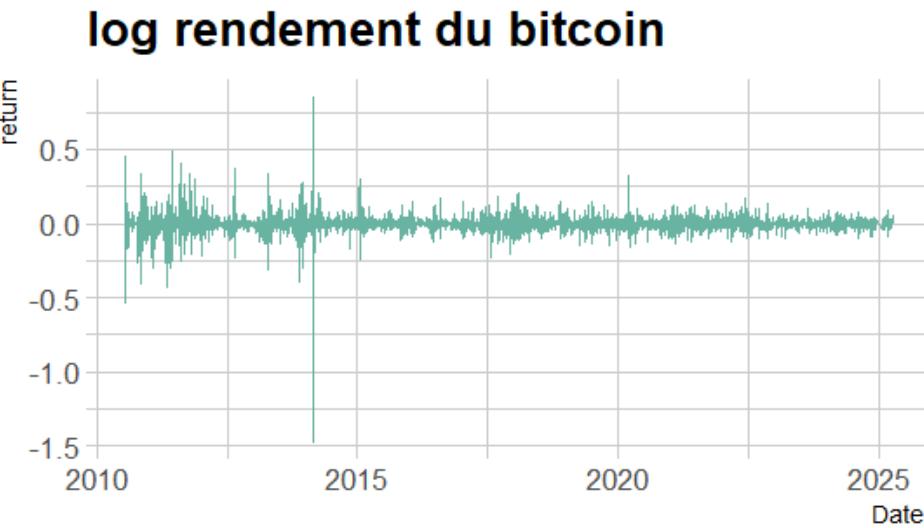
ce qui simplifie les cumuls, **symétrie** relative entre hausses et baisses de même amplitude ; pour des variations modestes, l'approximation  $\ell_t \approx r_t$  est valable. En pratique, ces propriétés facilitent la modélisation, notamment quand on postule (à tort ou à raison selon les actifs et l'horizon) une distribution proche de la normale.

**Lecture et interprétation des graphiques** À l'examen visuel des figures 3.8 et 3.9, plusieurs traits ressortent. La **volatilité est manifestement non constante** : les périodes calmes alternent avec des phases de forte agitation, où l'amplitude des variations s'accroît nettement (*volatility clustering*). Ensuite, on observe des **pics extrêmes** (hausses ou baisses soudaines) à différentes dates ;

**ACF et PACF des rendements** Soient  $\{r_t\}$  les rendements. L'autocorrélation empirique au retard  $k$  est

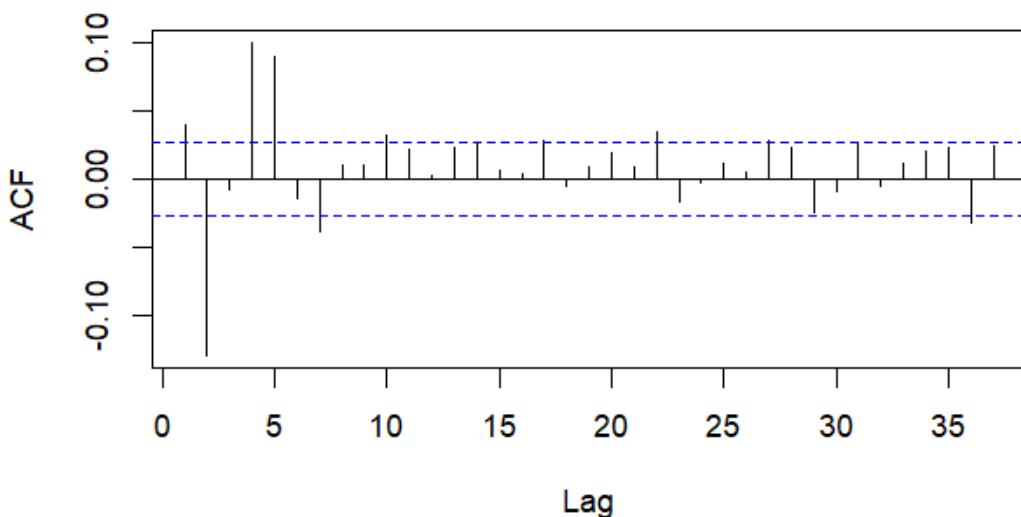
$$\hat{\rho}(k) = \frac{\sum_{t=k+1}^n (r_t - \bar{r})(r_{t-k} - \bar{r})}{\sum_{t=1}^n (r_t - \bar{r})^2}, \quad k \geq 0, \quad (3.5)$$

et la PACF correspond à l'autocorrélation *partielle* obtenue après avoir régressé  $r_t$  et  $r_{t-k}$  sur  $r_{t-1}, \dots, r_{t-k+1}$ . Dans un échantillon i.i.d., les barres de l'ACF/PACF devraient rester dans



**FIGURE 3.9** – Log-rendements du Bitcoin de 2010 à 2025  $\ell_t = \ln P_t - \ln P_{t-1}$ .

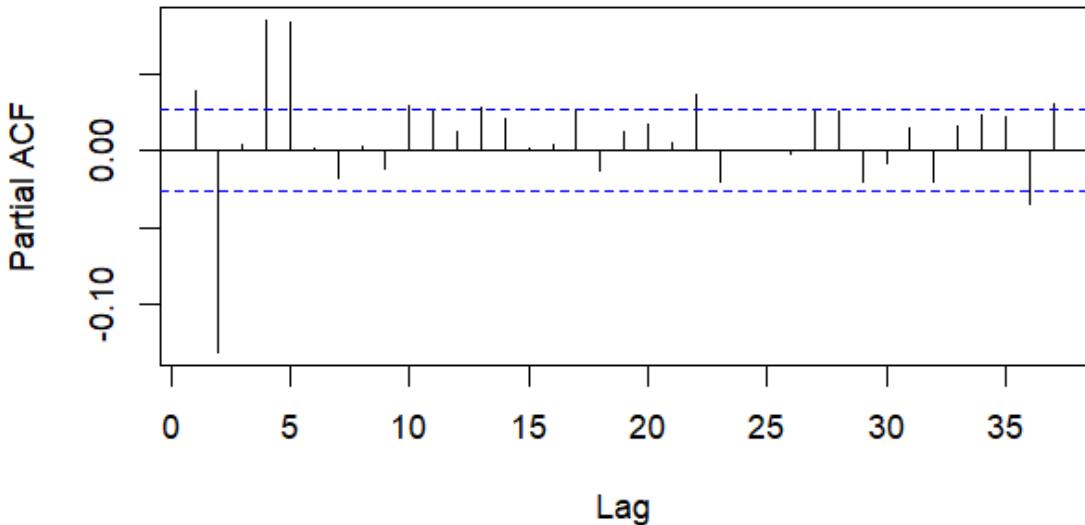
les bandes de confiance  $\pm 1.96/\sqrt{n}$ . Sur nos données, l’ACF et la PACF des rendements sont globalement faibles aux premiers retards (Figures 3.10 et 3.11), Mais l’on constate aussi qu’il y a aussi une partie de l’information qui est encore présente, c’est-à-dire une certaine dépendance dans le temps.



**FIGURE 3.10** – ACF rendements  $r_t$  Bitcoin 2010-2025.

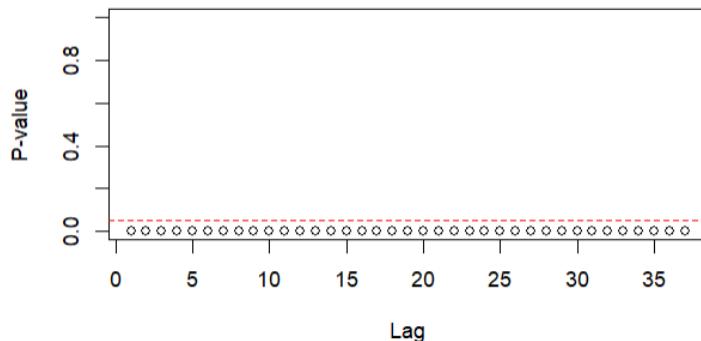
**Test de McLeod–Li (effets ARCH)** Pour détecter une variance conditionnelle (effets ARCH), on applique le portemanteau de Ljung–Box aux carrés  $r_t^2$ . La statistique

$$Q^*(m) = n(n+2) \sum_{k=1}^m \frac{\hat{\rho}_{r^2}(k)^2}{n-k} \quad (3.6)$$



**FIGURE 3.11 – PACF rendements  $r_t$ , Bitcoin 2010-2025.**

suit approximativement une  $\chi^2$  à  $m$  ddl sous  $H_0$  (absence d'ARCH). Des  $p$ -valeurs très faibles sur une plage de retards conduisent à rejeter  $H_0$ . Dans notre cas, le tracé des  $p$ -valeurs du test de McLeod–Li indique un rejet net pour quasiment tous les retards (figure 3.12), ce qui motive l'usage de modèles à variance conditionnelle.



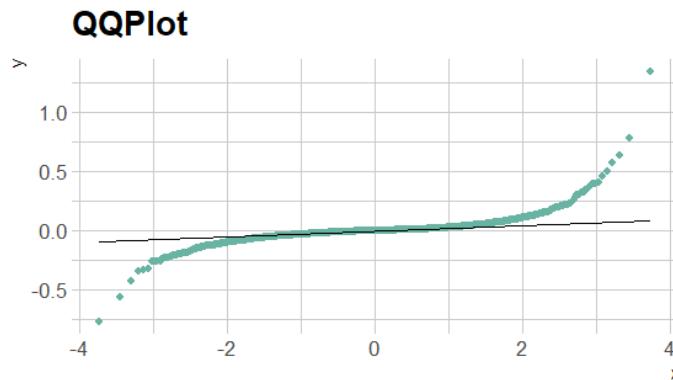
**FIGURE 3.12 – Test de McLeod-Li rendements Bitcoin 2010-2025 :  $p$ -valeurs par retard pour  $r_t^2$ .**

**Normalité : QQ-plot et test de Shapiro-Wilk** Le diagramme quantile-quantile (QQ-plot) confronte les quantiles empiriques aux quantiles théoriques de deux distribution<sup>9</sup> une droite suggère une normalité compatible, des écarts systématiques (forme en S, queues relevées) révèlent des queues épaisses et/ou de l'asymétrie. Le QQ-plot des rendements montre des écarts marqués aux extrémités (figure 3.13). Le test de Shapiro–Wilk (statistique  $W$ ) formalise ce diagnostic :

$$W = \frac{\left[ \sum_{i=1}^{\lfloor n/2 \rfloor} a_i (x_{(n-i+1)} - x_{(i)}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (3.7)$$

9. Ici on confronte les rendements à la loi normale

où  $x_{(i)}$  sont les données triées et les  $a_i$  proviennent des quantiles normaux théoriques. Sur notre échantillon,  $p$ -valeur  $< 2.2 \times 10^{-16}$  implique le *rejet* de la normalité (les rendements sont à queues épaisse), confortant l'usage de lois alternatives dans l'estimation.



**FIGURE 3.13 –** QQ-plot des rendements  $r_t$ . La courbure en S révèle des queues plus épaisse que la normale.

**Bilan.** Cette analyse univariée nous a permis d'établir plusieurs faits. Le premier, depuis sa date de création, le Bitcoin a subi une évolution conséquente, marquée par des périodes de hausses et de baisses et des changements structurels, notamment causés par des événements particuliers, tels que les annonces réglementaires, les événements technologiques, qu'ils soient liés à de la cybersécurité ou à de l'innovation et enfin des faits de société. Une analyse plus poussé au niveau des rendements, log rendements pour être plus exact, à déceler **une dépendance temporelle, une distribution non normale et une variance conditionnelle**.

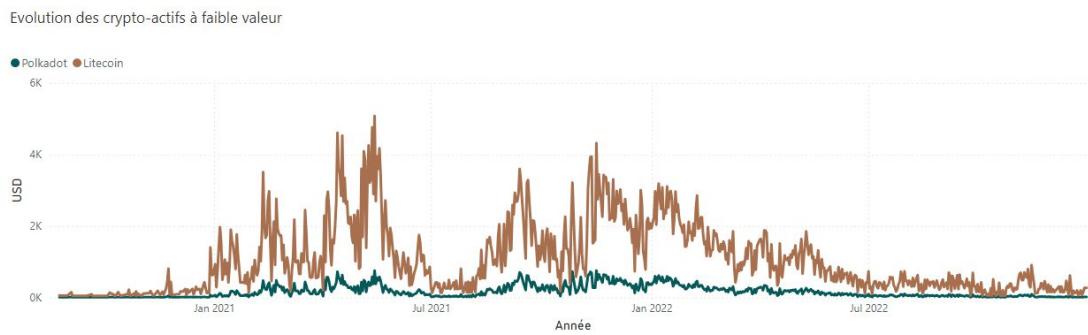
### 3.2.2 Analyse bivariée

Le monde de la cryptomonnaie est marqué par de multiples interactions. Des actifs qui, en apparence, semblent évoluer isolément parce que la demande leur est propre, entretiennent en réalité des relations étroites entre eux, mais aussi avec la popularité médiatique et, parfois, avec certains métaux précieux.

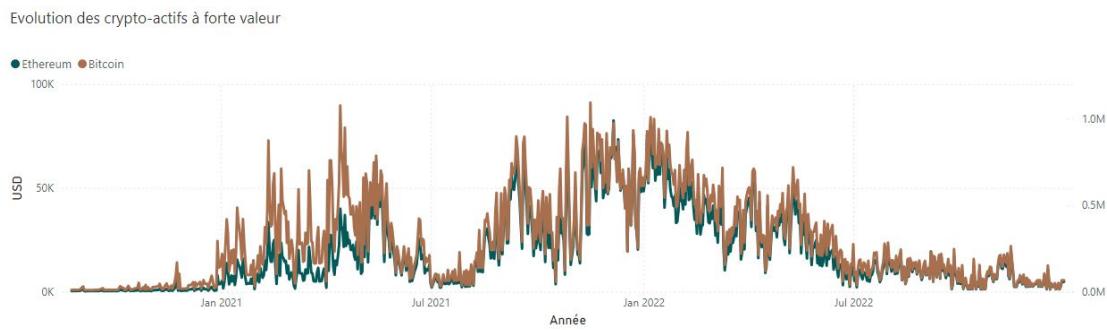
**Relations entre crypto-actifs** Pour cette analyse, nous nous référons au jeu de données Data/Kaggle/Set 5/data, qui contient des informations *minute* sur les prix de plusieurs cryptomonnaies. Après prétraitement, cinq actifs ont été retenus en raison de leur capitalisation et des irrégularités présentes dans les données (valeurs manquantes et granularité hétérogène<sup>10</sup>) : Bitcoin, Litecoin, Ethereum, Polkadot et Cardano. Les irrégularités proviennent principalement d'un manque de vérification des données collectées auprès de particuliers (Kaggle) et de différences dans les dates de mise en circulation des cryptomonnaies.

Les deux graphiques de prix figure 3.14 et figure 3.15 présentent, d'une part, l'évolution d'actifs à plus faible capitalisation (Polkadot et Litecoin) et, d'autre part, celle d'actifs à capitali-

10. les fréquence et les périodes ne sont pas souvent les mêmes



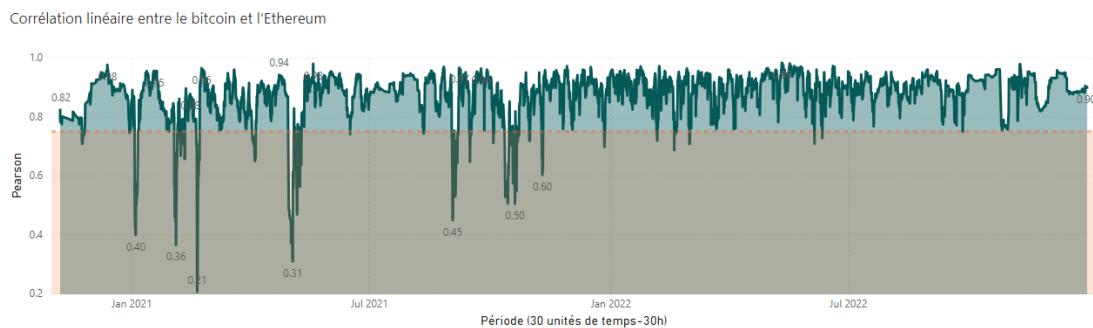
**FIGURE 3.14 – Evolution des actifs à faible capitalisation (Polkadot-Litecoin)**



**FIGURE 3.15 – Evolution des actifs à forte capitalisation Bitcoin-Ethèreum)**

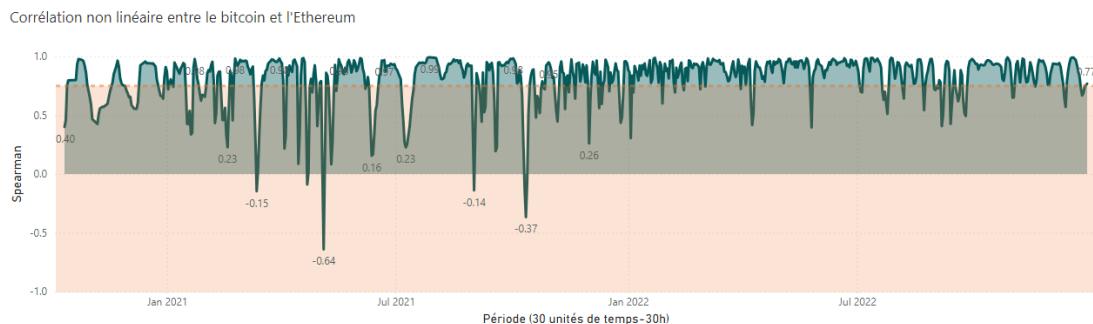
sation plus élevée (Ethereum et Bitcoin). Bien que ces crypto-actifs diffèrent par leur échelle de valeur, ils présentent un mouvement d'ensemble : deux grandes périodes de hausse (janvier 2021–juillet 2021 puis août 2021–juillet 2022), assorties d'une forte volatilité. Ces épisodes coïncident notamment avec des annonces marquantes : des prises de position des autorités américaines et la communication de Tesla autour de l'acceptation du Bitcoin comme moyen de paiement au printemps 2021, puis l'assouplissement de cette position à l'été 2021.

L'analyse qui suit porte sur deux volets complémentaires : la corrélation (linéaire et de rang) et la causalité au sens de Granger. L'objectif est d'identifier les facteurs potentiellement utiles à la prédiction du prix du Bitcoin.



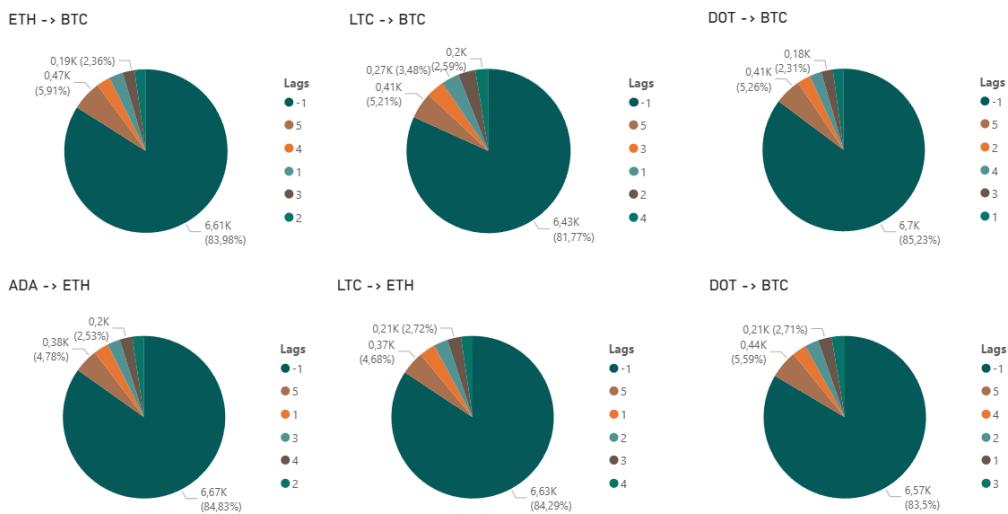
**FIGURE 3.16 – Test de corrélation linéaire entre le Bitcoin et l'Ethèreum**

**Test de corrélation (linéaire et non linéaire).** L'expérience est réalisée sur les données mentionnées ci-dessus. Les corrélations sont calculées à l'aide d'une fenêtre glissante de 30 points



**FIGURE 3.17 – Test de corrélation non linéaire entre le Bitcoin et l’Ethéreum**

(environ 30 heures, les fréquences de capture n’étant pas parfaitement homogènes selon les actifs), en se concentrant d’abord sur le couple Bitcoin–Ethereum (deux premières capitalisations). Le premier graphique met en évidence une corrélation de Pearson globalement positive sur la période étudiée ; la corrélation de Spearman (de rang) conduit à une lecture similaire. Les deux crypto-actifs semblent donc entretenir une relation étroite sur de nombreux intervalles.

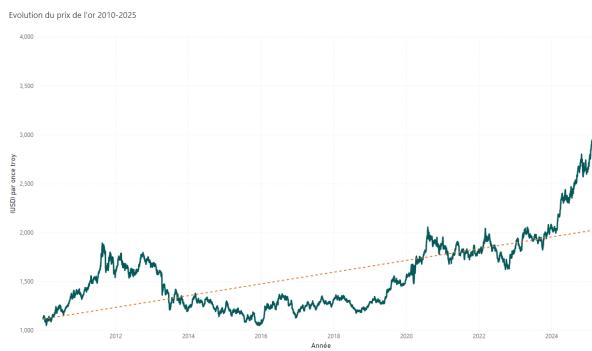


**FIGURE 3.18 – Test de causalité sur les cryptomonnaies**

**Test de causalité.** Pour approfondir, un test de causalité de Granger(DUFOUR 2002)<sup>11</sup> est appliqué sur la période d’étude, avec un nombre maximal de retards fixé à 5 (contrainte matérielle). Le test est conduit sur les *rendements* des cinq actifs . Par souci de lisibilité, la valeur –1 est utilisée lorsque la relation causale n’est pas significative entre deux actifs. Visuellement, la présence d’une causalité (au sens de Granger) n’apparaît qu’environ 15 % du temps ; on en déduit qu’en général les rendements d’une cryptomonnaie ne *causent* pas ceux des autres au sens strict du test.

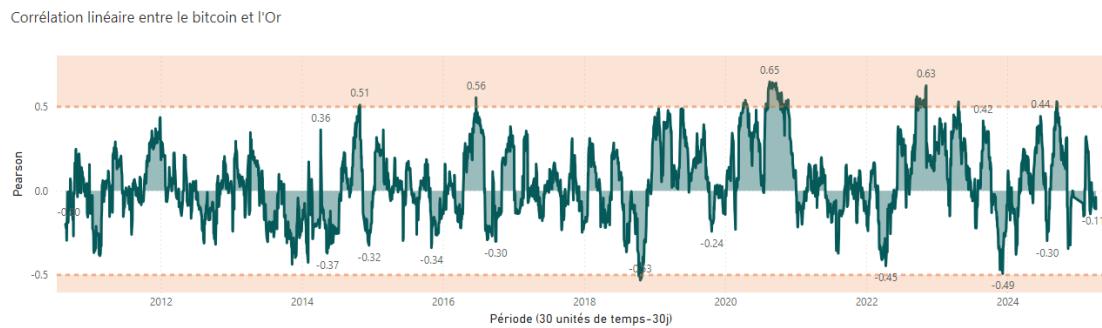
**Relations entre le Bitcoin et l’Or** L’or est traditionnellement perçu comme une valeur refuge et constitue un actif de référence pour de nombreux investisseurs souhaitant sécuriser une partie

11. Le test de Granger n'est applicable seulement si la série est stationnaire

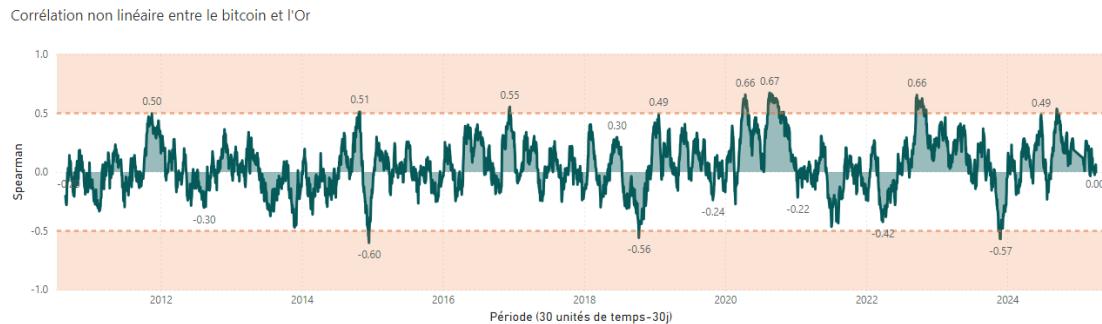


**FIGURE 3.19 – Evolution du prix de l’Or**

de leur exposition.

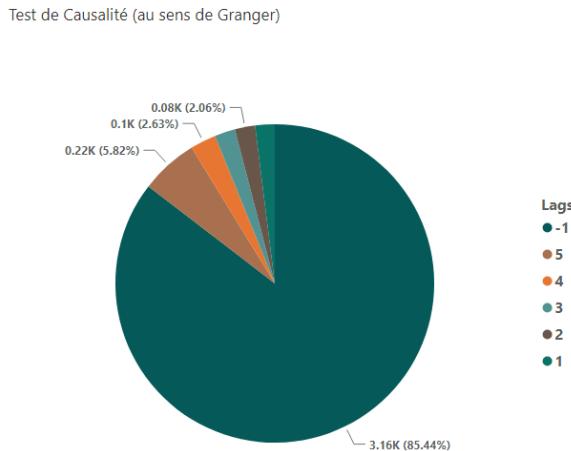


**FIGURE 3.20 – Corrélation linéaire entre le bitcoin et l’Or**



**FIGURE 3.21 – Corrélation non linéaire entre le bitcoin et l’Or**

**Analyse de corrélation (linéaire et non linéaire).** L’analyse est réalisée à partir des jeux de données Data/Investing.com/Set 1 (Investing.com) et Data/Kaggle/Set 1, sur environ quinze ans de données à fréquence journalière. Chaque observation est associée à une fenêtre glissante de 30 jours pour estimer la corrélation entre les rendements du Bitcoin et ceux de l’or. La première conclusion est l’absence de corrélation linéaire durablement forte entre ces deux actifs : on observe plutôt une alternance de phases de corrélation positive et négative. Il n’est donc pas possible de fonder une stratégie uniquement sur l’existence d’une corrélation stable entre Bitcoin et or.



**FIGURE 3.22 – Test de causalité entre l’or et le bitcoin**

**Test de causalité.** Le graphique récapitulatif du test de Granger entre le Bitcoin et l’or (fenêtre glissante de 30 jours sur quinze ans) indique qu’environ 85 % du temps, il n’existe pas de relation causale significative au sens du test. Ce pourcentage n’étant pas maximal, il subsiste néanmoins des périodes au cours desquelles les rendements de l’or peuvent apporter une information utile à la prédiction des rendements du Bitcoin.

### 3.3 Sources des données et architecture de collecte

À l’issue des analyses univariées et bivariées, nous retenons deux familles complémentaires. D’une part, des *variables endogènes* décrivant l’état du marché (prix, volumes) sur le Bitcoin et plusieurs grandes cryptomonnaies. D’autre part, des *variables exogènes* capturant le contexte d’information (sentiment des nouvelles, popularité grand public) et, lorsque pertinent, des événements macro-réglementaires ou technologiques. L’objectif est de fournir au modèle un socle factuel suffisamment riche tout en garantissant la cohérence temporelle entre séries.

#### 3.3.1 Organisation et provenance.

L’architecture est organisée par Data/<source>/Set<k>/, chaque *set* correspondant à un export stabilisé. Le socle marché du Bitcoin provient de Data/CryptoDataDownload/Set1/BTC\_for\_modelling.csv et contient notamment Close, Volume BTC et tradecount<sup>12</sup>. Les altcoins de référence (ETH, ADA, DOT, LTC) sont issus de Data/Kaggle/Set5/data/ sous la forme <SYMBOL>\_for\_modelling.csv. Le proxy de valeur refuge provient de Data/Investing.com/Set1/Gold\_for\_modelling.csv<sup>13</sup>. La popularité grand public est mesurée via Google Trends par Data/pytrends/Set1/BTC\_Popularity\_for\_modelling.csv.

**L’information issue des nouvelles (phase recherche)** est agrégée dans Data/CryptoNews/

12. Ces altcoins ont été sélectionnés sur la base de plusieurs propriétés, notamment l’irrégularité au niveau de la granularité offerte par chacun des cryptomonnaies, la période disponible, et les erreurs dans les données

13. Ici, l’or est utilisé car il est considéré comme un actif stable.

`Set1/news_for_modelling.csv` à partir de CryptoPanic : le jeu source couvre environ cinq années, et nous en extrayons une fenêtre centrée sur la période Covid-19 (fin 2019–2020). *Aucun modèle de langage n'est appliqué à ce stade* : nous utilisons uniquement les champs de votes positifs et négatifs (comptages) comme *proxy* de sentiment, agrégés par date.

### 3.3.2 Distinction recherche / déploiement.

Le pipeline ci-dessus décrit la constitution du panel pour la *recherche*. Pour le *déploiement*, un second pipeline construit un dataset distinct où l'information textuelle des nouvelles est *effectivement* passée par un modèle de sentiment spécialisé (FinBERT, via Hugging Face) afin de produire un score continu par article puis par heure. Cette différence de traitement est intentionnelle : les compteurs CryptoPanic suffisent pour explorer des relations globales, tandis que l'inférence supervisée (FinBERT) est réservée au produit final, avec des garanties d'inférence et de latence adaptées au déploiement.

### 3.3.3 Préparation et alignement.

Le traitement suit une logique homogène pour toutes les sources. Les horodatages sont convertis au format `datetime` et utilisés comme index de jointure. Sur les séries de prix, les log-rendements en pourcentage sont construits par

$$r_t = 100 \times \ln\left(\frac{P_t}{P_{t-1}}\right),$$

avec une notation explicite par actif (`return ETH`, `return ADA`, etc.) et, pour l'or, `return gold`. Les colonnes de volumes sont renommées de façon explicite (par exemple `volume gold`, `volume ETH`). Les séries exogènes sont harmonisées par renommage simple (`bitcoin` → `interest` pour Google Trends ; `datetime` → `date` pour le sentiment des nouvelles). Toutes les tables sont ensuite restreintes à la période commune définie par l'index du socle Bitcoin et fusionnées par concaténation colonne à colonne, de sorte que chaque ligne représente un même instant *cross-source*.

### 3.3.4 Jeux de sortie.

Le tableau final destiné à l'apprentissage est exporté sous `Data/Output/Data_for_modelling.csv`, qui rassemble le *panel* aligné (rendements, volumes, popularité, sentiment, etc.). Un second export `Data/Output/Data_BTC_Original_price.csv` conserve, à des fins de traçabilité, la série originale du Bitcoin avant toute sélection de variables. Cette séparation permet de réentraîner les modèles sur un panel propre, tout en gardant une référence directe à la source de vérité des prix<sup>14</sup>.

14. Les sources n'ont pas toutes la même granularité ni le même degré de contrôle en amont ; la consolidation gère ces différences par l'alignement temporel et l'utilisation d'indices (rendements, intérêts) plutôt que de niveaux bruts. Les irrégularités résiduelles (lacunes, mises en circulation décalées selon les actifs) sont prises en compte lors de la restriction à l'intersection des périodes, ce qui garantit l'absence de fuite d'information et une comparabilité stricte entre variables.

## 3.4 Jeu de données pour le déploiement : collecte et préparation

Cette section décrit le pipeline de constitution du *dataset d'entraînement* utilisé pour le *déploiement* des modèles. Contrairement au jeu *recherche* (où le signal d'actualité provenait uniquement des compteurs de votes CryptoPanic et d'actualité avec une période sélectionnée sur l'année 2019-2020), le présent pipeline extrait l'information textuelle des nouvelles et en dérive un score de sentiment supervisé, puis l'aligne à une base marché horaire multi-actifs du 5 Août 2024 au 25 Mai 2025.

### 3.4.1 Actualités et sentiment (CryptoPanic → FinBERT).

Les articles sont chargés depuis `Data/CryptoNews/Set1/cryptopanic_news.csv` (`title`, `description`, `newsDatetime`). Les champs textuels sont concaténés (`text = title + description`), horodatés en UTC et restreints à la fenêtre d'un an la plus récente<sup>15</sup>. Le score de sentiment est obtenu avec le modèle *FinBERT* (`ProsusAI/finbert`) : pour chaque texte, on calcule la distribution  $\hat{\mathbf{p}} = (p_{\text{neg}}, p_{\text{neu}}, p_{\text{pos}})$  via *softmax* et l'on retient

$$s_t = p_{\text{pos}} - p_{\text{neg}},$$

qui varie dans  $[-1, 1]$ . Les scores sont ensuite agrégés à l'**heure** (médiane par pas d'une heure après *floor* de `newsDatetime`) et réindexés sur une grille horaire complète  $[t_0, t_1]$ ; les heures sans article reçoivent la valeur 0 par construction, ce qui produit la série `market_sentiment`.

### 3.4.2 Prix et volumes (CoinGecko).

Les données marché sont récupérées via l'API CoinGecko (tête de requête authentifiée), pour un ensemble large de cryptomonnaies, sur la même fenêtre d'un an (`vs_currency = usd`, pas horaire). Chaque actif est converti en *DataFrame* indexé par l'horodatage UTC, réindexé sur la grille horaire, interpolé linéairement puis *back/forward-filled* pour combler les rares lacunes. Les colonnes sont renommées explicitement : `price_<SYMBOL>` et `volume_<SYMBOL>`. Afin d'éviter des colonnes trop lacunaires, on calcule pour chaque `price_<SYMBOL>` le pourcentage d'horodatages non nuls et l'on conserve uniquement les actifs dont la couverture est au moins de 90 % ; l'ensemble filtré constitue `df_filtered`.

### 3.4.3 Alignement et fusion.

On détermine la *période commune* entre la base marché et le sentiment agrégé ; les deux tables sont alors découpées sur  $[t_{\min}, t_{\max}]$  et fusionnées par l'index temporel horaire. Le résultat est un *panel* horaire multi-actifs où chaque ligne combine prix/volumes des cryptomonnaies retenues et le signal `market_sentiment`. Les valeurs manquantes résiduelles sont remplies à

15. La borne supérieure correspond au dernier `newsDatetime` observé ; la borne inférieure est fixée à –365 jours.

zéro uniquement pour les variables exogènes dérivées (par ex. heures sans nouvelles), afin de conserver un tenseur dense à l’entraînement.

**Sorties et traçabilité.** Les principaux artefacts produits par le pipeline sont sérialisés pour réutilisation :

- `df\news.parquet` — articles CryptoPanic enrichis (horodatage et `sentiment_score`);
- `df\prices.parquet` — base marché horaire (prix/volumes) après filtrage de couverture;
- `historical\crypto\data\_filtered.parquet` — version compacte des séries marché;
- `dataset\for\training.parquet` — panel final aligné (features du modèle).

Ce jeu *déploiement* sert d’entrée directe aux modèles opérationnels. Le choix d’une agrégation horaire, d’un filtrage de couverture et d’un score supervisé  $s_t$  vise à garantir la *robustesse* (moins de trous de données), la *cohérence temporelle* (même grille pour toutes les sources) et la *faible latence* (calcul du sentiment au fil de l’eau).

## 3.5 Conclusion

En conclusion, ce chapitre a clarifié le périmètre des données, la nature des variables utiles et les règles de construction mobilisées. Ces éléments forment une base solide pour la modélisation, l’évaluation et l’interprétation des résultats dans la suite du document.

---

---

# CHAPITRE 4

---

## MODÉLISATION

### Sommaire

---

4.1	Introduction . . . . .	43
4.2	Choix des modèles . . . . .	43
4.3	Pipeline de modélisation . . . . .	44
4.4	Modélisation . . . . .	45
4.4.1	ARIMA . . . . .	45
4.4.2	XGBoost . . . . .	56
4.4.3	GRU . . . . .	64
4.5	Evaluation des modèles . . . . .	73
4.6	Entraînement des modèles et suivi expérimental . . . . .	73
4.6.1	Analyse de la MFTR par crypto (modèles GRU) . . . . .	74
4.7	Foreward testing . . . . .	75
4.8	Conclusion . . . . .	76

---

## 4.1 Introduction

Lors du chapitre précédent nous avons établit le besoin en données. Dans ce chapitre nous présentons le procéssus de modélisation. L'objectif est d'expérimenter un cadre théorique et pratique de la prédition de la volatilité et des prix. Nous débuterons donc par définir le cadre de modélisation, appliquerons les différentes phases de celui-ci, évaluerons les modèles suivant la FinTSB, entraînerons ensuite les modèles pour le déploiement, enfin nous effectuerons une phase de forward testing pour évaluer en temps réelles nos modèles de prévision de moyenne temporelle.

## 4.2 Choix des modèles

Les analyses conduites au chapitre précédent nous ont permis de connaître les différents facteurs qui intervenaient dans la prévision des prix des cryptomonnaies. Arrivé à l'étape de modélisation, il est important de connaître les modèles adéquats qui vont prendre en entrée ces différentes données. Pour ce faire, plusieurs éléments sont à prendre en compte. Premièrement, la nature séquentielle de l'information. (D'après le résultat de l'ACF au chapitre précédent). C'est-à-dire qu'une partie de l'information est capturée entre les différents instants temporels. Ensuite, les éléments exogènes, notamment les événements sur le marché ou encore les informations sur l'état du marché, peuvent influencer la variation de ceux-ci. Et enfin, il est essentiel aussi de choisir des modèles qui sont capables d'expliquer l'ampleur des amplitudes.

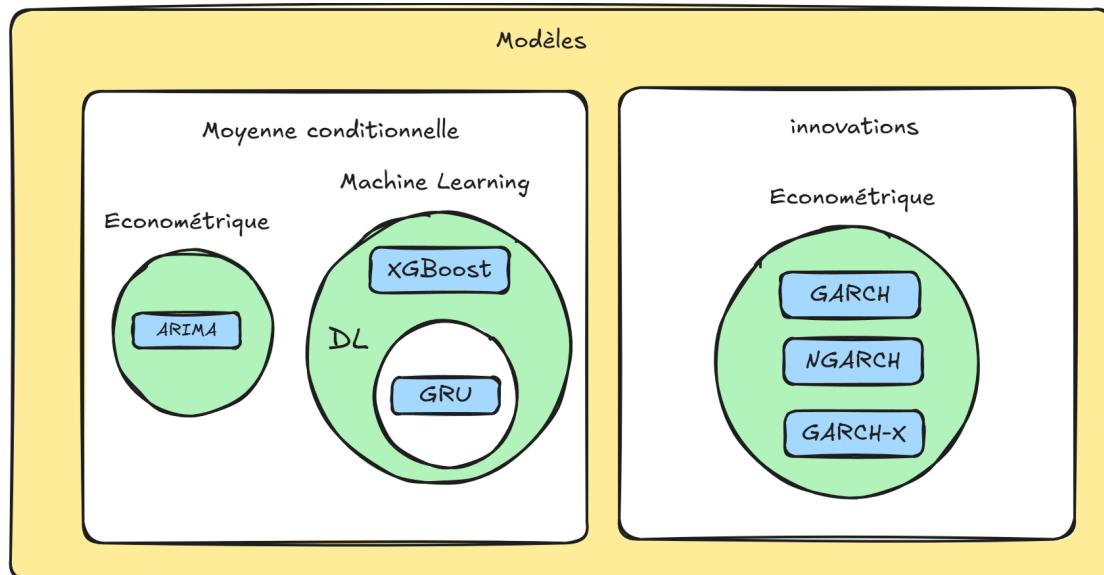
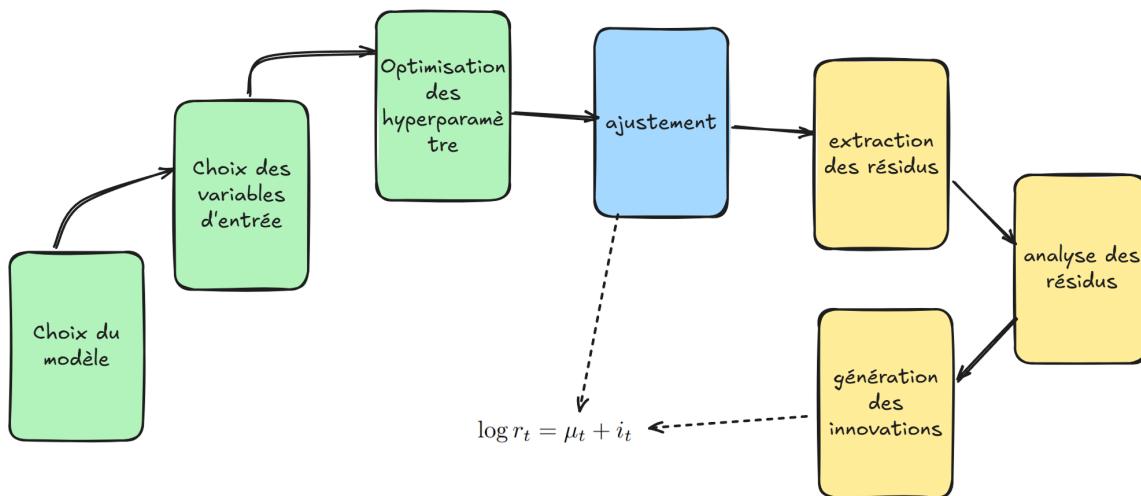


FIGURE 4.1 – Modèles de prévision

En se basant sur l'existant, il existe deux grandes familles de modèles, ou plus précisément trois. Nous avons les modèles économétriques et nous avons les modèles de machine learning & deep learning. Parmi les modèles économétriques, nous allons d'une part choisir le modèle ARIMA, connu pour sa capacité à capturer l'information séquentielle. Et les modèles GARCH, connus pour leur capacité à pouvoir expliquer l'ampleur ou encore la dépendance des amples.

Compte tenu d'une série temporelle. Enfin, parmi les modèles de machine learning, choisissons le modèle XGBoost, connu premièrement pour sa robustesse, sa capacité à pouvoir traiter des informations "sales". Enfin nous choisissons le modèle GRU, capable de traiter l'information séquentielle qui est plus parcimonieux que les modèles LSTM.

### 4.3 Pipeline de modélisation



**FIGURE 4.2 – Cadre de modélisation : des variables d'entrée à la moyenne  $\mu_t$ , puis des résidus  $\hat{i}_t$  à la variance conditionnelle  $\sigma_t^2$ .**

Le dispositif suit une logique commune à tous les algorithmes afin de préserver la comparabilité et d'éviter toute fuite d'information. La série des log-rendements est décomposée en

$$\log r_t = \mu_t + i_t, \quad i_t = \sigma_t z_t,$$

où  $\mu_t$  désigne la moyenne conditionnelle prévisible au temps  $t$  et  $i_t$  un processus d'innovation<sup>1</sup>.

Afin de proposer un cadre uniifié et standardisé, notre approche de prédiction se concentrera sur sept principales phases. La première, le choix du modèle, notamment le modèle qui fera la prédiction de la moyenne conditionnelle. Ensuite, le choix des variables d'entrée. Bien évidemment, nous avons fait le constat que certains modèles, tels que le modèle économétrique, ne sont capables qu'en général de pouvoir prendre l'information temporelle ou encore séquentielle, contrairement aux modèles de Machine Learning et de Deep Learning. Ces variables aussi sont prises d'après le résultat de l'étude dans le chapitre précédent. Ensuite, l'optimisation des paramètres. Cette étape est cruciale et permet de choisir les hyperparamètres ou encore la structure qui va venir maximiser l'efficacité du modèle. Le modèle est ensuite ajusté sur les différentes données. Cet ajustement est évalué. Bien évidemment, le modèle n'est pas parfait et ne le sera pas, car il est impossible de pouvoir prédire le futur à la perfection. Mais notre rôle sera donc d'examiner ses résidus. Les résidus notés  $\hat{i}_t = \log r_t - \hat{\mu}_t$  sont extraits, ensuite analysés pour comprendre la loi sous-jacente. Ces résidus sont ensuite générés après bonne compréhension de

1.  $\sigma_t^2 = \text{Var}(i_t | \mathcal{F}_{t-1})$  la variance conditionnelle,  $z_t$  étant centré de variance unitaire (d'une loi à déterminer)

cette loi. la *génération des innovations* consiste à calibrer  $\hat{\sigma}_t$  via une famille GARCH (GARCH, NGARCH, GARCH-X) puis à former des prévisions cohérentes en combinant  $\hat{\mu}_{t+h}$  et  $\hat{\sigma}_{t+h}$ .

Le choix de ces trois familles de modèles GARCH a été choisi principalement après une étude sur la modélisation de log rendement par les différentes extensions des modèles GARCH. Notamment, il a été conclu que les extensions telles que GJR-GARCH, I-GARCH, et etc. ont approximativement les mêmes performances, comme présenté dans la figure 4.3. Raison pour laquelle, au lieu de se cantonner à des extensions classiques, il a été préférable de choisir des modèles qui sont plus innovants et plus alignés avec nos besoins.

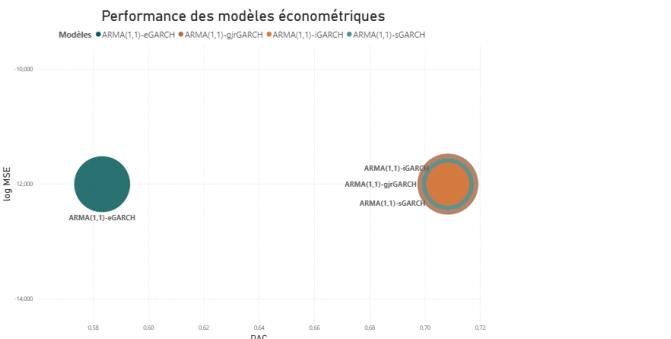


FIGURE 4.3 – Performance des modèles économétriques

## 4.4 Modélisation

### 4.4.1 ARIMA

Pour le début de cette expérience, il est important de fixer des paramètres qui nous seront utiles. Premièrement, le label length, qui est la prédiction à un pas de temps. Ensuite, la window label, qui est la prédiction totale sur une semaine, définie à 164 heures (une semaine). Et enfin, la graine pour la reproductibilité, qui est fixée à 42. Le jeu de données ou les jeux de données utilisés sont au nombre de deux. Premièrement, celui issu du chapitre précédent pour la recherche. Ensuite, le jeu de données original contenant les prix du bitcoin. Pour la traçabilité et la transformation inverse dans l'optique de retrouver les prix originaux à partir des logs rendements prédits.

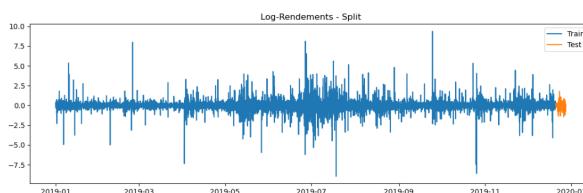


FIGURE 4.4 – Split du dataset pour le modèle ARIMA

**Choix des hyperparamètres** Le modèle ARIMA est défini par 3 paramètres, P, D et Q. P pour le retard (Autoregressive), D pour l'ordre de différenciation et Q pour la moyenne mobile (Moving Average). Après l'utilisation de la fonction Auto-ARIMA, les différents ordres souhaitables sont les suivants. 1 pour le Moving Average, 1 pour l'autoregressif et 0 pour l'ordre d'intégration. Ce

résultat n'est pas étonnant puisque, en général, ces paramètres sont parcimonieux et, d'après l'analyse du chapitre précédent, il était visible que la série temporelle n'avait pas besoin d'être intégrée, car en apparence, elle ne présente pas de tendance.

Variable dépendante	<code>return</code>
Nombre d'observations	8500
Échantillon	01-01-2019 → 12-21-2019
Modèle	ARIMA(1, 0, 1)
Type de covariance	<code>opg</code>
Log-vraisemblance	-9195.598
AIC	18399.196
BIC	18427.387
HQIC	18408.816

Lecture directe du résumé `statsmodels` (SARIMAX).

**TABLE 4.1 – ARIMA(1, 0, 1) sur les log-rendements : métainformations d'ajustement**

Paramètre	Est.	Std.err	<i>z</i>	<i>p</i> -val.	2.5%	97.5%
const	-0.01	0.01	-1.01	0.31	-0.02	0.01
ar.L1	-0.58	0.16	-3.66	0.00	-0.90	-0.27
ma.L1	0.60	0.16	3.82	0.00	0.29	0.91
sigma2	0.51	0.00	233.88	0.00	0.51	0.51

Intervalle à 95 %. Les lignes `ar.L1` et `ma.L1` désignent les composantes AR et MA au retard 1.

**TABLE 4.2 – ARIMA(1, 0, 1) : coefficients estimés**

Ljung–Box $Q(L1)$	0.89	( $\text{Prob}(Q) = 0.35$ )
Hétéroscédasticité (H)	1.68	( $\text{Prob}(H)$ , bilatéral = 0.00)
Jarque–Bera (JB)	235, 444.12	( $\text{Prob}(JB) = 0.00$ )
Asymétrie (Skew)	0.05	
Kurtosis	28.73	

Statistiques et *p*-valeurs reprises du résumé `statsmodels`.

**TABLE 4.3 – ARIMA(1, 0, 1) : diagnostics sur les résidus**

**Ajustement** Les tableaux 4.1, 4.2 et 4.3 rassemblent les résultats d'ajustement de l'ARIMA(1, 0, 1) sur les log-rendements. Le premier tableau regroupe des informations utiles surtout pour la comparaison de plusieurs spécifications (tailles d'échantillon, log-vraisemblance, AIC/BIC/H-QIC) ; nous ne les exploitons pas davantage ici<sup>2</sup>.

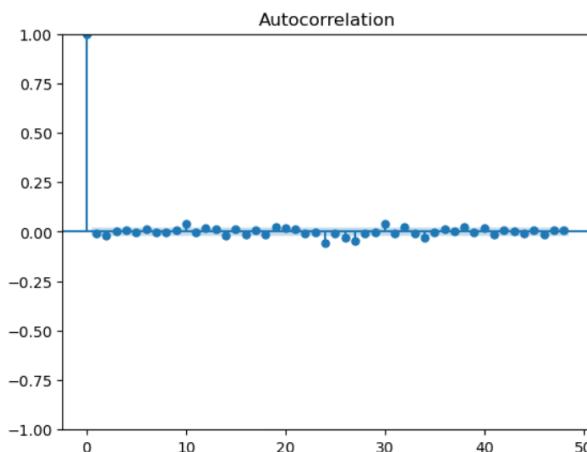
Le second tableau présente les paramètres estimés. La constante apparaît *non significative* ( $p = 0,313 > 0,05$ ), ce qui est cohérent avec des log-rendements globalement centrés. En revanche, les composantes d'*autorégression* AR(1) et de *moyenne mobile* MA(1) sont significatives (valeurs- $p < 0,001$ ), indiquant une dépendance de court terme captée par le couple  $(\phi_1, \theta_1)$ . La variance

2. Ces informations sont utiles pour la comparaison des modèles

de l'innovation, notée  $\sigma^2$ , est estimée à 0,5096 et est hautement significative : elle mesure la dispersion des erreurs une fois la moyenne conditionnelle retirée.

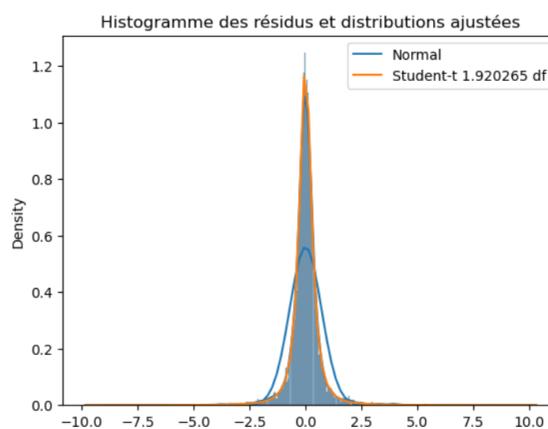
Le troisième tableau regroupe les diagnostics sur les résidus. Le test de Ljung–Box au retard 1 ne rejette pas l'absence d'autocorrélation de premier ordre ( $p = 0,35$ ), mais l'hétérosécédasticité est marquée (statistique  $H = 1,68$ ,  $p \approx 0$ ) et la normalité est nettement rejetée (Jarque–Bera très élevé, kurtosis = 28,73). Ces éléments confirment la pertinence d'un second étage de modélisation de la variance appliqué aux résidus.

**Analyse des résidus** L'analyse débute avec l'autocorrelation function.



**FIGURE 4.5 – Graphique de la fonction d'autocorrélation ARIMA**

**ACF** La figure ci-dessus (figure 4.30) présente l'autocorrelation function sur les résidus issus de l'ajustement du modèle ARIMA. Il apparaît alors très brièvement que, à certains lags, l'autocorrélation dépasse le niveau de signification. Une partie de l'information n'a pas été captée par le modèle ARIMA, car cette dépendance temporelle persiste.



**FIGURE 4.6 – Histogramme des résidus et distribution ajustées - ARIMA**

**Ajustement Graphique** La figure ici-dessus (figure 4.6) présente l'ajustement sur nos résidus par deux lois, notamment la loi normale et la loi de student avec un degré de liberté égal à 1, 92.

Il apparaît de manière claire que la loi de student s'ajuste le mieux sur nos résidus par rapport à la loi normale.

Pour pouvoir confirmer nos dire nous passons des tests statistiques plus fiables .

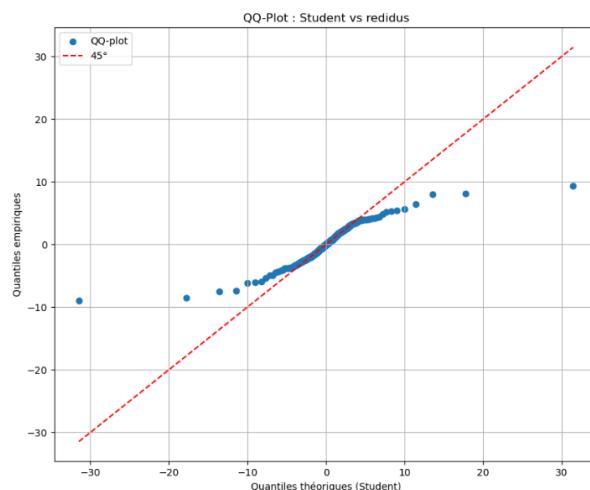
Modèle	AD stat. <sup>a</sup>	Log-vrais.	AIC	$k$	ddl (t)
Residus	407.16	-9195.60	18 395.00	2.00	-
Student- <i>t</i>	71.02	-6350.52	12 707.00	3.00	1.92

<sup>a</sup> AD = statistique d'Anderson–Darling calculée *contre la normalité*. Pour les autres métriques les résidus sont simuler comme loi normale en récupérant les paramètres par ajustement; la comparaison principale repose sur la log-vraisemblance et l'AIC.

**TABLE 4.4** – Comparaison de lois pour les résidus de l'ARIMA(1, 0, 1)

À présent, nous utilisons la distance d'Anderson et Darling<sup>3</sup>. Ainsi, nous évaluons la distance entre la loi de student et la loi normale, et nos résidus de la loi normale. Il apparaît très clairement sur le tableau (tableau 4.4) que la distance entre nos résidus de la loi normale est très grande que celle de student par rapport à la loi normale.

Ayant l'information selon laquelle nos résidus sont très éloignés de la loi normale, nous nous appesantirons donc sur la loi de Student. Mais pour s'assurer que la loi de Student est adaptée à nos besoins, nous analyserons en profondeur les queues de distribution, car d'après le graphique sur la densité, nous avons l'impression que l'ajustement est parfait.

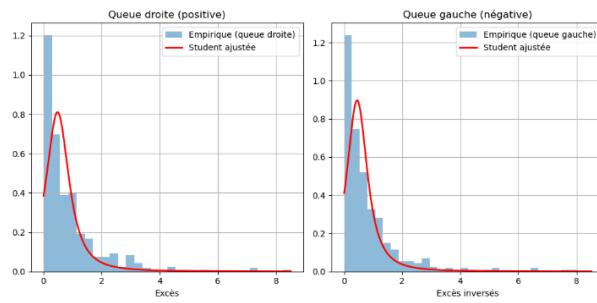


**FIGURE 4.7** – QQ-Plot entre la loi de student et les résidus - ARIMA

Mais ce graphique peut être trompeur. Comme le révèle le graphique du QQPlot(figure 4.7), on observe une forme en S, ce qui veut dire que la distribution d'origine, notamment celle des résidus, est possède des queues plus épaisse que la distribution de la loi de Student.

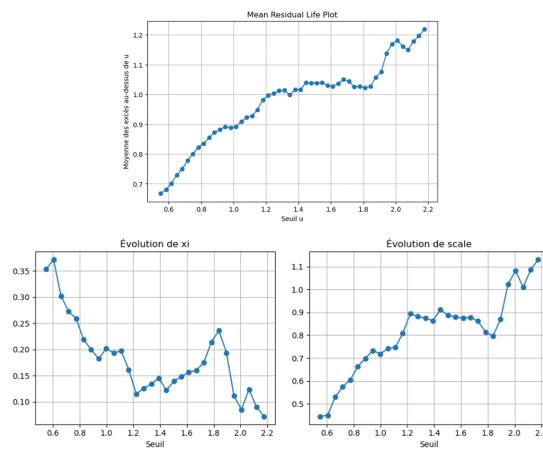
Il apparaît donc que la loi de Student n'est pas adaptée (figure 4.8) pour simuler les événements extrêmes qui peuvent souvent arriver sur le marché de la cryptomonnaie.

3. une version modifiée de la distance de Kolmogorov-Minov, et qui va prendre en compte les écarts au niveau des queues



**FIGURE 4.8 – Ajustement des queues par la loi de student - ARIMA**

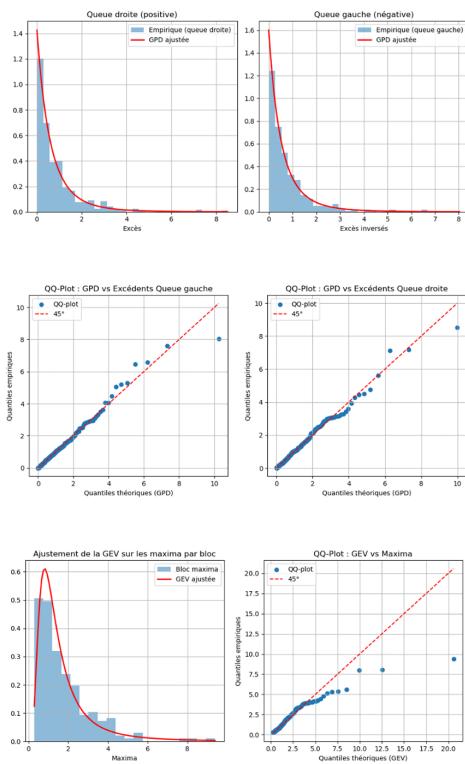
Ayant un problème au niveau des queues, nous allons nous tourner vers des lois adaptées. Ici, nous pouvons nous aider de la théorie des valeurs extrêmes qui nous aide avec deux principales lois, notamment la loi des extrêmes généralisées et la loi du pareto généralisé.



**FIGURE 4.9 – Evolution des paramètres de la loi GPD en fonction du seuil - ARIMA**

Bien. Dans le but d'utiliser la loi des extrêmes avec les deux familles de lois, notamment la loi des extrêmes généralisées et la loi de Pareto généralisée, nous utilisons une méthode constructive. En ce qui concerne la loi de Pareto généralisée, qui est une méthode par seuil, il est important de choisir le bon seuil avant toute analyse. Pour ce faire, nous allons analyser premièrement le graphique de la moyenne des excès (figure 4.9), c'est-à-dire choisir une plage de valeur du seuil où la fonction moyenne des excès paraît linéaire. Pour ce faire, nous allons tester une plage de valeur entre le quantile 90% et le quantile 99%. Ensuite, nous allons étudier la stabilité des différents paramètres, notamment le paramètre de forme et le paramètre d'échelle. D'après la figure la moyenne des excès, il apparaît qu'une plage comprise entre 0.8 et 1,5 est souhaitable. Au niveau de la figure de l'évolution du paramètre de forme, on note une stabilité apparente entre 1.2 et 1.6 par rapport aux différents changements dans la figure. Au niveau de l'évolution du paramètre d'échelle, il y a une stabilité aussi apparente entre 1.2 et 1.4. Après quoi, l'instabilité s'installe. En effectuant une intercession de toutes ces valeurs, le seuil **1,3** est retenu.

En ce qui concerne la loi des extrêmes généralisées, qui est une méthode par bloc, nous choisissons un paramètre de 40 pour le nombre d'éléments par bloc. Ce choix est justifié car un nombre trop peu pourrait entraîner une mauvaise convergence ou encore de mauvais résultats sur la loi des extrêmes généralisés et un nombre trop élevé entraînerait peu de données.



**FIGURE 4.10 – Analyse de l’ajustement des lois extrêmes sur les queues des résidus - ARIMA**

D’après la figure figure 4.10, il apparaît clairement que la loi qui s’ajuste le mieux au cas des résidus est la loi du Pareto généralisé. Car on observe un décalage important dans le QQPlot et dans la densité de la loi des extrêmes généralisés, qui, au contraire, au niveau de la loi du Pareto généralisé, s’ajuste bien à nos données. En conclusion, deux lois sont nécessaires afin de pouvoir reproduire les innovations. Notamment la loi de Stuten pour pouvoir simuler les résidus sur un intervalle intermédiaire ou assez courant, et la loi de Pareto généralisée pour pouvoir simuler les résidus pour les événements extrêmes. Pour pouvoir confronter ou mixer ces différentes lois, nous allons utiliser la loi des mélanges, qui consiste à effectuer une combinaison convexes des différentes lois sous-jacentes.

**Démonstration** On travaille avec des innovations standardisées  $Z$  (centrées, variance unitaire). On fixe deux seuils (potentiellement différents)

$$t_L = -u_{\text{neg}} < 0, \quad t_R = +u_{\text{pos}} > 0,$$

et l’on découpe l’axe en trois zones : centre  $[t_L, t_R]$ , queue gauche  $(-\infty, t_L)$ , queue droite  $(t_R, +\infty)$ .

**GPD et excès positifs (POT).** Pour  $x > t_R$ , l’excès vaut  $y_R = x - t_R \geq 0$ ; pour  $x < t_L$ ,  $y_L = t_L - x \geq 0$ . On ajuste une GPD  $(\xi, \beta)$  sur  $y \geq 0$  :

$$g(y ; \xi, \beta) = \frac{1}{\beta} \left( 1 + \xi \frac{y}{\beta} \right)^{-1/\xi - 1}, \quad \text{support : } 1 + \xi y / \beta > 0.$$

La densité en  $x$  devient

$$g_L(x) = g(t_L - x ; \xi_L, \beta_L) \mathbf{1}_{\{x < t_L\}}, \quad g_R(x) = g(x - t_R ; \xi_R, \beta_R) \mathbf{1}_{\{x > t_R\}}.$$

**Centre (Student tronquée).** Soit  $s_V(x)$  une Student- $t$  de moyenne 0 et variance 1 (facteur d'échelle  $\sqrt{(\nu - 2)/\nu}$ ). La version tronquée et normalisée sur  $[t_L, t_R]$  est

$$s_{\text{tr}}(x) = \frac{s_V(x)}{F_V(t_R) - F_V(t_L)} \mathbf{1}_{\{t_L \leq x \leq t_R\}}.$$

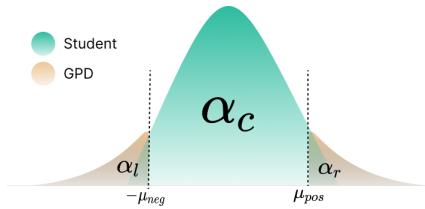


FIGURE 4.11 – loi des mélanges

**Définition de la densité splicée et conditions** On cherche une densité par morceaux

$$f(x) = \alpha_C s_{\text{tr}}(x) + \alpha_L g_L(x) + \alpha_R g_R(x), \quad \alpha_* \geq 0, \quad \alpha_C + \alpha_L + \alpha_R = 1. \quad (4.1)$$

**Continuité aux jonctions.** Comme  $s_{\text{tr}}$  s'annule hors de  $[t_L, t_R]$ , la continuité en  $t_L$  et  $t_R$  impose

$$\alpha_C s_{\text{tr}}(t_L) = \alpha_L g_L(t_L) = \alpha_L / \beta_L, \quad \alpha_C s_{\text{tr}}(t_R) = \alpha_R g_R(t_R) = \alpha_R / \beta_R,$$

soit

$$\alpha_L = \alpha_C s_{\text{tr}}(t_L) \beta_L, \quad \alpha_R = \alpha_C s_{\text{tr}}(t_R) \beta_R. \quad (4.2)$$

**Normalisation globale.** En injectant (4.2) dans  $\alpha_C + \alpha_L + \alpha_R = 1$ , on obtient

$$\boxed{\alpha_C = \frac{1}{1 + s_{\text{tr}}(t_R) \beta_R + s_{\text{tr}}(t_L) \beta_L}, \quad \alpha_R = \frac{s_{\text{tr}}(t_R) \beta_R}{1 + s_{\text{tr}}(t_R) \beta_R + s_{\text{tr}}(t_L) \beta_L}, \quad \alpha_L = \frac{s_{\text{tr}}(t_L) \beta_L}{1 + s_{\text{tr}}(t_R) \beta_R + s_{\text{tr}}(t_L) \beta_L}.} \quad (4.3)$$

Ces poids assurent la continuité et somment à 1.

**Forme finale par morceaux** En développant (4.1) :

$$f(x) = \begin{cases} \alpha_L g(t_L - x ; \xi_L, \beta_L), & x < t_L, \\ \alpha_C \frac{s_V(x)}{F_V(t_R) - F_V(t_L)}, & t_L \leq x \leq t_R, \\ \alpha_R g(x - t_R ; \xi_R, \beta_R), & x > t_R. \end{cases}$$

Le terme  $t_L - x$  à gauche vient de la définition de l'excès positif  $y_L = t_L - x \geq 0$ .

### Vraisemblance

**(a) Sur données standardisées  $Z_t$ .** Pour  $\{z_t\}_{t=1}^T$  et  $\vartheta = (\nu, t_L, t_R, \xi_L, \beta_L, \xi_R, \beta_R)$  :

$$\ell_Z(\vartheta) = \sum_{t=1}^T \log f(z_t ; \vartheta), \quad \log f(z_t ; \vartheta) = \begin{cases} \log \alpha_L + \log g(t_L - z_t ; \xi_L, \beta_L), & z_t < t_L, \\ \log \alpha_C + \log s_{\text{tr}}(z_t), & t_L \leq z_t \leq t_R, \\ \log \alpha_R + \log g(z_t - t_R ; \xi_R, \beta_R), & z_t > t_R. \end{cases}$$

**(b) En GARCH, sur résidus non standardisés  $\varepsilon_t = \sqrt{h_t} Z_t$ .** Conditionnellement à  $\mathcal{F}_{t-1}$  et pour  $h_t(\theta_H) > 0$  :

$$f_{\varepsilon_t | \mathcal{F}_{t-1}}(x ; \vartheta, \theta_H) = \frac{1}{\sqrt{h_t(\theta_H)}} f\left(\frac{x}{\sqrt{h_t(\theta_H)}} ; \vartheta\right),$$

d'où la log-vraisemblance conditionnelle

$$\ell(\vartheta, \theta_H) = \sum_{t=1}^T \left[ \log f\left(\frac{\varepsilon_t}{\sqrt{h_t(\theta_H)}} ; \vartheta\right) - \frac{1}{2} \log h_t(\theta_H) \right].$$

En pratique, on ajuste  $(t_L, t_R, \xi_{\pm}, \beta_{\pm}, \nu)$  sur les  $z_t = \varepsilon_t / \sqrt{h_t}$ .

Pour trouver le facteur nous part de la décomposition

$$\varepsilon_t = \sqrt{h_t(\theta_H)} Z_t, \quad Z_t \text{ a pour densité } f(\cdot ; \vartheta), \quad Z_t \perp\!\!\!\perp \mathcal{F}_{t-1}.$$

Conditionnellement à  $\mathcal{F}_{t-1}$ , la quantité  $h_t(\theta_H)$  est *connue* (mesurable), donc on peut la considérer comme une constante  $h_t > 0$ . Pour  $x \in \mathbb{R}$ ,

$$\Pr(\varepsilon_t \leq x | \mathcal{F}_{t-1}) = \Pr\left(Z_t \leq \frac{x}{\sqrt{h_t}}\right) = F\left(\frac{x}{\sqrt{h_t}}\right),$$

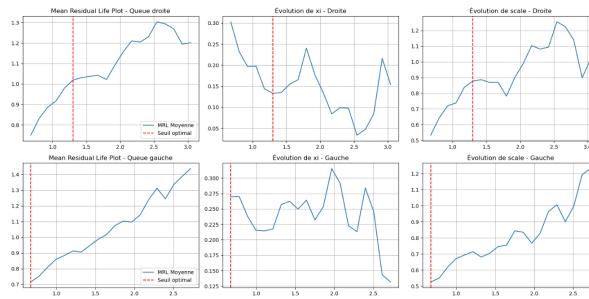
où  $F$  est la fonction de répartition de  $Z_t$ . En dérivant par rapport à  $x$ ,

$$f_{\varepsilon_t | \mathcal{F}_{t-1}}(x) = \frac{d}{dx} F\left(\frac{x}{\sqrt{h_t}}\right) = f\left(\frac{x}{\sqrt{h_t}} ; \vartheta\right) \cdot \frac{d}{dx}\left(\frac{x}{\sqrt{h_t}}\right) = \frac{1}{\sqrt{h_t}} f\left(\frac{x}{\sqrt{h_t}} ; \vartheta\right).$$

On peut retrouver la même formule en utilisant la formule de changement de variable (pour la densité). car on sait que la fonction  $y = g(X)$  où  $g(x) = \sqrt{h_t} * x$  est monotone et dérivable (sachant l'information jusqu'à l'instant  $t$ ).

**Support et signe de  $\xi$**  Pour  $\xi > 0$  (queues à la Pareto), le support de la GPD est  $[0, \infty)$ ; Donc au niveau des extrémités on aura pas de problèmes en terme de support au niveau des queues.

À partir de l'algorithme, nous avons trouvé les différents poids des mélanges. Nous avons notamment le poids du centre qui est environ 0, 68 et qui est un peu normal. Ici, la loi de Student prend une bonne partie de la distribution. Ensuite, nous voyons que c'est la loi de Pareto



**FIGURE 4.12 –** Paramètres pour les queues de distribution - ARIMA

**TABLE 4.5 –** Paramètres GPD par queue (résidus ARIMA)

	Queue droite ( $x > t_R$ )	Queue gauche ( $x < t_L$ )
Seuil <sup>a</sup>	1.30	-0.67
Paramètre de forme $\xi$	0.13	0.27
Paramètre d'échelle $\beta$	0.88	0.53

<sup>a</sup>  $t_R$  et  $t_L$  sont les seuils de la densité splicée Student-GPD.

généralisée sur la queue gauche qui prend plus d'ampleur, donc qui est plus susceptible de survenir. Et enfin, nous avons la queue droite qui a un poids égal à 0, 15.

**TABLE 4.6 –** Poids du splice et seuils associés (résidus ARIMA)

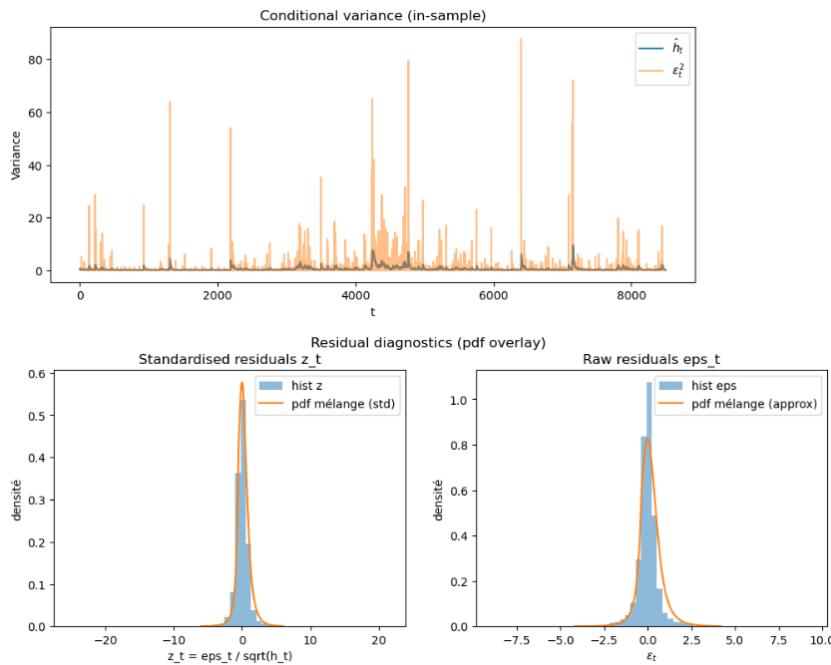
$\alpha_C$	$\alpha_L$	$\alpha_R$	$t_L$	$t_R$	$u_{\text{neg}} =  t_L $	$u_{\text{pos}} = t_R$
0.68	0.17	0.15	-0.67	1.30	0.67	1.30

Somme des poids  $\alpha_C + \alpha_L + \alpha_R \simeq 1$  (à l'arrondi près).

**TABLE 4.7 –** GARCH(1, 1) — paramètres estimés et métriques (résidus ARIMA)

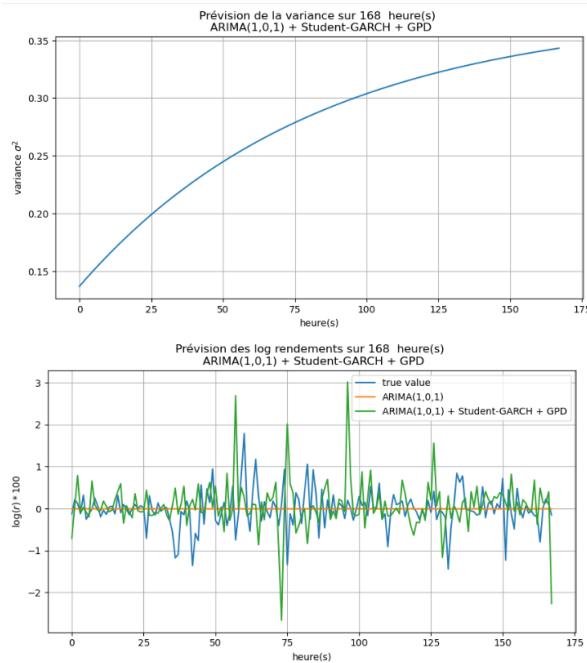
Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.07
$\beta$	0.92
Persistante ( $\beta + \alpha$ )	0.99
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.35
Demi-vie (périodes)	54.07
Taille de l'échantillon	8500.00
Log-vraisemblance négative	5872.44
AIC	11 750.88
BIC	11 772.03

**GARCH** Après un bon ajustement sur le modèle GARCH avec la loi mixte (Student - GPD), les figures ci figure 4.13 présentes représentent les résultats de cet ajustement sous nos données. La première figure présente l'effet des différents chocs sur la variance conditionnelle. On peut notamment voir son impact au cours du temps sur les données d'entraînement. Ensuite, nous pouvons aussi apercevoir les distributions de nos différents résidus ajustés à notre loi de mélange.



**FIGURE 4.13 – Résultats d’ajustement du modèle GARCH (Loi Mixte) - ARIMA**

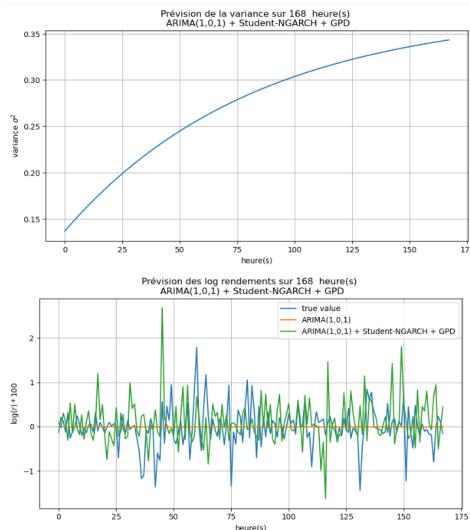
Nous voyons donc qu’il y a un ajustement presque parfait pour nos résidus standardisés. Par contre, au niveau de nos résidus mis à l’échelle, c’est-à-dire les innovations retrouvées, il y a un sous-ajustement, c’est-à-dire la variance prédictée est surestimé par rapport au cas réel.



**FIGURE 4.14 – Variances conditionnelles et rendements prédicts loi mixte GARCH - ARIMA**

La figure ci-dessus présente la variance prédictée et les rendements prédicts retrouvés avec le framework établi dès le départ, notamment en additionnant la moyenne conditionnelle à chaque instant et les innovations générées. Nous voyons donc une évolution croissante de la variance dans le futur et une bonne adaptation du modèle à pouvoir capturer les différentes amplitudes futures. Mais on note aussi des dépassemens importants entre la 90e et la 100e itération des

rendements prédis. Cependant, en général, on voit que le modèle GARCH vient combler la lacune du modèle ARIMA afin de pouvoir capturer les différentes amplitudes qui n'ont pas été expliquées par le modèle ARIMA.



**FIGURE 4.15** – Variances conditionnelles et rendements prédis loi mixte NGARCH - ARIMA

**TABLE 4.8** – NGARCH(1, 1) — paramètres estimés et métriques (résidus ARIMA)

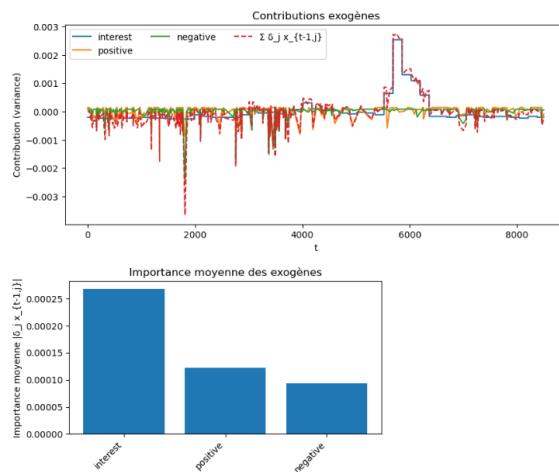
Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.07
$\beta$	0.92
$\theta$	-0.05
Persistante ( $\beta + \alpha(1 + \theta^2)$ )	0.99
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.34
Demi-vie (périodes)	52.28
Taille de l'échantillon	8500.00
Log-vraisemblance négative	5872.24
AIC	11 752.49
BIC	11 780.68

**NGARCH** Nous apercevons une allure presque similaire. Mais bien à noter que la variance qui est présentée ou prédite est une moyenne conditionnelle(l'expérance conditionnelle de la variance future). C'est de cette valeur que se rapprochent les simulations de Monte Carlo. le paramètre  $\beta$  est proche de 1 signe d'une bonne mémoire de volatilité.  $\theta$  le paramètre d'asymétrie<sup>4</sup> est négatif donc les chocs négatifs contribue fortement à la volatilité.Une persistante presque égale à 1, une volatilité inconditionnelle égale à 0.38 et une demi vie longue (environ 52 heure soit une journée et demi) présage d'une volatilité persistante et de grappe de volatité de plus en plus longues.Le modèle s'affiche alors comme très pessimiste en affichant un comportement violent.

4. On utilise le modèle NAGARCH qui prend en compte l'effet de choc négatif

**TABLE 4.9 – GARCH-X — paramètres estimés et métriques (résidus ARIMA)**

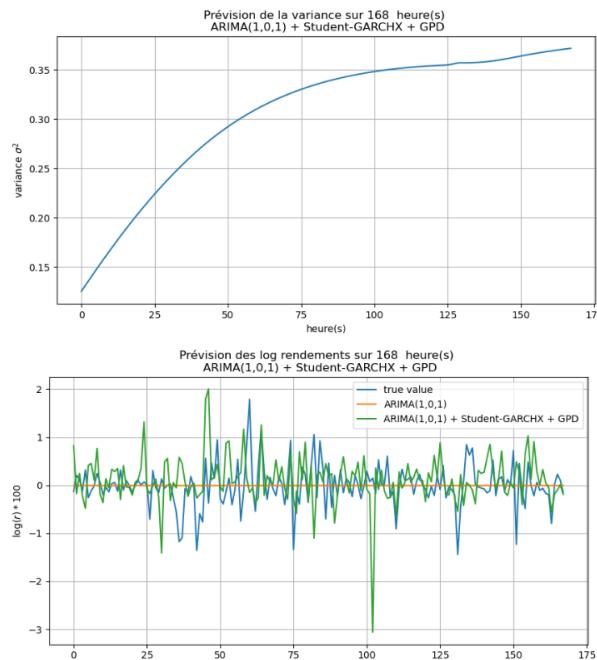
Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.04
$\beta$	0.91
$\gamma$	0.04
$\delta_0$	0.00
$\delta_1$	0.00
$\delta_2$	0.00
Persistante ( $\alpha + \beta$ )	0.95
$E Z  (\text{std})$	0.68
Variance inconditionnelle (approx. baseline)	0.38
Demi-vie (périodes)	13.98
Taille de l'échantillon	8500.00
Log-vraisemblance négative	5834.60
AIC	11 683.21
BIC	11 732.54


**FIGURE 4.16 – Paramètre ajusté GARCH-X- ARIMA**

**GARCH-X** D'après le graphique paramètre ajusté du modèle GARCH-X, nous pouvons apercevoir l'influence des différentes variables exogènes sur la variance conditionnelle. Des variables exogènes c'est la variable “interest”(popularité) qui apporte le plus de contribution à l'évolution de la variance. A contrario les votes négatifs réduisent la variance. Une interprétation de ces résultats serait que le marché devient agité lorsqu'une crypto aurait le vent en poupe et que lorsque les acteurs sont réticents vit à vit d'un évènement il ne préféreraient pas se lancer et miser sur la prudence. Une demi vie de 14 heures présage un marché très changeant avec des grappes de volatilité court

#### 4.4.2 XGBoost

**Recherche des hyperparamètres optimaux** La recherche des hyperparamètres s'effectue à travers une GridSearch sur 5 paramètres tableau 4.10. La panoplie de valeur testé est petite à cause des limitations en termes de puissance de calcul.

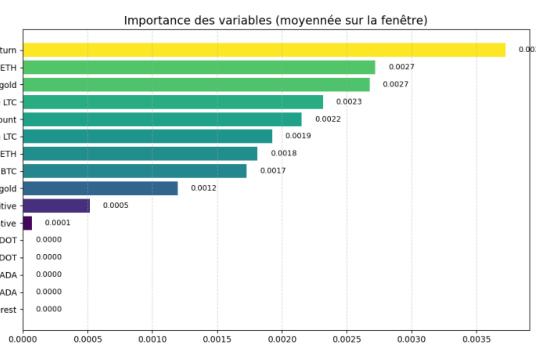


**FIGURE 4.17 – Variances conditionnelles et rendements prédites loi mixte GARCH-X - ARIMA**

**TABLE 4.10 – XGBoost — espace de recherche (Grid)**

Hyper-paramètre	Valeurs testées
n_estimators	{100, 300, 500}
max_depth	{3, 5, 7}
learning_rate	{0.01, 0.05, 0.1}
subsample	{0.8, 1.0}
colsample_bytree	{0.8, 1.0}
<b>Nombre de combinaisons</b>	<b><math>3 \times 3 \times 3 \times 2 \times 2 = 108</math></b>

Après une gridsearch<sup>5</sup> de 954 secondes (16 minutes), les hyperparamètres optimaux trouvés sont les suivants. Premièrement, la profondeur est égale à 3, ce qui signifie que, en général, le modèle ne sur-apprend pas<sup>6</sup>. Un pas d'apprentissage de 0, 01, qui est relativement moyen. Enfin, la proportion utilisée au niveau des features est égale à 80%, ce qui ajoute de la non colinéarité entre les différents arbres.

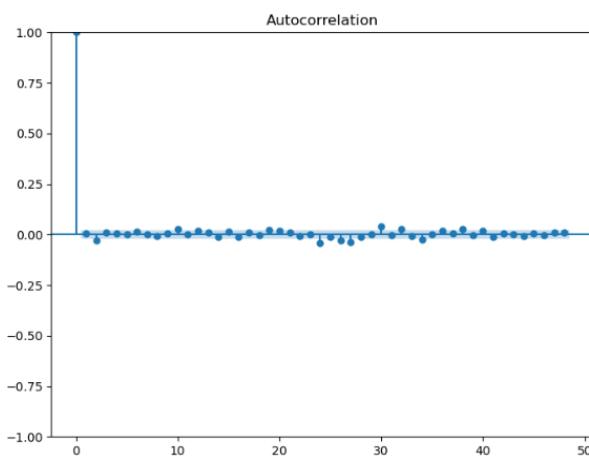


**FIGURE 4.18 – Importance des variables XGBoost**

- 5. Recherche combinatoire
- 6. une profondeur trop grande impliquerait un ajustement trop important sur les données

L'avantage avec ces modèles tels que les modèles à arbre, c'est leur explicativité, la capacité à pouvoir expliquer les résultats qu'ils fournissent. La figure ci présente représente l'importance des différentes variables, bien évidemment moyennées sur la fenêtre. Le premier constat est que la majorité de l'information est déduite à partir des variables sur l'état du marché de certaines grandes cryptomonnaies, telles que le Bitcoin et l'Ethereum. Par contre, les mouvements de marché ne sont pas expliqués par des variables de marché telles que le Polkadot et les altcoins alternatifs. Ensuite, nous voyons que les votes positifs et les votes négatifs ont une plus grande réponse que la popularité de la crypto selon le modèle.

**Analyse des résidus** L'analyse débute avec l'autocorrelation function.



**FIGURE 4.19** – Graphique de la fonction d'autocorrélation XGBoost

**ACF** La figure ci-dessus (figure 4.19) présente l'autocorrelation function sur les résidus issus de l'ajustement du modèle XGBoost. Comme pour le modèle ARIMA précédemment étudié, il apparaît que une partie de l'information n'a pas été captée par le modèle XGBoost, des relations persistent encore à différents instants temporels.

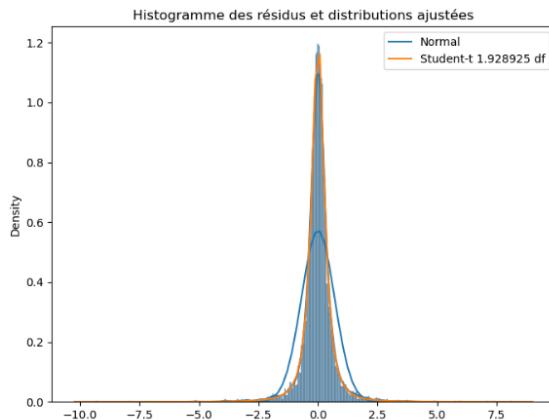
**Ajustement Graphique** L'histogramme des résidus et les distributions ajustées paraissent similaires qu'au cas du modèle ARIMA. Une loi normale qui s'ajuste mal, une colline bien plus affaissé que la montagne créée par les résidus Tandis que la loi de student s'ajuste parfaitement au résidu (en apparence).

Modèle	AD stat. <sup>a</sup>	Log-vrais.	AIC	k	ddl (t)
Résidus	387.09	-8930.99	17 865.98	2.00	-
Student- <i>t</i>	91.84	-6231.18	12 468.36	3.00	1.92

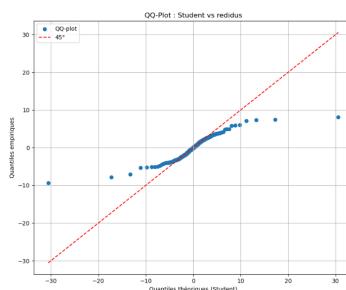
<sup>a</sup> AD = statistique d'Anderson–Darling calculée *contre la normalité*.

**TABLE 4.11** – Comparaison de lois pour les résidus de XGBoost

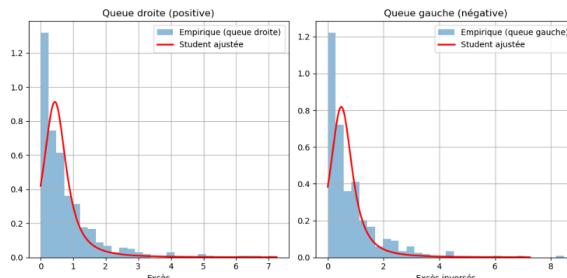
En sommant les informations fournies par le tableau de la statistique de Anderson et Darling, le QQ-plot entre les résidus du modèle XGBoost et de la loi de Student, et enfin de l'ajustement



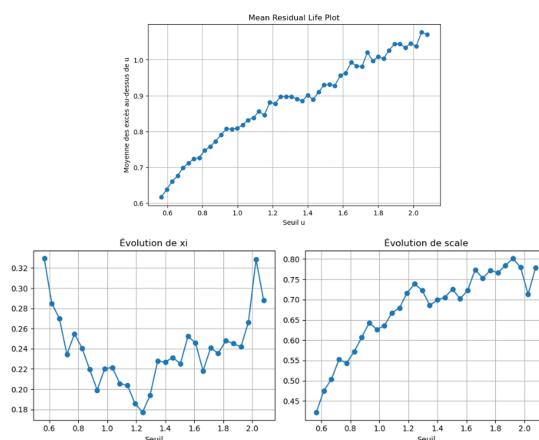
**FIGURE 4.20 – Histogramme des résidus et distribution ajustés - XGBoost**



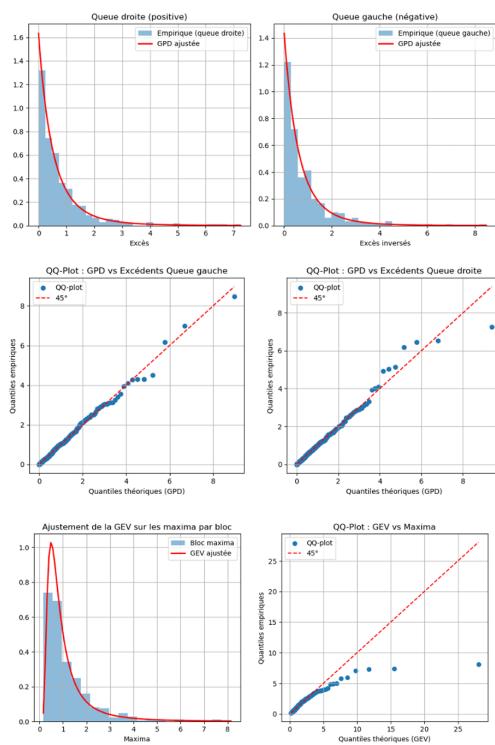
**FIGURE 4.21 – QQ-Plot entre la loi de student et les résidus - XGBoost**



**FIGURE 4.22 – Ajustement des queues par la loi de student - XGBoost**



**FIGURE 4.23 – Evolution des paramètres de la loi GPD en fonction du seuil - XGBoost**



**FIGURE 4.24** – Analyse de l’ajustement des lois extrêmes sur les queues des résidus - XGBoost

au niveau des queues de la loi de Student, on peut conclure que la loi de Student est parfaite pour ajuster les valeurs intermédiaires, mais elle a du mal à pouvoir ajuster parfaitement les événements extrêmes. Alors, nous travaillerons de manière similaire qu’au cas précédent.

**TABLE 4.12** – Poids du splice et seuils associés (résidus XGBoost)

$\alpha_C$	$\alpha_L$	$\alpha_R$	$t_L$	$t_R$	$u_{\text{neg}} =  t_L $	$u_{\text{pos}} = t_R$
0.86	0.07	0.07	-1.76	1.81	1.76	1.30

Somme des poids  $\alpha_C + \alpha_L + \alpha_R \simeq 1$  (à l’arrondi près).

Grâce à la méthode de la loi des mélanges démontrée ci-haut, les différents poids de mélange sont automatiquement estimés. Le poids attribué à la zone centrale est de 0, 86 (bien supérieur à la valeur obtenue pour le modèle ARIMA). Cela se justifie par le fait que le modèle XGBoost a absorbé certaines variations contrairement au modèle ARIMA. Ainsi, les cas rares sont moins expliqués par la loi de Pareto généralisée. Bien évidemment, ce fait aussi peut être illustré par les seuils optimaux trouvés, qui sont bien plus éloignés que pour le modèle ARIMA.

**TABLE 4.13** – Test ADF sur les résidus (trend = constant)

<b>Statistique</b>	-14.467
<b>p-value</b>	< 0.001
<b>Retards (lags)</b>	37
<b>Valeurs critiques</b>	-3.43 (1%), -2.86 (5%), -2.57 (10%)

$H_0$  : racine unitaire ;  $H_1$  : stationnarité faible.

Décision : rejet de  $H_0$  (série stationnaire) au seuil usuel.

**TABLE 4.14 –** Test LM d'hétéroscédaicité ARCH (lags = 48) sur les résidus XGBoost

<b>LM statistic</b>	440.81274886751044
<i>p</i> -value	$6.4268 \times 10^{-65}$

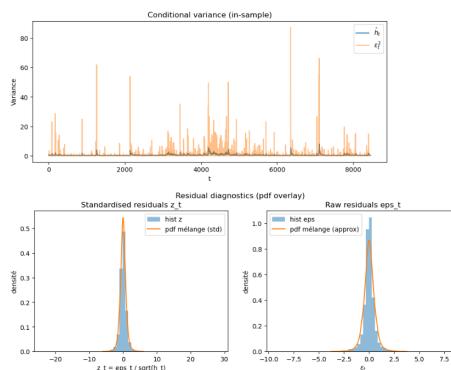
$H_0$  : absence d'effets ARCH.

Décision : rejet très net de  $H_0 \Rightarrow$  présence d'hétéroscédaicité conditionnelle.

Le test d'augmented dickey-fuller (ADF) présente une p-value inférieur aux seuil de 5% donc l'hypothèse null (présence d'une racine unitaire<sup>7</sup>) est rejetté. Le test d'hétérosquédaicité sur les résidus (par la méthode du multiplicateur de Lagrange) se conclu avec une p-value < 5%, donc l'hypothèse d'absence d'effet ARCH est rejetté. A travers ces deux derniers résultats l'utilisation du modèle GARCH est justifié.

**TABLE 4.15 –** GARCH(1, 1) — paramètres estimés et métriques (résidus XGBoost)

Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.06
$\beta$	0.92
Persistance ( $\beta + \alpha$ )	0.98
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.15
Demi-vie (périodes)	29.76
Taille de l'échantillon	8452.00
Log-vraisemblance négative	5578.30
AIC	11 162.59
BIC	11 183.72

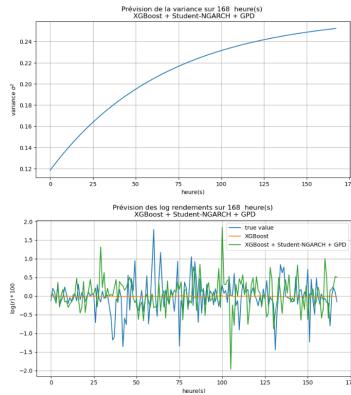

**FIGURE 4.25 –** Résultats d'ajustement du modèle GARCH (Loi Mixte) - XGBoost

**GARCH** Le tableau 4.15 présente les informations d'ajustement du modèle GARCH. On observe une variance inconditionnel petite (0.15), une persistanc forte  $0.98 \approx 1$  et une démi vie<sup>8</sup> de 30 heures(plus d'une journée). le modèle aparait calme et posé au premier abord Cependant il présente une explosion et une agitation prolongée dans le future. Le graphique figure 4.25

7. Absence de stationnarité

8. temps nécessaire pour qu'un choc réduise de 50% la valeur de la variance

est issu de l'ajustement de notre modèle GARCH de loi mixte sur nos résidus. D'après la figure de l'effet des chocs sur la variance, on voit une bonne corrélation entre les chocs et la variance conditionnelle future. De plus, nous voyons que les innovations générées par notre loi s'ajustent au mieux à nos résidus. Cependant, nous avons le même cas de figure que celui du modèle ARIMA. La variance générée est bien plus grande que celle réelle, comme le confirme le graphique figure 4.25.



**FIGURE 4.26** – Variances conditionnelles et rendements prédicts loi mixte NGARCH - XGBoost

**TABLE 4.16** – NGARCH(1, 1) — paramètres estimés et métriques (résidus XGBoost)

Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.07
$\beta$	0.92
$\theta$	-0.11
Persistance ( $\beta + \alpha(1 + \theta^2)$ )	0.99
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.26
Demi-vie (périodes)	46.68
Taille de l'échantillon	8500.00
Log-vraisemblance négative	5576.95
AIC	11 161.90
BIC	11 190.07

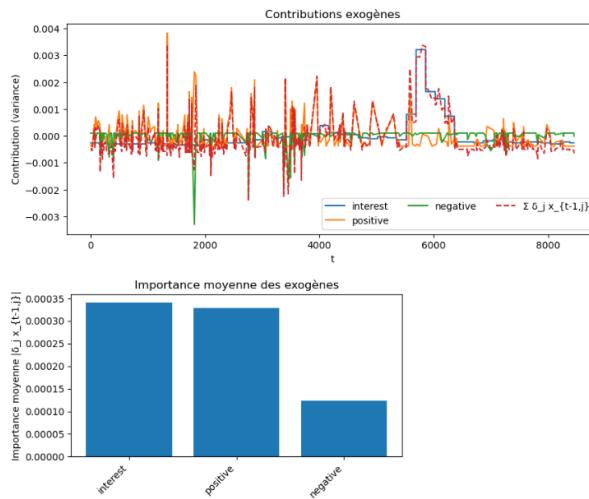
**NGARCH** Le tableau figure 4.13 présente les résultats d'ajustement du modèle NGARCH sur les résidus XGBoost. Le paramètre d'asymétrie  $\theta$  est négatif et bien supérieur à celui du modèle NGARCH-ARIMA. Les chocs négatifs ont bien plus d'influence ici. une volatilité incinditionnelle de 0.26, une démi-vie de 46 heures et une persistance d'environ 1 présente un modèle avec des grappes de volatilité longues.

En regardant attentivement le graphique de la variance conditionnelle avec le modèle NGARCH, on aperçoit que la variance converge vers 0.26, une valeur bien inférieure que celle obtenue à partir des résidus du modèle ARIMA. Ce qui montre qu'une bonne partie de la variation au niveau du modèle XGBoost a été absorbée. Mais lorsqu'on essaie de regarder les clusters au niveau de la prévision des rendements, on voit qu'il existe des parties où la variance est extrêmement forte et où les chocs ont plus d'ampleur. En regardant de plus près sur le tableau

des informations issues de l'ajustement, nous observons une demi-vie de 46, 68 environ 48 heures, soit deux jours, pour lesquels l'effet d'un choc sur la variance conditionnelle diminue de 50%, une valeur plus modérée que celle du modèle ARIMA.

**TABLE 4.17 – GARCH-X — paramètres estimés et métriques (résidus XGBoost)**

Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.04
$\beta$	0.91
$\gamma$	0.03
$\delta_0$	0.00
$\delta_1$	0.00
$\delta_2$	0.00
Persistance ( $\alpha + \beta$ )	0.95
$\mathbb{E} Z  (\text{std})$	0.69
Variance inconditionnelle (approx. baseline)	0.21
Demi-vie (périodes)	14.18
Taille de l'échantillon	8452.00
Log-vraisemblance négative	5544.35
AIC	11 102.70
BIC	11 151.99



**FIGURE 4.27 – Paramètre ajusté GARCH-X- XGBoost**

**GARCH-X** La figure 4.27 représente la contribution des variables exogènes dans la prévision de la variance conditionnelle. Dans le premier graphique, nous notons une contribution assez présente. Deux variables ont un apport conséquent, notamment la variable *interest* qui représente la popularité de la cryptomonnaie, ensuite la variable *positive* qui représente le nombre de votes sur le site CryptoPanic.com. Nous savons que le modèle XGBoost a une bonne capacité, ou bien une meilleure capacité à capter certaines variations, contrairement au modèle ARIMA. On peut expliquer une grande partie de cette implication des variables exogènes par la quantité d'informations qui sont non expliquées dans la moyenne conditionnelle. La demi-vie ici est

d'environ 15 heures, soit une demi-journée. Ceci montre aussi que ce modèle prend en compte les variations rapides du marché, contrairement aux autres modèles qui donnent une demi-vie bien plus grande.

#### 4.4.3 GRU

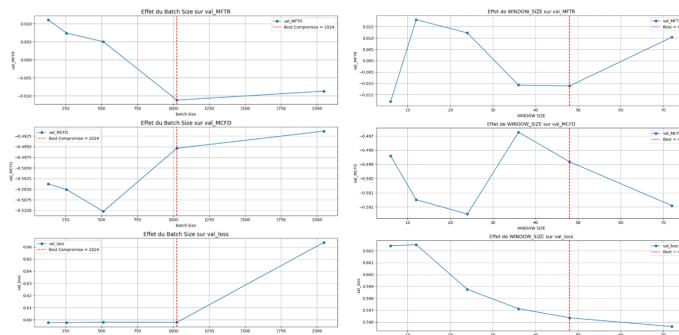


FIGURE 4.28 – Recherche window size et batch size optimale-GRU

**Recherche des hyperparamètres optimaux** La recherche des hyperparamètres optimaux définit à la fois la structure et les algorithmes de descente (Adam) pour notre réseau de neurones, nous avons débuté par une GridSearch. Après une première GridSearch, nous avons obtenu la tangente hyperbolique comme fonction d'activation optimale .La couche intermédiaire est une couche GRU, constituée de 50 neurones, avec une régularisation L1. Enfin, une couche de Dropout (pour éviter le surajustement) est appliquée avant la sortie pour éviter le surapprentissage. Ayant effectué plusieurs tests, nous avons obtenu de mauvais résultats (et la gridsearch devenait un peu trop couteuse). Alors, nous avons changé de stratégie.Nous avons décidé de choisir la fenêtre optimale de données, c'est-à-dire la dépendance maximale à prendre en compte à l'entrée du modèle. Pour ce faire, nous avons choisi une méthode qui prend en compte les différentes métriques de la FinTSB, notamment la MFTR, la MCFD et la MAE. Toutes répartis avec des poids bien précis, en donnant une importance première à la MFTR, car les traders et les gestionnaires sur le marché recherchent du bénéfice. Ensuite, la MCFD qui va donner un poids au modèle pour des prédictions dans le bon sens, et enfin la MAE. En utilisant cette méthode de score pondérée<sup>9</sup>, nous obtenons une fenêtre optimale de 48 heures, soit deux jours. Ensuite, nous avons essayé de rechercher le batch size optimal, toujours dans cet esprit. Grâce à cela, nous avons obtenu un batch size optimal de 1024. En utilisant ces deux dernières méthodes, nous avons amélioré drastiquement les performances du modèle.

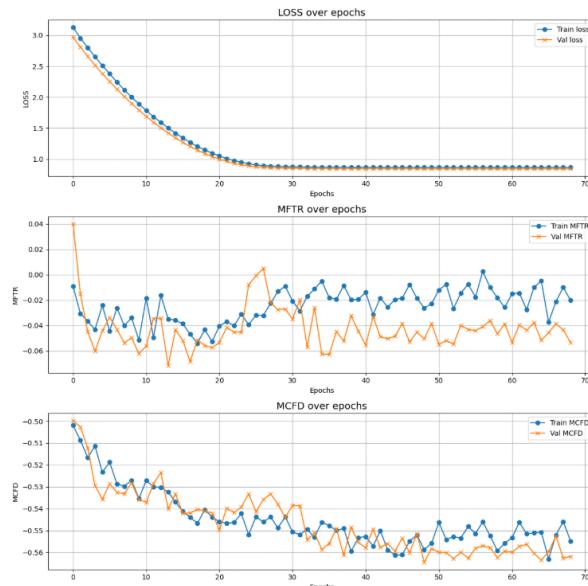
TABLE 4.18 – GRU — meilleur WINDOW\_SIZE (sélection par score pondéré)

Paramètre	Valeur	val_MFTR	val_MCFD	val_loss	compromise_score
WINDOW_SIZE	48.00	-0.01	-0.50	0.60	0.34

9. 40% MFTR, 35% MCFD, 35% MAE

**TABLE 4.19 – GRU — meilleur batch\_size (sélection par score pondéré)**

Paramètre	Valeur	val_MFTR	val_MCFD	val_loss	compromise_score
batch_size	1024.00	-0.01	-0.50	0.60	0.28


**FIGURE 4.29 – Ajustement du modèle GRU**

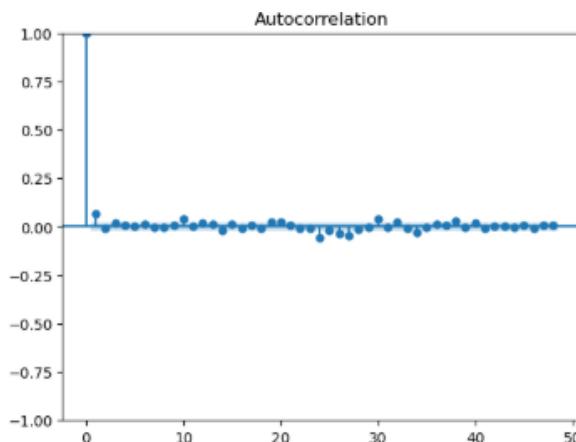
**Ajustement du modèle** Sur la trajectoire d’entraînement, la losse d’entraînement et de validation, notamment la MAE, décroît rapidement, jusqu’à la 30 epoch, puis se stabilise autour d’un même palier, sans écart durable entre les deux courbes. Les métriques de validation, toutefois, montrent un comportement divergent. Au niveau de la MFTR<sup>10</sup>. Sur la MFTR on observe une décroissance jusqu’à la 20<sup>e</sup> epoch ensuite une stabilisation continu. Enfin, au niveau de la MCFD, on voit un mouvement baissier et régulier, parallèle entre les valeurs d’entraînement et les valeurs de validation, pour atteindre un plateau commun à -55%. Confirmant ainsi la stabilité du modèle à pouvoir capturer les bons signaux, mais à ne pas réellement engager du profit sur de nouvelles valeurs.

**Analyse des résidus** L’analyse débute avec l’autocorrelation function.

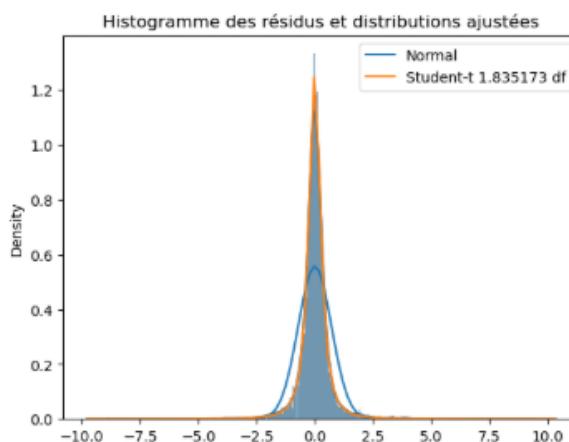
**ACF** La figure ci-dessus (figure 4.30) présente l’autocorrelation function sur les résidus issus de l’ajustement du modèle GRU. Le constat est similaire aux précédents modèles, il y’a encore de l’information dans les résidus qui n’a pas été capturé par le modèle.

**Ajustement Graphique** D’après l’histogramme des résidus et distributions ajustés, on constate toujours avec la même ardeur que le modèle Student s’ajuste au mieux à nos données. On remarque aussi que le degré de liberté, qui est égal ici à 1,83, est inférieur à celui des résidus du modèle ARIMA et du modèle XGBoost. Nous savons que plus ce degré de liberté est grand, notamment supérieur à 3, plus il se rapproche de la loi normale. Ici, on peut donc conclure

10. Ici la valeur utilisé est l’opposé de la MFTR, afin de créer un problème de minimisation au lieu de maximisation, le cas est similaire pour la MCFD



**FIGURE 4.30 – Graphique de la fonction d'autocorrélation GRU**



**FIGURE 4.31 – Histogramme des résidus et distribution ajustés - GRU**

que nos données s'éloignent de plus en plus de la loi normale avec le modèle GRU. On peut donc interpréter cela par le fait que le modèle GRU a absorbé une grande partie des valeurs intermédiaires, mais n'arrive pas à expliquer les valeurs extrêmes.

Modèle	AD stat. <sup>a</sup>	Log-vrais.	AIC	k	ddl (t)
Residus	428.11	-9168.42	18 340.84	2.00	–
Student-t	61.42	-6210.20	12 426.50	3.00	1.92

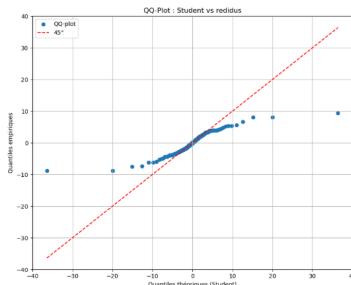
<sup>a</sup> AD = statistique d'Anderson–Darling calculée *contre la normalité*.

**TABLE 4.20 – Comparaison de lois pour les résidus de GRU**

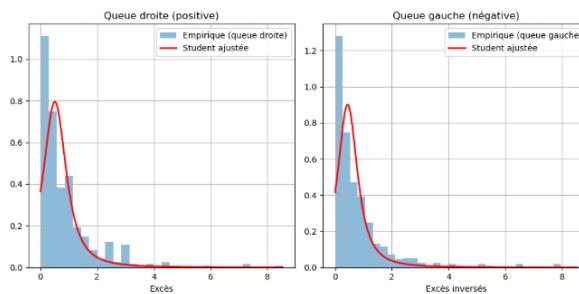
En analysant les données du tableau 4.20 de comparaison de loi pour les résidus du GRU, on constate que nos résidus sont très éloignés de la loi normale en se référant à la distance de Anderson et Darling, qui est plus grande que celle du student.

Le QQPlot au niveau de nos résidus et de la loi de Student nous donne aussi raison. On voit un écart assez important, un peu plus grand que celui du QQPlot dans les autres modèles

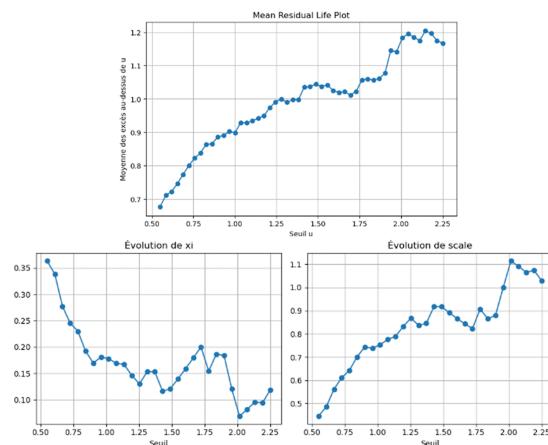
A partir du graphique de l'évolution des paramètres de la loi de GPD en fonction du seuil, nous voyons que la fonction moyenne des excès est linéaire jusqu'à 1, 30. En ce qui concerne le graphique de l'évolution du paramètre de forme et du paramètre d'échelle, on voit une



**FIGURE 4.32 – QQ-Plot entre la loi de student et les résidus - GRU**



**FIGURE 4.33 – Ajustement des queues par la loi de student - GRU**

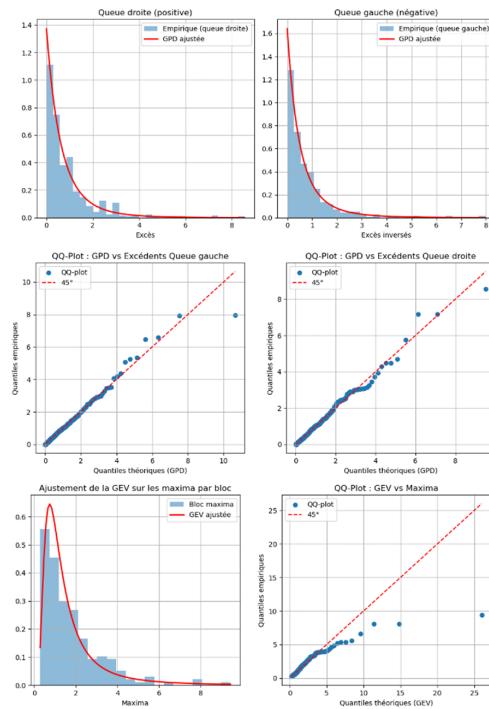


**FIGURE 4.34 – Evolution des paramètres de la loi GPD en fonction du seuil - GRU**

stabilisation entre 1, 25 et 1, 40 pour chacun. Donc, nous allons prendre la valeur de 1, 3 comme seuil pour cette analyse. Un bref aperçu sur la figure d'analyse d'ajustement des lois extrêmes sur les queues des résidus nous permet d'observer un ajustement presque parfait de la loi de Pareto généralisée sur nos résidus comparément à la loi des extrêmes généralisées où on observe un dépassement sur la densité et un écart important au niveau du QQ-plot.

Avant d'aller plus loin, nous effectuons quelques analyses pour assurer et justifier l'utilisation de modèles de variances conditionnelles, notamment en faisant un test de stationnarité et un test d'hétéroscédasticité.

Le test utilisé pour la stationnarité est le test augmented dickey fuller, avec lequel on obtient une p-value extrêmement inférieure au seuil critique de 5%. Donc, l'hypothèse nulle est rejetée. Il existe bien donc une stationnarité faible dans nos résidus. De plus, en observant aussi les résultats du test du multiplicateur de la range, on observe aussi une p-value qui est extrêmement inférieure au seuil du 5%. Donc, il y a une présence d'hétéroscedasticité conditionnelle.



**FIGURE 4.35 – Analyse de l’ajustement des lois extrêmes sur les queues des résidus - GRU**

**TABLE 4.21 – Test ADF sur les résidus (trend = constant)**

<b>Statistique</b>	-14.645
<b>p-value</b>	< 0.001
<b>Retards (lags)</b>	37
<b>Valeurs critiques</b>	-3.43 (1%), -2.86 (5%), -2.57 (10%)

$H_0$  : racine unitaire ;  $H_1$  : stationnarité faible.

Décision : rejet de  $H_0$  (série stationnaire) au seuil usuel.

**TABLE 4.22 – Test LM d’hétéroscedasticité ARCH (lags = 48) sur les résidus**

<b>LM statistic</b>	484.339
<b>p-value</b>	$1.96 \times 10^{-73}$

$H_0$  : absence d’effets ARCH.  
Décision : rejet très net de  $H_0$   
 $\Rightarrow$  présence d’hétéroscedasticité conditionnelle.

**TABLE 4.23 – Poids du splice et seuils associés (résidus GRU)**

$\alpha_C$	$\alpha_L$	$\alpha_R$	$t_L$	$t_R$	$u_{\text{neg}} =  t_L $	$u_{\text{pos}} = t_R$
0.90	0.05	0.05	-2.29	2.64	2.29	2.64

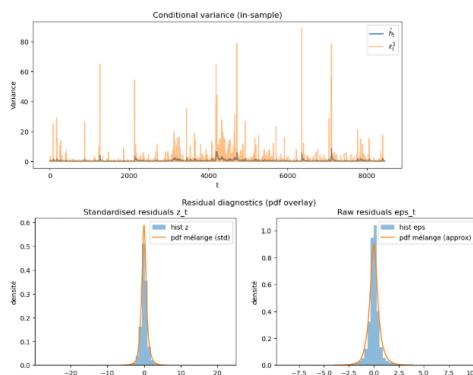
Somme des poids  $\alpha_C + \alpha_L + \alpha_R \simeq 1$  (à l’arrondi près).

Un constat est assez flagrant au niveau des poids de la loi de mélange. La grande importance donnée à la loi de student (90% en terme de contribution), On note aussi une faible contribution symétrique de la loi de paréto généralisé. Ceci peut se traduire par le fait que le modèle GRU, comparé aux deux modèles précédents, a une bonne capacité d'absorber certaines variations. Les variations qui ne sont pas expliquées sont celles des grandes amplitudes, qui sont expliquées notamment en grande partie par le modèle student et par le modèle de Pareto généralisé de manière équitable sur les deux queues. Enfin, nous observons aussi que le seuil optimal trouvé pour la loi de Pareto généralisé sur la queue gauche est bien supérieur par rapport au symétrique à celle de la queue droite. Donc, une majorité des cas extrêmes sur les pertes sont expliqués par la loi de Pareto généralisé.

**TABLE 4.24 – GARCH(1, 1) — paramètres estimés et métriques (résidus GRU)**

Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.06
$\beta$	0.92
Persistante ( $\beta + \alpha$ )	0.98
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.18
Demi-vie (périodes)	34.72
Taille de l'échantillon	8452.00
Log-vraisemblance négative	5563.78
AIC	11 133.56
BIC	11 154.69

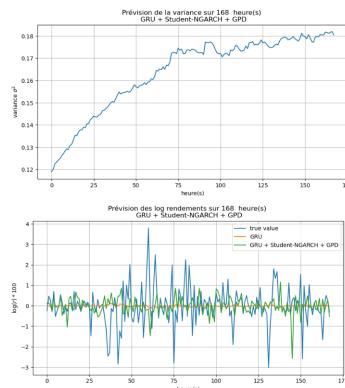
**GARCH** Le tableau 4.24 présente trois composantes qui tire notre attention : une persistance forte (0.98), une variance inconditionnelle faible (certaines amplitudes sont expliquées par le modèle GRU), et une demi-vie très longue, les grappes de volatilité sont longues et la volatilité perdure plus longtemps.



**FIGURE 4.36 – Résultats d'ajustement du modèle GARCH (Loi Mixte) - GRU**

La figure 4.36 résulte de l'ajustement sur nos données par le modèle GARCH avec notre loi mixte. Comme remarqué sur les deux précédents modèles, les résidus standardisés et l'ajustement par le modèle fonctionnent très bien. Tandis qu'on voit toujours au niveau des résidus mis à l'échelle une variance surestimée. Car l'ajustement montre un voile avec un étirement inférieur à

celui des données réelles. Bien évidemment, ceci n'est pas une mauvaise chose, dans la mesure où le but est de pouvoir surestimer le risque derrière afin d'éviter aux traders et aux gestionnaires d'actifs de pouvoir sous-estimer le risque possibles.



**FIGURE 4.37 – Variances conditionnelles et rendements prédicts loi mixte NGARCH - GRU**

**TABLE 4.25 – NGARCH(1, 1) — paramètres estimés et métriques (résidus GRU)**

Paramètre / métrique	Valeur
$\omega$	0.06
$\alpha$	0.07
$\beta$	0.92
$\theta$	0.14
Persistante ( $\beta + \alpha(1 + \theta^2)$ )	0.98
Variance inconditionnelle ( $\omega/(1 - \text{pers.})$ )	0.19
Demi-vie (périodes)	35.18
Taille de l'échantillon	8452.00
Log-vraisemblance négative	5581.03
AIC	11 170.07
BIC	11 198.24

**NGARCH** Le paramètre d'asymétrie  $\theta$  est positif (0.14) donc les chocs positifs contribuent fortement à l'augmentation de la variance. Une demi longue ici nous suggère des grappes de volatilité longue et des périodes de turbulence persistantes.

La figure figure 4.37 présente les variations conditionnelles et rendements prédicts par le modèle NGARCH sous les résidus GRU avec une méthode de Monte Carlo. C'était l'occasion de pouvoir essayer une nouvelle méthode. La méthode traditionnelle était de considérer la variance conditionnelle en moyenne. La méthode avec Monte Carlo permet non seulement d'ajouter de l'incertitude et de pouvoir tester plusieurs simulations. Il est bien vrai que lorsque le nombre de simulations tend vers l'infini alors, les résultats de Monte Carlo se convergent vers la variance conditionnelle future moyenne. En regardant de plus près la figure, la forme se rapproche le plus de ce que l'on connaît habituellement. On constate aussi que la variance converge vers 0, 18, qui est une valeur bien faible de ce qu'on a l'habitude de voir. Ceci se confirme aussi dans la figure représentant les rendements prédicts. Nous voyons alors que les différentes amplitudes sur les rendements prédicts sont sous-estimées par rapport aux rendements réels.

**TABLE 4.26 – GARCH-X — paramètres estimés et métriques (résidus GRU)**

Paramètre / métrique	Valeur
$\omega$	0.00
$\alpha$	0.04
$\beta$	0.91
$\gamma$	0.04
$\delta_0$	0.00
$\delta_1$	0.00
$\delta_2$	0.00
Persistante ( $\alpha + \beta$ )	0.95
$E Z  (\text{std})$	0.69
Variance inconditionnelle (approx. baseline)	0.25
Demi-vie (périodes)	13.47
Taille de l'échantillon	8452.00
Log-vraisemblance négative	5546.19
AIC	11 106.37
BIC	11 155.67

**GARCH-X** La figure suivante montre l'effet des différentes variables exogènes sur la volatilité, en montrant un grand impact de la variable exogène intérêt, qui est reliée à la popularité de la cryptomonnaie sur Internet. Ensuite, On observe une importance décroissante entre les votes négatifs et positifs sur le site CryptoPanic.com. Nous pouvons aussi observer à partir du tableau 4.26 les différents paramètres estimés et des métriques du modèle GARCH-X des résidus XGBoost. Une demi-vie assez très petite, de l'ordre d'une demi-journée. Et une persistance aux alentours de 0.95. La mémoire quant à elle, égale à 0.91, montre une grande importance de la volatilité passée sur ces valeurs futures. Tandis que les variables exogènes, leur implication, pour être plus exact, est très minime, de l'ordre de  $10^{-3}$ .

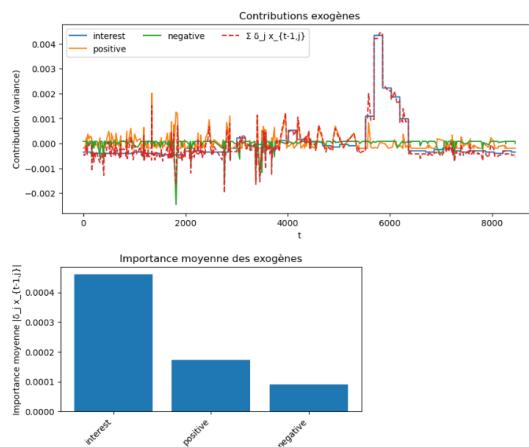

**FIGURE 4.38 – Paramètre ajusté GARCH-X- GRU**

TABLE 4.27 – Tableau comparatif (Modèles × Métriques)

Modèle	Erreur (↓)		Classement (↑)		Portefeuille		Risque	
	MAE	MSE	MCFD	MFTR	ARR	ANV	MIDD	ASR
ARIMA(1, 0, 1)	<b>0.3017</b>	<b>0.1983</b>	0.1677	0.0125	-0.5006	<b>0.0000</b>	0.0132	<b>-362867.56</b>
ARIMA + Student-GARCH + GPD	0.5083	0.5865	0.4910	0.0163	916.42	0.5399	0.0355	12.91
ARIMA + Student-NGARCH + GPD	0.5135	0.4508	0.5150	-0.0436	18218.99	0.4701	0.0223	21.12
ARIMA + Student-GARCH-X + GPD	0.4696	0.4514	0.4850	<b>0.0304</b>	<b>18604.26</b>	0.4752	0.0392	20.94
GRU	0.3036	0.2027	<b>0.7066</b>	0.0025	-0.8585	0.0459	0.0370	<b>-42.54</b>
GRU + Student-GARCH + GPD	0.4188	0.3504	0.5329	0.0409	16.44	0.4069	0.0440	7.23
<b>GRU + Student-NGARCH + GPD</b>	<b>0.5429</b>	<b>0.5432</b>	<b>0.5569</b>	<b>0.0005</b>	<b>521.32</b>	<b>0.5538</b>	<b>0.0396</b>	<b>11.58</b>
GRU + Student-GARCHX + GPD	0.4668	0.4147	0.5269	-0.0076	399.53	0.4360	0.0295	13.97
XGBoost	<b>0.3018</b>	<b>0.1982</b>	0.3772	0.0394	-0.4720	0.0056	0.0122	<b>-114.18</b>
XGBoost + Student-GARCH + GPD	0.4343	0.3383	0.4910	-0.0491	1.09	0.3541	0.0398	2.26
XGBoost + Student-NGARCH + GPD	0.4537	0.3878	0.5210	-0.0348	66.39	0.3950	0.0367	10.86
XGBoost + Student-GARCHX + GPD	0.4460	0.3518	0.4431	0.0123	-0.9623	<b>0.4206</b>	0.0799	-7.58

**Légende :** (↓) à minimiser ; (↑) à maximiser. **Vert** = meilleur sur un critère à maximiser ; **Rouge** = meilleur (plus petit) sur un critère à minimiser ; **Jaune** = explosion numérique détectée ; **Violet** = meilleur compromis (choix manuel).

## 4.5 Evaluation des modèles

Globalement sur le tableau 4.27, les modèles purs de moyenne (ARIMA, XGBoost) obtiennent les plus faibles erreurs ( $MAE \approx 0.302$ ,  $MSE \approx 0.198$ ), mais leurs métriques de portefeuille sont médiocres ou instables (ARR négatif, ASR très dégradé), ce qui montre que minimiser l'erreur ponctuelle ne garantit pas un profil rendement-risque exploitable. À l'inverse, l'ajout d'une volatilité conditionnelle avec innovations à queues épaisse (Student + GPD) dégrade un peu l'erreur mais améliore nettement la qualité de portefeuille : par exemple, ARIMA + (Student-)GARCHX + GPD et ARIMA + (Student-)NGARCH + GPD affichent les ARR les plus élevés ( $\approx 18.6k$  et  $18.2k$ ) avec des ASR positifs et élevés (21), signe d'un meilleur risque ajusté.

Dans la famille réseaux, GRU seul présente le meilleur MCFD (0.7066) mais, comme les modèles purement moyens, conduit à des métriques de portefeuille peu robustes. Les hybrides GRU + GARCH/NGARCH + GPD rééquilibrent le compromis : GRU + Student-GARCH + GPD atteint le MFTR le plus élevé (0.0409) avec un ASR correct (7.23), tandis que GRU + Student-NGARCH + GPD (marqué “meilleur compromis”) accepte une variance plus forte (ANV 0.5538) pour des drawdowns contents et un ASR raisonnable (11.6).

## 4.6 Entraînement des modèles et suivi expérimental

Pour le déploiement deux modèles sont sélectionnés : GRU et XGBoost, ceci en raison d'une part de leur capacité intrinsèque<sup>11</sup>. Les modèles sont entraînés à partir d'un *dataset* consolidé (*dataset\_for\_training.parquet*). Après construction des log-rendements  $\log(P_t/P_{t-1}) \times 100$  et agrégation des volumes *volume\_\** avec un indicateur de sentiment *market\_sentiment*, la matrice d'entrée *X* est mise à l'échelle par un *RobustScaler*. Chaque crypto-actif possède sa propre cible *return\_<symbole>* - un modèle est ajusté par crypto.

**TABLE 4.28 –** Jeu de données et variables utilisées pour l'entraînement

Fichier	<i>dataset_for_training.parquet</i>
Cibles (y)	<i>return_&lt;crypto&gt;</i> (une cible par crypto)
Variables explicatives (X)	<i>return_*</i> , <i>volume_*</i> , <i>market_sentiment</i>
Transformation	log-rendements ( $\log(P_t/P_{t-1}) \times 100$ ), mise à l'échelle <i>RobustScaler</i> sur X

**GRU** Les entrées sont segmentées en séquences glissantes de longueur 48 avec horizon de prédiction 1 pas. Le réseau a l'architecture GRU → Dropout → Dense (1) avec régularisation  $\ell_1$  sur la couche GRU. L'optimisation est réalisée avec Adam et une perte MAE. Deux métriques sont suivies et utilisées sous forme « perte » pour la recherche de compromis : –MFTR (moyenne du

11. GRU capture bien les dépendances à long termes avec une bonne MCFD, et XGBoost pour sa capacité à absorber les dépendances non linéaires, sans compter ces bonnes performances sur la MFTR

signe cohérent avec le rendement) et  $-\text{MCFD}$  (taux d'orientation correcte). Un *early-stopping* sur la perte d'entraînement (patience = 3) prévient le sur-apprentissage.

**TABLE 4.29 – Hyperparamètres GRU**

Seq. (WINDOW)	Batch	Epochs	Hidden	Dropout	Perte	Métriques	Early-stop
48	1024	50	50	0.2	MAE	$-\text{MFTR}, -\text{MCFD}$	patience 3

**XGBoost** Un modèle par crypto est ajusté sur  $X$  non séquencé, avec objectif MSE. Les réglages retenus sont résumés dans le tableau 4.30.

**TABLE 4.30 – Hyperparamètres XGBoost**

Objective	Max depth	Learning rate	Subsample	Colsample_bytree	Boost rounds
reg:squarederror	3	0.01	0.8	0.8	300

**Suivi et export.** Tous les entraînements sont suivis dans MLflow (tracking\_uri locale file `./mlruns`, expériences `Crypto_GRU_Training` et `Crypto_XGBoost_Training`). Chaque modèle est exporté pour Triton Inference Server : les GRU au format *TensorFlow SavedModel* et les XGBoost via le backend *FIL* (`XGBoost.json`), avec génération du `config.pbtxt` adapté. Le scaler est sauvegardé pour la mise à l'échelle en production.

**TABLE 4.31 – Artefacts produits pour l'inférence**

Modèles GRU	<code>saved_models/gru/gru_&lt;crypto&gt;/1/model.savedmodel</code> + <code>config.pbtxt</code>
Modèles XGBoost	<code>saved_models/XGBoost/XGBoost_&lt;crypto&gt;/1/XGBoost.json</code> + <code>config.pbtxt</code>
Scaler	<code>saved_models/scalers/scaler.pkl</code>
Tracking	répertoires <code>./mlruns/</code> (paramètres, métriques, artefacts)

#### 4.6.1 Analyse de la MFTR par crypto (modèles GRU)

La **MFTR** (*Mean Forecasting Trading Return*) mesure le *rendement moyen de trading* que l'on obtiendrait en suivant les signaux du modèle. Soit un signal  $s_t \in \{-1, 0, +1\}$  déduit de la prévision  $\hat{r}_{t+1}$  (par exemple  $s_t = \text{sign}(\hat{r}_{t+1})$ ). Le rendement de trading d'une unité de position, sans frais, est

$$r_{t+1}^{\text{tr}} = s_t r_{t+1},$$

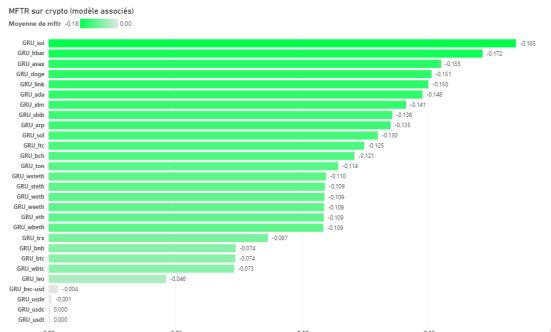
et la métrique s'écrit

$$\text{MFTR} = \frac{1}{N} \sum_{t=1}^N r_{t+1}^{\text{tr}}.$$

La MFTR est donc *dans les mêmes unités que les rendements* (par heure dans nos jeux de données) et traduit un gain moyen par pas de temps lorsque l'on suit mécaniquement le signal du modèle.

Lors de l'entraînement<sup>12</sup>, nous avons journalisé dans MLflow la quantité –MFTR pour raison de minimisation, d'où des valeurs affichées *négatives*. Pour revenir à l'échelle naturelle, il suffit d'inverser le signe :

$$\text{MFTR}_{\text{rapport}} = -\overline{\text{mftr}}_{\text{MLflow}}.$$



**FIGURE 4.39** – MFTR moyenne par crypto pour les modèles GRU. Les barres montrent –MFTR tel qu'enregistré dans MLflow ; plus la barre est à gauche, plus la MFTR vraie est élevée.

Le figure 4.39 fait ressortir une hétérogénéité par actif. Les cryptos comme SUI, HBAR, AVAX, DOGE, LINK ou ADA exhibent les meilleurs rendements de trading moyens (barres les plus négatives en  $-MFTR$ , donc MFTR vraie la plus forte). À l'inverse, les *stablecoins* (USDT, USDC, BSC-USD) restent proches de zéro, ce qui est cohérent avec leur faible directionnalité structurelle. On rappelle que la MFTR dépend de l'horizon de prévision, des règles de positionnement  $\{s_t\}$  et des coûts de transaction (non inclus ici).

## 4.7 Foreward testing

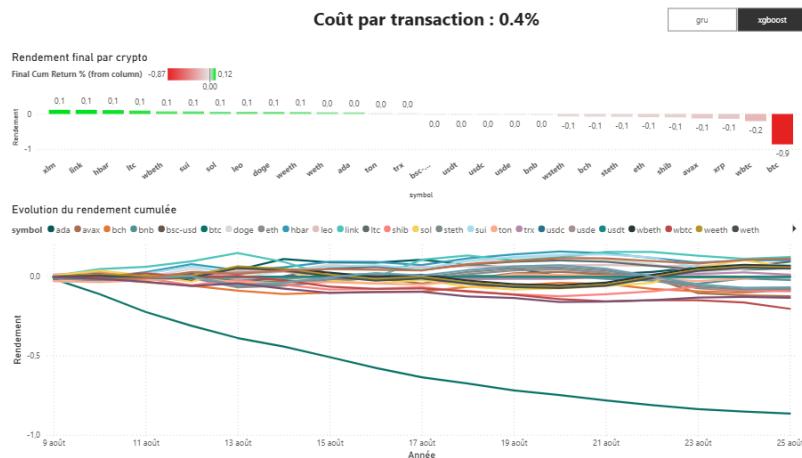
Nous avons mené une évaluation *prospective out-of-sample* sur la période du 9 au 25 août, avec un horizon de prévision d'un pas. À chaque date  $t$ , un signal de rendement était produit pour  $t+1$ , puis converti en position selon des règles fixées à l'avance. Les coûts de transaction ont été intégrés dans toutes les mesures, avec un taux uniforme de 0,4 % par transaction. Deux moteurs de signaux ont été testés en parallèle et sur le même univers de cryptomonnaies (incluant des *stables*) : un modèle séquentiel de type GRU et un modèle tabulaire de type XGBOOST. L'objectif n'était pas de « rejouer » l'historique hors ligne, mais d'observer, au fil de l'eau, la capacité de chaque moteur à générer une performance nette des coûts sur une fenêtre courte.

La figure composite Fig. 4.40 réunit, pour chacun des deux moteurs, deux vues complémentaires : en haut de chaque sous-figure, le rendement cumulé final par actif ; en bas, les trajectoires du rendement cumulé sur la période. Les résultats XGBOOST sont présentés en Fig. 4.40a, tandis que la figure figure 4.40b illustre le comportement du GRU.

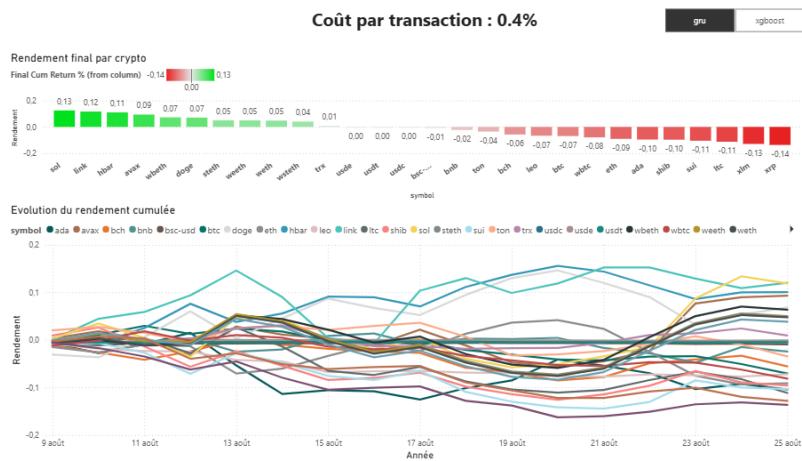
À la lecture de la figure figure 4.40a, On observe une échelle de valeur entre 0.1 et -0.9 la majorité des actifs reste proche de zéro, mais un titre enregistre une dérive négative marquée, proche de -0.87<sup>13</sup> sur la fenêtre.

12. lors de l'entraînement, pour des raisons inconnues les résultats d'entraînements des modèles XGBoost n'ont pas été journalisé

### 13 celle du bitcoin



(a) Forward testing (XGBoost) – rendement final par crypto (haut) et trajectoires cumulées (bas).



(b) Forward testing (GRU) – rendement final par crypto (haut) et trajectoires cumulées (bas).

**FIGURE 4.40** – Comparatif du forward testing (9–25 août, coût 0, 4 %/transaction) : (a) XGBoost en haut ; (b) GRU en bas.

La figure figure 4.40b montre, à l'inverse, le modèle gru montre une plage de valeur plus contrôlé compris entre 0.13% et -0.14%.

Le modèle GRU apparaît donc comme un modèle qui limite ses gains et contrôle moins ses pertes<sup>14</sup>, tandis que le modèle GRU stabilise ses gains et ses pertes dans un intervalle régulier.

En conclusion, à coûts et cadence identiques, pour une plus grande robustesse du GRU sur cette fenêtre.

## 4.8 Conclusion

Au terme de ce chapitre nous avons effectuer les différents taches évoquer au tout début de cette modélisation et ouvrons la porte pour la prochaine phase de notre projet : la gestion de risque et plus précisément la tarification des options.

14. avec une explosion à perte de stratégie sur le bitcoin

---

---

# CHAPITRE 5

---

## TARIFICATION DES OPTIONS

### Sommaire

---

5.1	Introduction . . . . .	78
5.2	Gestion de risque . . . . .	78
5.3	Comparaison NGARCH et Black–Scholes (maturité une semaine) . . . . .	80
5.4	Simulation du payoff et du P&L d'un call européen . . . . .	82
5.5	Effet du strike sur la probabilité d'obtenir un P&L positif . . . . .	83
5.6	Décroissance convexe de la prime $C(K)$ en fonction du strike . . . . .	83
5.7	Conclusion . . . . .	84

---

## 5.1 Introduction

La phase de modélisation issue d'une méthode hybride nous a donné un avantage phare. Car la propriété probabiliste nous permet de faire des simulations et calculer ainsi calculer des métriques de risques utiles aux acteurs du marché. Mais notre mission à travers ce chapitre concerne principale la tarification des options qui est un levier de couverture contre le risque.

## 5.2 Gestion de risque

Les options sur le marché financier en général sont des contrats qui donnent à leurs détenteurs le droit d'acheter ou de vendre à une date donnée et à un prix donné un actif sur le marché. Et pour un vendeur d'options, il est important de pouvoir trouver le juste prix, c'est-à-dire la prime que vont lui donner les clients, pour que même dans un cas défavorable ou bien dans la grande majorité des cas, ils puissent se couvrir contre le risque sous-jacent.

L'objectif est de valoriser des options européennes à partir d'une dynamique de prix qui articule une moyenne conditionnelle issue du GRU et une volatilité conditionnelle  $h_t$  décrite par un NGARCH. L'estimation des paramètres est réalisée sous la mesure réelle  $\mathbb{P}$ , avec des innovations à queues épaisses (loi splicée Student–GPD) pour mieux refléter l'empirique. Pour la tarification, nous adoptons l'approche classique de *valuation localement neutre au risque* (LRNVR) qui garde la dynamique de variance et remplace l'innovation par une gaussienne sous  $\mathbb{Q}$ . Cette hypothèse, courante en GARCH discret, permet de préserver la martingale du prix escompté  $\mathbb{Q}$  Duan1995 ; RITCHKEN et TREVOR 1999. Bien évidemment nous n'avons pas de problème en ce qui concerne la différence de distribution car le principe élément utilisé dans ce style de tarification est le risque sous-djacent (la volatilité) autrement dit l'ampleur des amplitudes est transférer. Cette méthode porte aussi un avantage sur des méthodes classiques tels que celle de black & scholes qui sont basé sur une volatilité constante.

Sous la probabilité risque neutre  $\mathbb{Q}$ , à pas  $\Delta t$ <sup>1</sup>, la dynamique des rendements logarithmiques retient le drift sans risque et la même variance conditionnelle qu'en estimation :

$$\log \frac{S_{t+1}}{S_t} = (r - q) \Delta t - \frac{1}{2} h_t \Delta t + \sqrt{h_t \Delta t} e_{t+1}, \quad e_{t+1} \sim \mathcal{N}(0, 1) \text{ i.i.d. sous } \mathbb{Q}, \quad (5.1)$$

tandis que la variance suit le NGARCH estimé (paramètres figés aux valeurs de tableau 4.25) :

$$h_{t+1} = \omega + \beta h_t + \alpha h_t (e_{t+1} - \theta)^2. \quad (5.2)$$

La moyenne  $\mu_t$  prédite par le GRU sert à fixer l'état initial sous  $\mathbb{P}$  et à contrôler la cohérence des simulations historiques ; sous  $\mathbb{Q}$ , le drift pertinent est  $(r - q)^2$  comme dans (5.1).

Le prix d'un call européen de strike  $K$  et maturité  $T$  s'obtient par espérance neutre au risque :

$$C_0 = e^{-rT} \mathbb{E}_{\mathbb{Q}} [(S_T - K)^+], \quad (5.3)$$

1. Ici  $\Delta t$  est égale à 1 (une heure)

2. le drift est égale à  $r_f$  dans notre cas le taux de rendement sans risque et  $q = 0$  (car absence de dividende)

que l'on approxime par Monte Carlo en générant  $M$  trajectoires de  $(S_t, h_t)$  selon (5.1)–(5.2) à partir de  $S_0$  et  $h_0$  estimés :

$$\hat{C}_0 = e^{-rT} \frac{1}{M} \sum_{m=1}^M (S_T^{(m)} - K)^+,$$

Où  $\tilde{S}_t$  (prix actualisé) est une martingale sous la probabilité risque neutre.

## Vérification de la propriété de martingale

Considérons la dynamique du prix de l'actif sous la probabilité risque-neutre  $\mathbb{Q}$ , pour un pas  $\Delta t$  :

$$\log \frac{S_{t+\Delta t}}{S_t} = (r - q)\Delta t - \frac{1}{2}h_t\Delta t + \sqrt{h_t\Delta t}e_{t+\Delta t}, \quad e_{t+\Delta t} \sim \mathcal{N}(0, 1) \text{ i.i.d. sous } \mathbb{Q}. \quad (5.4)$$

En réécrivant, on obtient :

$$S_{t+\Delta t} = S_t \exp\left((r - q)\Delta t - \frac{1}{2}h_t\Delta t + \sqrt{h_t\Delta t}e_{t+\Delta t}\right). \quad (5.5)$$

Conditionnellement à l'information disponible  $\mathcal{F}_t$ , la variance  $h_t$  est connue et  $e_{t+\Delta t}$  est une variable gaussienne indépendante. On applique alors la fonction génératrice des moments d'une loi normale : si  $Z \sim \mathcal{N}(0, 1)$ , on a  $\mathbb{E}[e^{aZ}] = e^{\frac{1}{2}a^2}$ . Ainsi :

$$\mathbb{E}_{\mathbb{Q}}[S_{t+\Delta t} | \mathcal{F}_t] = S_t \exp\left((r - q)\Delta t - \frac{1}{2}h_t\Delta t\right) \mathbb{E}_{\mathbb{Q}}\left[\exp\left(\sqrt{h_t\Delta t}e_{t+\Delta t}\right) | \mathcal{F}_t\right] \quad (5.6)$$

$$= S_t \exp\left((r - q)\Delta t - \frac{1}{2}h_t\Delta t\right) \exp\left(\frac{1}{2}h_t\Delta t\right) \quad (5.7)$$

$$= S_t e^{(r-q)\Delta t}. \quad (5.8)$$

On introduit maintenant le prix actualisé tenant compte du dividende  $q$  :

$$\tilde{S}_t = e^{-(r-q)t} S_t. \quad (5.9)$$

Alors :

$$\mathbb{E}_{\mathbb{Q}}[\tilde{S}_{t+\Delta t} | \mathcal{F}_t] = e^{-(r-q)(t+\Delta t)} \mathbb{E}_{\mathbb{Q}}[S_{t+\Delta t} | \mathcal{F}_t] \quad (5.10)$$

$$= e^{-(r-q)(t+\Delta t)} S_t e^{(r-q)\Delta t} \quad (5.11)$$

$$= e^{-(r-q)t} S_t \quad (5.12)$$

$$= \tilde{S}_t. \quad (5.13)$$

On conclut donc que  $(\tilde{S}_t)_{t \geq 0}$  est une **martingale sous  $\mathbb{Q}$** . Cela confirme que la dynamique proposée est bien conforme au principe d'absence d'arbitrage.

Pour une mise en pratique de notre algorithme, nous allons simuler deux contrats, un contrat A et un contrat B.

### 5.3 Comparaison NGARCH et Black–Scholes (maturité une semaine)

Lors de la première version de nos simulations, la variance  $h_t$  issue des modèles GARCH/N-GARCH avait été interprétée comme une variance de log-rendements en *décimal*, alors que le modèle avait été estimé sur des **log-rendements multipliés par 100** (en “%”). Or, si

$$r_{t+1}^{(\%)} = 100 \cdot \log\left(\frac{S_{t+1}}{S_t}\right) = \mu_t + \sqrt{h_t} Z_{t+1}, \quad Z_{t+1} \sim \mathcal{N}(0, 1),$$

alors la propagation du prix doit utiliser la variance *décimale*

$$h_t^{(S)} = \frac{h_t}{100^2}.$$

Sous la mesure risque-neutre  $\mathbb{Q}$ , on simule donc

$$\log\left(\frac{S_{t+1}}{S_t}\right) = (r - q)\Delta t - \frac{1}{2}h_t^{(S)} + \sqrt{h_t^{(S)}} Z_{t+1},$$

tandis que la récurrence de variance (GARCH/NGARCH/GARCH-X) reste inchangée dans l’unité d’estimation. Après correction ( $h_t^{(S)} = h_t/10,000$ ), les primes redeviennent d’un ordre de grandeur réaliste et les tests de cohérence (martingale, limite Black–Scholes) sont satisfaits.

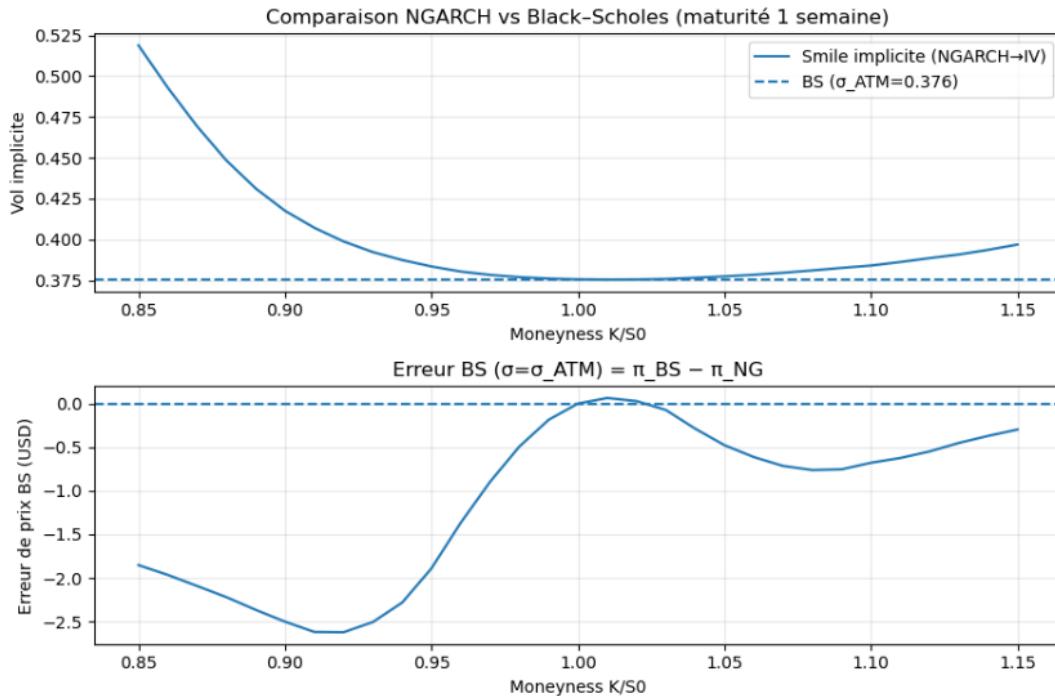
Sous l’hypothèse d’un rendement log-normal avec une volatilité constante, le prix d’un call européen s’écrit

$$C^{\text{BS}}(S_0, K, T, r, q; \sigma) = S_0 e^{-qT} \Phi(d_1) - K e^{-rT} \Phi(d_2), \quad d_{1,2} = \frac{\ln(S_0/K) + (r - q \pm \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}},$$

où  $\Phi$  désigne la fonction de répartition de la loi normale standard. Ce résultat classique, dû à BLACK et SCHOLES (1973) et approfondi par MERTON (1973), sert d’étoile pour confronter toute approche dynamique plus riche.

Nous travaillons à l’horizon d’une semaine ( $T = 168$  heures) en posant  $r = q = 0$ , ce qui revient à confondre ATM-spot et ATM-forward à ce pas de temps. Le modèle NGARCH a été estimé sur des log-rendements exprimés en pourcentage ; dans la simulation sous  $\mathbb{Q}$ , la variance horaire utilisée pour propager le log-prix est donc  $h_t/100^2$ , tandis que la récurrence de variance reste dans l’unité d’estimation. La comparaison procède ainsi : pour une grille de strikes  $K$  centrée autour de  $S_0$ , on simule  $S_T$  sous NGARCH puis on calcule le prix  $\pi_{\text{NG}}(K) = e^{-rT} \mathbb{E}^{\mathbb{Q}}[(S_T - K)^+]$ . À chaque strike, on inverse ensuite la formule de Black–Scholes afin d’obtenir la volatilité implicite  $\sigma_{\text{impl}}(K)$  telle que  $C^{\text{BS}}(S_0, K, T, r, q; \sigma_{\text{impl}}(K)) = \pi_{\text{NG}}(K)$ . En parallèle, on fixe une unique volatilité  $\sigma_{\text{ATM}}$  en égalant le prix BS au prix NGARCH au strike ATM  $K_{\text{ATM}} \approx S_0$ . Cette volatilité unique sert à produire une courbe BS « plate » que l’on confronte aux prix NGARCH. L’écart résiduel est mesuré par l’erreur  $E_{\text{BS}}(K) = C^{\text{BS}}(S_0, K, T, r, q; \sigma_{\text{ATM}}) - \pi_{\text{NG}}(K)$ .

Le panneau supérieur de la figure 5.1 met en évidence une courbe de volatilité implicite qui s’incline à gauche et se cambre légèrement à droite. Autrement dit, les vols sont plus



**FIGURE 5.1** – Haut : smile implicite  $\sigma_{\text{impl}}(K)$  obtenu en inversant les prix NGARCH et, en pointillé, Black–Scholes calibré à l’ATM ( $\sigma = \sigma_{\text{ATM}}$ ). Bas : erreur de prix  $E_{\text{BS}}(K) = \pi_{\text{BS}}(K; \sigma_{\text{ATM}}) - \pi_{\text{NG}}(K)$ .

élevées pour des strikes bas (skew négatif) et remontent modestement dans l’aile droite, ce qui révèle une distribution risque-neutre asymétrique et dotée de queues plus épaisses que la log-normale. La ligne horizontale issue de Black–Scholes calibré à l’ATM reste, elle, par construction, approximativement plate : un modèle à volatilité constante ne génère ni inclinaison ni smile. Le panneau inférieur confirme ce diagnostic sous la forme d’une erreur de prix qui s’annule au centre, puis devient négative pour les strikes dans la monnaie (BS sous-estime la valeur des calls ITM) et redevient négative plus loin dans l’OTM. Entre les deux, une légère zone positive peut apparaître, signe d’une surévaluation locale lorsque la courbure implicite de NGARCH croise la courbe plate de BS. L’ensemble montre qu’une seule volatilité ne suffit pas à suivre la variation des prix avec le strike, alors que la dynamique NGARCH la reproduit naturellement.

En dehors de tout modèle particulier, la forme de  $K \mapsto C(K)$  est gouvernée par deux identités « modèle-libres » : la dérivée première vaut  $-\partial C / \partial K = e^{-rT} \mathbb{Q}(S_T > K)$ , ce qui relie la pente au prix d’un digital, et la dérivée seconde vaut  $\partial^2 C / \partial K^2 = e^{-rT} f_Q(K)$ , c’est-à-dire la densité risque-neutre au strike (à l’escompte près). Ces relations, établies de manière fondatrice par BREEDEN et LITZENBERGER (1978) et reprises dans le *CCBS Handbook* de la Bank of England, expliquent que toute asymétrie ou épaisseur de queue dans  $f_Q$  se reflète en inclinaison et en convexité de la courbe d’IV. NGARCH, par sa volatilité conditionnelle et son terme d’asymétrie, induit précisément ce type de densité ; Black–Scholes, avec une volatilité unique, ne le peut pas.

Calibré au centre, Black–Scholes fournit une approximation correcte à l’ATM, mais décroche dès que l’on s’éloigne en strike. La version NGARCH, en revanche, restitue le skew et le léger smile observés sur l’ensemble de la grille, avec des erreurs de prix plus faibles hors ATM. Dans un contexte où la surface implicite n’est pas plate, la méthode NGARCH apparaît ainsi préférable

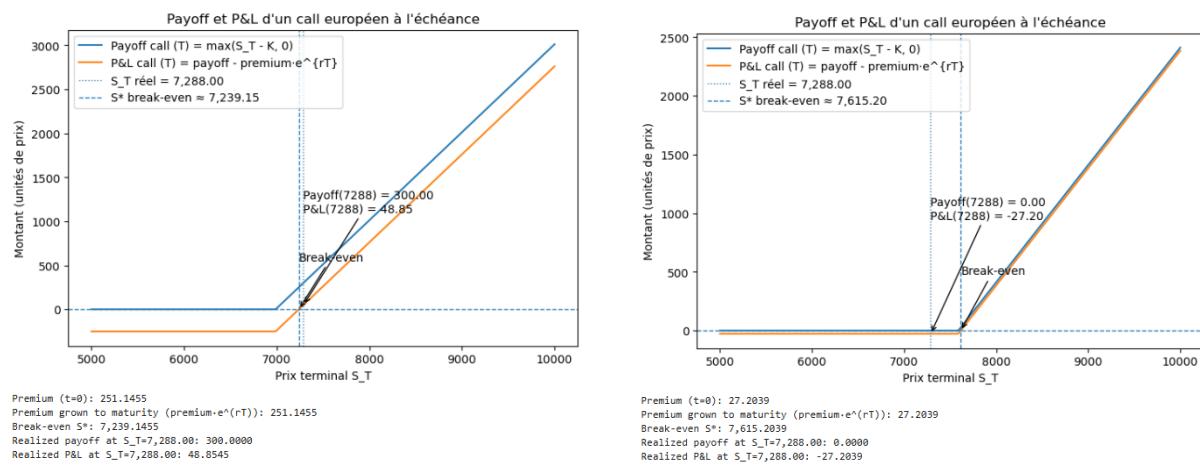
pour obtenir des prix cohérents sur la maturité considérée, sans recourir à une collection de volatilités ad hoc.

## 5.4 Simulation du payoff et du P&L d'un call européen

Nous illustrons à l'échéance  $T = 168$  heures (une semaine,  $r = q = 0$ ) le *payoff* d'un call max( $S_T - K, 0$ ) (courbe bleue) et le *profit & perte* (P&L) de l'acheteur (courbe orange),

$$\text{P\&L}(T) = \max(S_T - K, 0) - C_0 e^{rT}.$$

La verticale de gauche marque le strike  $K$ , celle de droite le point mort  $S^* = K + C_0 e^{rT}$ ; l'horizontale  $y = 0$  sépare gains et pertes. Dans les deux exemples, le prix terminal observé vaut  $S_T = 7288$  (avec  $S_0 = 7156,01$ ).



(a) Call dans la monnaie : Contrat A ( $K = 6988$ )

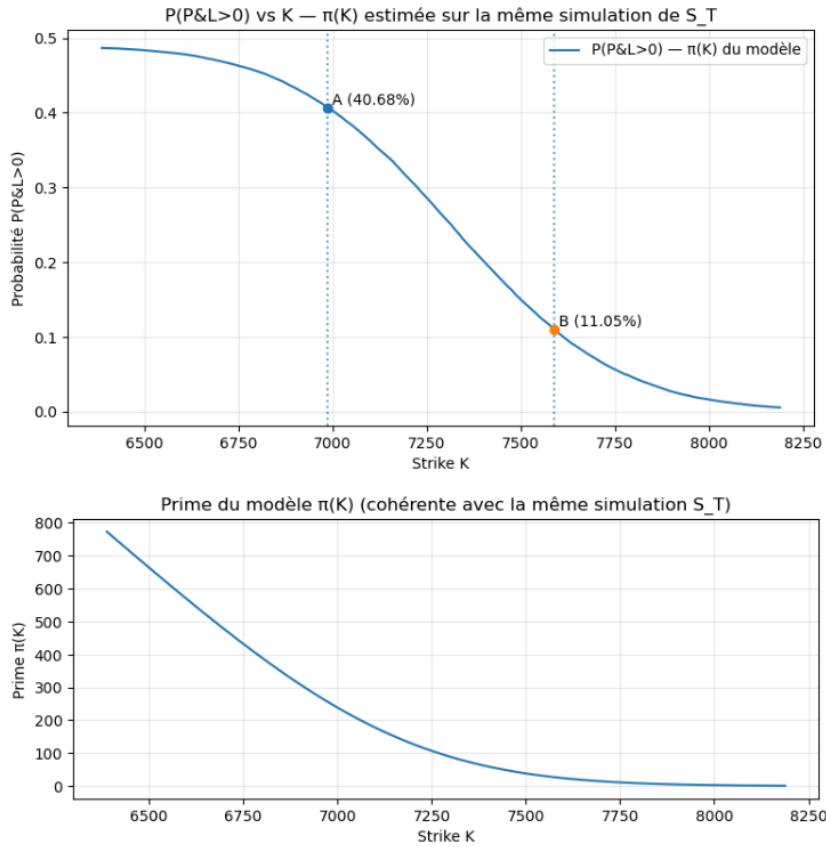
(b) Call hors de la monnaie : Contrat B ( $K = 7588$ )

**FIGURE 5.2 – Payoff (bleu) et P&L (orange) d'un call européen à  $T$**

**Contrat A (ITM).** On obtient une prime  $C_0 \simeq 251,15$  et un point mort  $S^* \simeq 7239,15$ . Comme  $S_T = 7288 > S^*$ , l'option finit *dans la monnaie* : le payoff vaut  $S_T - K = 300$  et le P&L de l'acheteur est  $\approx 300 - 251,15 = +48,85$ .

**Contrat B (OTM).** La prime est beaucoup plus faible,  $C_0 \simeq 27,20$ , d'où un point mort  $S^* \simeq 7615,20$ . Ici  $S_T < K$  : l'option expire sans valeur et l'acheteur perd exactement sa prime capitalisée, soit un P&L  $\approx -27,20$ .

Géométriquement, la courbe *P&L* (orange) est une translation vers le bas de la courbe de payoff (bleu) d'un montant  $C_0 e^{rT}$ ; sa pente est nulle pour  $S_T < K$  et égale à 1 pour  $S_T \geq K$ . Les valeurs numériques peuvent varier légèrement d'une simulation Monte Carlo à l'autre (erreur statistique, graine).



**FIGURE 5.3 – Probabilité  $\mathbb{P}(P\&L > 0)$  (haut) et prime  $\pi(K)$  (bas), en cohérence avec la *même* simulation de  $S_T$ . Les points A et B marquent  $K = 6988$  et  $K = 7588$ .**

## 5.5 Effet du strike sur la probabilité d'obtenir un P&L positif

La courbe supérieure montre que  $\mathbb{P}(P\&L > 0)$  est **décroissante** en  $K$  lorsque la prime utilisée est celle du modèle  $\pi(K)$  calculée sur la même simulation de  $S_T$  : le point mort  $S^*(K) = K + \pi(K)$  se décale vers la droite quand  $K$  augmente, plus vite que la baisse de  $\pi(K)$  (courbe inférieure). Sur la figure, on lit par exemple :

- **A** ( $K = 6988$ ) :  $\pi(K) \approx 246,4$  et  $\mathbb{P}(P\&L > 0) \approx 40,7\%$  ;
- **B** ( $K = 7588$ ) :  $\pi(K) \approx 25,5$  et  $\mathbb{P}(P\&L > 0) \approx 11,0\%$ .

Autrement dit, *abaisser le strike* accroît la probabilité d'un P&L positif, mais au prix d'une *prime plus élevée* : c'est le compromis classique entre risque et coût d'entrée. Les deux graphiques confirment en outre les propriétés attendues :  $\pi(K)$  est décroissante et convexe en  $K$ , et la probabilité associée diminue avec  $K$  pour une même distribution de  $S_T$ .

## 5.6 Décroissance convexe de la prime $C(K)$ en fonction du strike

Pour une maturité  $T$  fixée, sous la mesure risque-neutre, le prix d'un call s'écrit

$$C(K) = e^{-rT} \mathbb{E}^{\mathbb{Q}}[(S_T - K)^+] = e^{-rT} \int_K^{\infty} (s - K) f_Q(s) ds, \quad (5.14)$$

où  $f_Q$  désigne la densité (risque-neutre) du prix terminal  $S_T$ .<sup>3</sup>

En dérivant (5.14) par rapport à  $K$  (interversion sous des conditions usuelles),

$$\frac{\partial C}{\partial K} = -e^{-rT} \int_K^\infty f_Q(s) ds = -e^{-rT} \mathbb{Q}(S_T > K) \leq 0, \quad (5.15)$$

$$\frac{\partial^2 C}{\partial K^2} = e^{-rT} f_Q(K) \geq 0. \quad (5.16)$$

L'identité (5.15) montre que le prix du call est *décroissant* en  $K$ ; (5.16) identifie sa *courbure* à la densité risque-neutre (à l'escompte près), d'où la *convexité*. Autrement dit,  $-\partial C/\partial K$  coïncide avec le prix d'un digital (probabilité d'être dans la monnaie), tandis que  $\partial^2 C/\partial K^2$  coïncide avec la densité implicite en  $K$ .

Ces deux égalités sont les formes « modèle-libres » du résultat de BREEDEN et LITZENBERGER (1978) : la seconde dérivée de la fonction de prix du call par rapport au strike engendre le prix (densité) des revendications élémentaires. Le *CCBS Handbook* en donne une démonstration opérationnelle : à partir de (5.14), ils obtiennent

$$\frac{\partial C}{\partial K} = -e^{-rT} \int_K^\infty f_Q(x) dx, \quad \frac{\partial^2 C}{\partial K^2} = e^{-rT} f_Q(K),$$

puis en déduisent la récupération de  $f_Q$  à partir de la courbure des prix.<sup>4</sup>

Enfin, même quand la densité n'admet pas de version  $f_Q$  lisse, l'argument d'absence d'arbitrage par « butterfly » impose que  $K \mapsto C(K)$  soit *convexe* et *non croissante*; sinon, un portefeuille butterfly offrirait un payoff non négatif pour un coût négatif.<sup>5</sup>

## 5.7 Conclusion

Arrivée au terme de ce chapitre, il était question pour nous de tarifier les options<sup>6</sup> dans un cadre de forte volatilité. En utilisant une base sur le modèle NGARCH nous avons pu obtenir une tarification plus fiable comparément à des méthodes tels que Black & scholes . Nous avons simulé le payoff d'un call européen, analyser l'effet du strike sur la probabilité d'obtenir un P&L positif et démontrer la décroissance convexe de la prime  $C(K)$  en fonction du strike.

---

3. Cette écriture figure mot pour mot dans le *CCBS Handbook* de la Bank of England, puis ils en déduisent les dérivées en  $K$  (p. 12).

4. Résultat issus de Bank of England, *Deriving option-implied probability densities for foreign exchange markets*, p. 12 (équations affichées pour  $\partial C/\partial K$  et  $\partial^2 C/\partial K^2$ ). Chez BREEDEN et LITZENBERGER (1978), on le retrouve dans la Section II et l'égalité où  $c_{XX}$  (la seconde dérivée du call par rapport au strike  $X$ ) donne le prix de la claim élémentaire ; cf. p. 10 du PDF, eq. (5), et la discussion encadrant l'équation (3) p. 7.

5. Preuve par arbitrage dans des notes de cours standard (par ex. « Convexity of call-option prices. Butterfly spreads », Univ. of Texas)

6. Call dans notre cas

---

## CONCLUSION GÉNÉRALE

Ce projet avait pour ambition de mieux comprendre et d'outiller la décision sur un marché singulier : celui des cryptomonnaies, caractérisé par une forte instabilité, des queues épaisse et des ruptures fréquentes de régime. La problématique a été abordée de bout en bout, depuis la collecte et l'alignement des données jusqu'à la mise à disposition d'analyses opérationnelles dans une application web. L'objectif initial — améliorer la prévision des prix et de la volatilité tout en proposant des instruments concrets de gestion du risque — a guidé l'ensemble des choix méthodologiques et techniques.

La contribution scientifique et technique se décline à trois niveaux complémentaires. D'abord, une démarche de modélisation en deux étages a été mise en place : estimation de la moyenne conditionnelle avec des modèles économétriques et apprenants (ARIMA, XGBoost, GRU), puis modélisation de l'hétérosécédasticité résiduelle à l'aide de familles GARCH (GARCH, NGARCH, GARCH-X). Ce couplage permet d'isoler les composantes systématiques des rendements avant de caractériser la dynamique de la variance, ce qui améliore l'interprétabilité et la robustesse. Ensuite, des expériences de validation ont été conduites dans un cadre cohérent avec l'usage réel : séparation temporelle stricte, prise en compte de coûts de transaction et comparaison des modèles sur des métriques multiples (erreur de prédiction, classement et performance de portefeuille). Enfin, une couche applicative a été développée pour relier les sorties des modèles aux besoins des utilisateurs : tableau de bord marché, simulation et gestion de portefeuilles, analyses de risque et scénarios.

Les résultats obtenus confirment l'intérêt de l'approche hybride. Sur la moyenne, les modèles apprenants capturent mieux les non-linéarités locales, tandis que les modèles économétriques facilitent l'analyse diagnostique et l'explication des signaux. Sur la variance, l'introduction d'une dynamique non linéaire via NGARCH permet de reproduire plus finement les épisodes de volatilité marquée. La comparaison avec Black–Scholes a mis en évidence les limites d'une volatilité constante pour tarifer en contexte crypto ; le recours à une volatilité conditionnelle fournit des prix plus cohérents avec les sourires implicites observés et les distributions empiriques des rendements. D'un point de vue « produit », la plateforme livre un enchaînement continu des données vers la décision : ingestion périodique, prévisions horaires, mesures de risque et visualisations prêtées à l'emploi.

Le cheminement n'a pas été exempt de difficultés. La première tient à l'instationnarité des séries : changements de régimes, sauts et dépendances de longue mémoire compliquent la généralisation hors échantillon. La seconde concerne la qualité et l'alignement des sources : synchroniser prix, volumes et signaux d'actualité nécessite des règles robustes de préparation. La troisième relève de l'industrialisation : orchestrer entraînement, suivi expérimental, exposition d'inférences et déploiement à coût raisonnable impose des compromis entre sophistication des modèles, simplicité d'exploitation et contraintes de ressources. Ces obstacles ont conduit à privilégier des pipelines explicites, des tests systématiques et une architecture de service mesurée.

Au-delà des livrables, ce travail a été l'occasion d'un véritable progrès personnel. Sur le plan scientifique, il a consolidé la pratique des modèles de séries temporelles financières, le dialogue entre statistique et apprentissage, et l'évaluation rigoureuse centrée « métier ». Sur le plan d'ingénierie, il a renforcé les compétences en conception d'API, gestion d'environnements, conteneurisation et exposition d'inférence.

# ANNEXE A

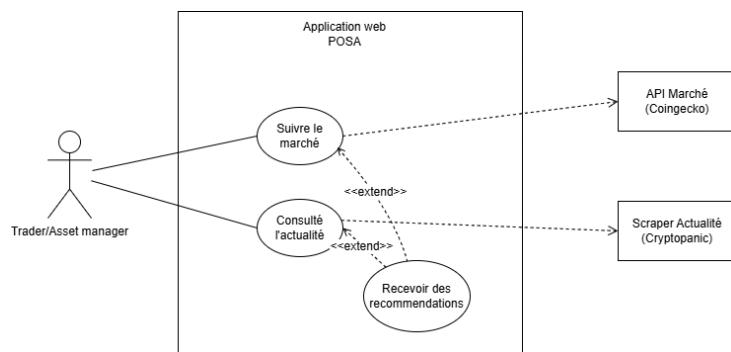
---

## ANALYSE DES BESOINS

### Introduction

Dans le cadre du développement de notre solution, il est apparu essentiel de préciser avec rigueur les besoins qui en découlent. En effet, toute conception logicielle et technique s'appuie d'abord sur l'identification claire des problèmes rencontrés, avant de traduire ces derniers en besoins fonctionnels et non fonctionnels. Ce chapitre présente donc une lecture structurée des besoins issus des trois problèmes majeurs de notre projet, à savoir la prise de décision dans un environnement incertain, la difficulté à modéliser les dynamiques de marché, et la gestion des risques associée à l'optimisation des stratégies. L'objectif est de mettre en évidence les attentes auxquelles la solution doit répondre, tant sur le plan de la fonctionnalité que sur celui de la performance et de la fiabilité.

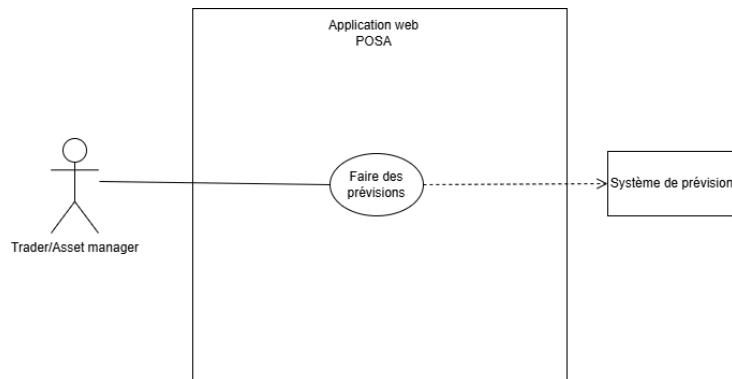
### Corps du chapitre



**FIGURE A.1 –** prise de décision dans un environnement incertains - diagramme cas d'utilisation

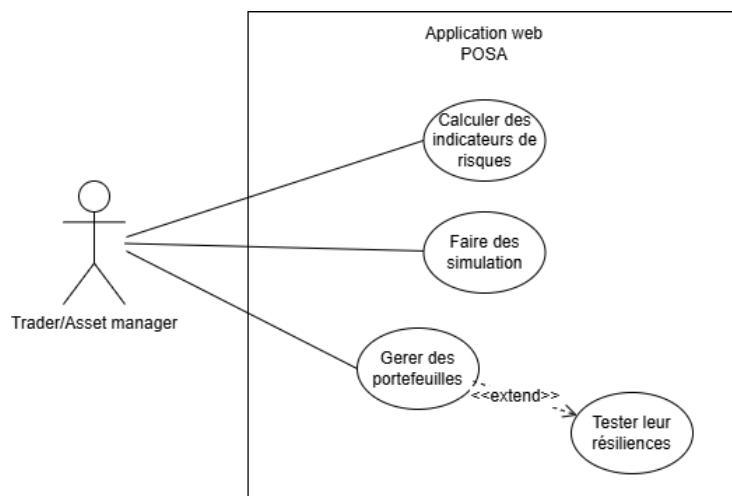
La première problématique identifiée concerne la prise de décision dans un environnement marqué par une forte incertitude(figure A.1). Les marchés financiers et, plus particulièrement, les marchés de crypto-actifs se caractérisent par une variabilité élevée et des comportements souvent

imprévisibles. Dans ce contexte, trois besoins se sont affirmés. Il s'agit tout d'abord de disposer d'outils permettant de suivre en continu le comportement du marché et les relations entre actifs. Il est également nécessaire de suivre l'actualité en temps réel, puisque les annonces institutionnelles, les évolutions réglementaires ou encore les tendances médiatiques influencent directement les dynamiques de prix. Enfin, la prise de décision doit être accompagnée de recommandations ou de conseils automatisés, permettant d'éclairer l'utilisateur face à la complexité de l'information disponible.



**FIGURE A.2 – modéliser les dynamiques de marché - diagramme cas d'utilisation**

Le second problème porte sur la difficulté à modéliser les dynamiques de marché (figure A.2). La complexité des séries temporelles financières, marquées par la volatilité, la non-stationnarité et les phénomènes extrêmes, exige des modèles capables de capturer ces comportements. Le besoin fonctionnel principal s'exprime donc en termes de prévision : il s'agit de pouvoir anticiper l'évolution du marché à court terme, en combinant des approches économétriques et des techniques d'apprentissage profond. Cette dimension prédictive constitue le cœur de la valeur ajoutée de la solution, car elle fournit un éclairage prospectif indispensable à l'utilisateur.



**FIGURE A.3 – gestion de risque et optimisation des stratégies - diagramme cas d'utilisation**

Le troisième problème concerne la gestion des risques et l'optimisation des stratégies (figure A.3). Dans ce domaine, les besoins sont multiples. Il est nécessaire de calculer différents indicateurs de risque, tant au niveau des actifs individuels que des portefeuilles agrégés. Ces mesures permettent d'évaluer l'exposition aux pertes potentielles et de comparer différentes stratégies. Par ailleurs,

la solution doit offrir des outils de simulation et de gestion de portefeuille, permettant de tester différentes allocations et d'évaluer leurs performances. L'optimisation des stratégies constitue ainsi un levier essentiel pour transformer l'information en actions concrètes.

Au-delà de ces besoins fonctionnels, certains besoins non fonctionnels sont venus préciser les exigences techniques. La question de l'authentification et de la sécurité s'est traduite par l'intégration d'un système d'authentification via Google et l'utilisation de jetons JWT pour sécuriser les communications. L'expérience utilisateur a également été placée au centre des préoccupations, avec une interface fluide et intuitive facilitant la prise en main. Sur le plan computationnel, la parcimonie s'est imposée comme un impératif. L'inférence des modèles, effectuée via Triton Server, repose sur un chargement dynamique : les modèles sont activés uniquement au moment des prédictions, puis déchargés afin d'économiser les ressources. Enfin, l'usage de bibliothèques optimisées telles que Pandas et NumPy, exploitant les performances du langage C en arrière-plan, a permis de garantir des temps de calcul compatibles avec les contraintes de production.

## Conclusion

L'analyse des besoins met en évidence la cohérence entre les problèmes identifiés et les solutions envisagées. Les besoins fonctionnels définissent le cœur de la valeur ajoutée attendue : accompagner la décision, prévoir les dynamiques de marché et gérer efficacement les risques. Les besoins non fonctionnels, quant à eux, garantissent la robustesse, la sécurité et l'efficacité de l'implémentation. Cette étape constitue ainsi un socle indispensable pour la conception de l'architecture logicielle et technique, qui sera détaillée dans le chapitre suivant.

## ANNEXE B

# ARCHITECTURE GLOBALE DE LA SOLUTION

## B.1 Introduction

Après avoir identifié les besoins fonctionnels et non fonctionnels, il est nécessaire de présenter la structure logicielle et technique de la solution développée. L'objectif de ce chapitre est de donner une vision d'ensemble de l'architecture, en mettant en évidence les principales briques qui composent le système et la manière dont elles interagissent. Deux représentations complémentaires sont proposées : une vue globale des composants et une vue détaillant le flux des données et le processus de prédiction.

## B.2 Vue globale de l'architecture

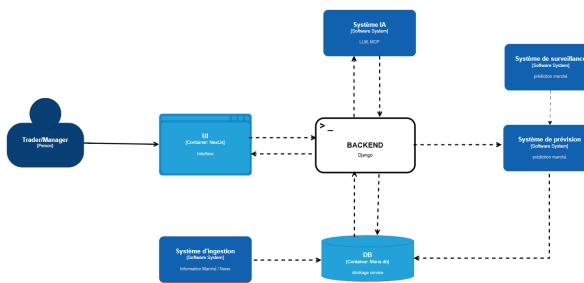
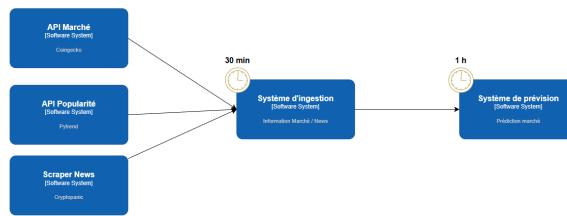


FIGURE B.1 – architecture globale - Vue A : Contexte

La première vue figure B.1 illustre les composants principaux de la plateforme et leurs interactions. L'utilisateur, qu'il s'agisse d'un trader ou d'un gestionnaire, accède à l'application à travers l'interface développée avec Next.js. Cette interface communique directement avec le *backend*, implanté sous Django, qui centralise la logique applicative et sert de point d'orchestration entre les différents modules.

Les services d'ingestion, organisés en conteneurs dédiés, collectent les données de marché et les actualités à intervalles réguliers, avant de les stocker dans une base de données relationnelle MariaDB. D'autres services prennent en charge la génération de prédictions ainsi que la surveillance du marché, en interaction constante avec le *backend*. Ce dernier est également relié à des modules d'intelligence artificielle, permettant d'intégrer des capacités de traitement avancées telles que les modèles séquentiels ou les approches par apprentissage automatique. L'ensemble forme une architecture modulaire où chaque composant assume une responsabilité bien définie, facilitant ainsi la maintenance et l'évolution du système.

### B.3 Flux de données et processus de prédition



**FIGURE B.2 – architecture globale - Vue A :Flux de données et ML**

La seconde vue figure B.2 met en évidence le parcours suivi par les données, depuis leur collecte jusqu'à leur restitution à l'utilisateur. Les sources externes incluent des API de marché comme CoinGecko, des indicateurs de popularité via PyTrend, ainsi qu'un service de récupération d'actualités tel que CryptoPanic. Ces flux sont agrégés toutes les trente minutes par le système d'ingestion, qui assure la normalisation et la mise en cohérence des informations collectées.

À chaque heure, les données consolidées sont transmises au module de prédition, chargé de produire des anticipations sur l'évolution du marché. Les résultats sont ensuite enregistrés dans la base et mis à disposition à travers l'API du *backend*. L'interface utilisateur peut ainsi proposer des visualisations en temps réel, combinant données de marché, analyses de sentiment et prévisions, afin d'apporter une aide concrète à la prise de décision. Cette organisation en flux périodiques garantit un équilibre entre fraîcheur des informations et maîtrise des coûts computationnels.

### B.4 Conclusion

Ces deux représentations permettent d'appréhender l'architecture dans sa globalité et de comprendre le cheminement des données à travers le système. La vue globale met en lumière les briques logicielles et leur articulation, tandis que la vue orientée données souligne le rôle des processus d'ingestion et de prédition. Les aspects techniques détaillés, incluant la structure des données, les algorithmes de modélisation et les mécanismes de déploiement, seront approfondis dans le chapitre consacré à la couche technique.

## ANNEXE C

# PARCOURS UTILISATEUR ET INTERFACES

## Introduction

Au-delà de la conception technique et des besoins fonctionnels, la valeur d'un système se mesure à la fluidité de l'expérience qu'il propose à ses utilisateurs. Dans notre cas, l'application cible principalement les traders et les gestionnaires de portefeuille, et leur efficacité dépend de la simplicité avec laquelle ils accèdent aux données de marché, créent des portefeuilles, évaluent les risques et interagissent avec l'assistant. Ce chapitre présente donc le parcours utilisateur type, depuis son arrivée sur la plateforme jusqu'à l'utilisation des fonctionnalités avancées, en s'appuyant à la fois sur un diagramme de flux d'utilisation et sur des captures d'écran des principales interfaces.

## Parcours utilisateur global

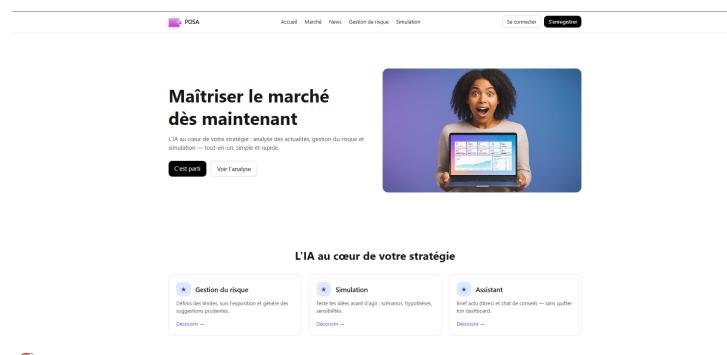
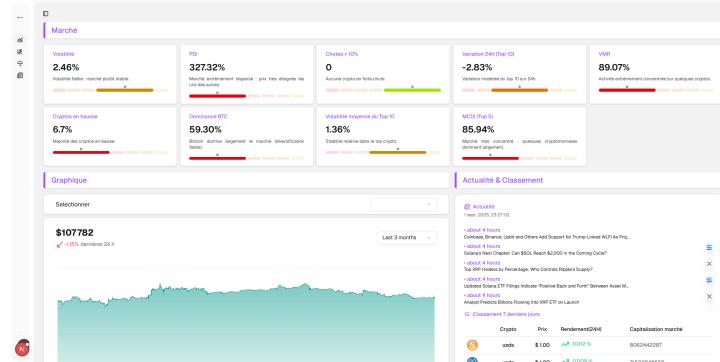


FIGURE C.1 – Page d'accueil : Capture d'écran

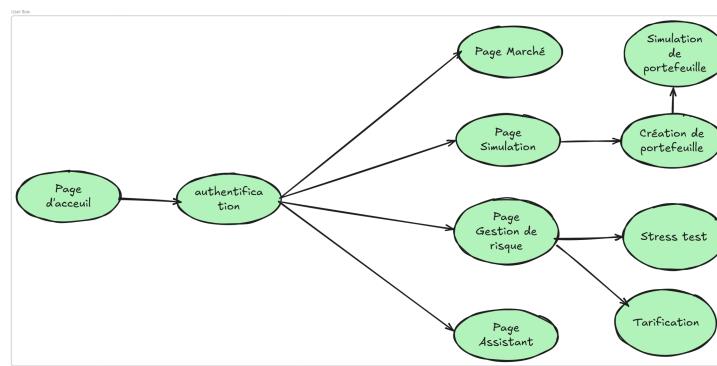
L'utilisateur découvre d'abord la plateforme via une page d'accueil (figure C.1), qui met en avant les principales fonctionnalités et propose une action claire : se connecter ou s'enregistrer.

Une fois authentifié, il est redirigé vers la page *Marché*, qui constitue le point d'entrée principal du système.



**FIGURE C.2 – Page Marché : Capture d'écran**

Sur cette page, l'utilisateur accède à un tableau de bord présentant les indicateurs clés de marché, les évolutions des cryptomonnaies, les prévisions horaires et les actualités (figure C.2). Il peut ensuite naviguer vers la page *Simulation*, où il dispose d'outils avancés pour analyser le marché et gérer ses portefeuilles. Le diagramme de flux utilisateur présenté en figure C.3 illustre ce cheminement global, depuis l'authentification jusqu'aux différentes sections fonctionnelles.



**FIGURE C.3 – Parcours utilisateur**

## Interfaces principales

### Page Marché

La page marché (figure C.2) concentre les informations essentielles permettant de saisir rapidement l'état du marché. Les indicateurs tels que la volatilité, la dispersion, la dominance du Bitcoin ou encore la variation quotidienne du Top 10 donnent un aperçu instantané de la situation. L'utilisateur peut consulter les actualités les plus récentes ainsi que le classement des cryptomonnaies par capitalisation, prix ou rendement. Enfin, un module graphique lui permet de suivre l'évolution temporelle d'une crypto-monnaie spécifique.

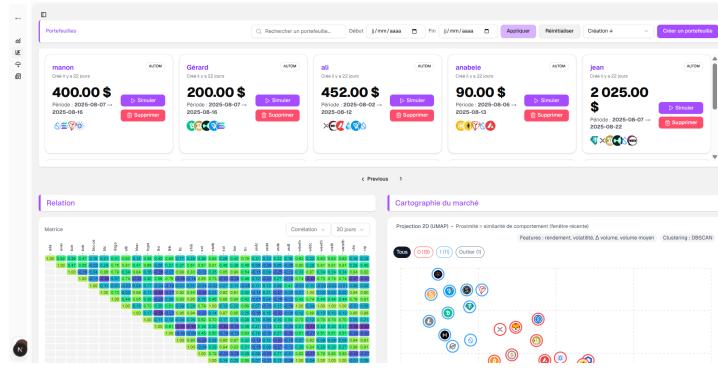


FIGURE C.4 – Page Simulation : Capture d’écran

## Simulation et gestion de portefeuilles

La page simulation (figure C.4) offre une double perspective. D'une part, elle propose une liste des portefeuilles déjà créés, consultables et simulables à la demande. D'autre part, elle intègre des outils d'analyse avancée tels qu'une matrice de corrélation et une cartographie du marché construite à partir d'UMAP et de DBSCAN, permettant d'identifier des groupes de cryptomonnaies ayant des comportements similaires.

La création d'un portefeuille s'effectue en deux étapes successives. Dans un premier temps, l'utilisateur renseigne les informations générales : nom, budget, période de détention, méthode d'allocation et stratégie d'optimisation . Dans un second temps, il définit la répartition des actifs si l'option manuelle est choisie . Une fois le portefeuille validé, il peut être simulé, et la page de résultats présente alors la valeur cumulée, la performance de chaque actif, ainsi que des indicateurs de risque détaillés tels que la volatilité, le maximum de drawdown ou le ratio de Sharpe .

## Gestion du risque

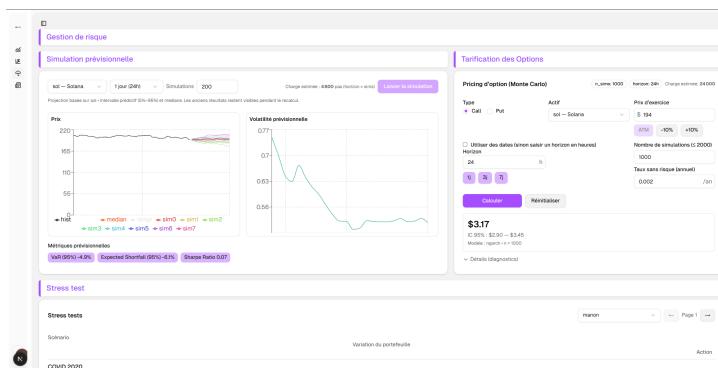
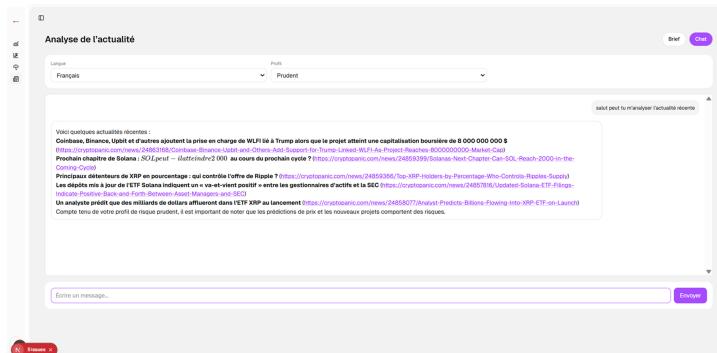


FIGURE C.5 – Page Gestion de risque : Capture d’écran

La page gestion du risque (figure C.5) met à disposition plusieurs modules d'évaluation avancée. L'utilisateur peut réaliser des analyses prévisionnelles sur une crypto-monnaie donnée, effectuer des tarifications d'options à l'aide de modèles économétriques et d'apprentissage profond (GRU, NGARCH, simulations Monte Carlo), ou encore soumettre ses portefeuilles à différents scénarios de stress test (historiques, par facteurs ou uniformes). Cette section complète

les fonctionnalités de simulation en offrant une vision plus approfondie de l'exposition aux risques.

## Assistant



**FIGURE C.6 – Page Assistant : Capture d’écran**

Enfin, la page Assistant (figure C.6) introduit une dimension interactive et personnalisée. L’agent virtuel fournit un résumé des actualités récentes et permet de dialoguer avec l’utilisateur via un chat. Grâce à l’accès aux données du marché, aux performances et aux portefeuilles, l’assistant peut contextualiser ses réponses et proposer des recommandations adaptées au profil de l’utilisateur (prudent, dynamique, etc.). Cet outil constitue un prolongement naturel de la plateforme, en transformant l’information en conseil directement exploitable.

## Conclusion

Le parcours utilisateur décrit dans ce chapitre illustre la cohérence entre les besoins fonctionnels identifiés et l’architecture logicielle mise en place. De la consultation du marché à la simulation de portefeuilles, en passant par l’analyse du risque et l’assistant conversationnel, chaque étape a été pensée pour accompagner le trader ou le gestionnaire dans ses décisions. La clarté du flux utilisateur et la richesse des interfaces constituent ainsi un atout majeur pour l’adoption et l’utilisation efficace du système. Le chapitre suivant détaillera les aspects techniques de mise en œuvre de ces fonctionnalités.

---

---

## ANNEXE D

---

# COUCHES TECHNIQUES DÉTAILLÉES

## Introduction

L'architecture globale présentée précédemment doit être détaillée afin de mieux comprendre le rôle de chaque composant et les choix techniques qui ont guidé leur implémentation. Ce chapitre explore les différentes couches de la solution, en commençant par les données et en finissant par le déploiement.

## D.1 Couche de données

### D.1.1 Description des entités

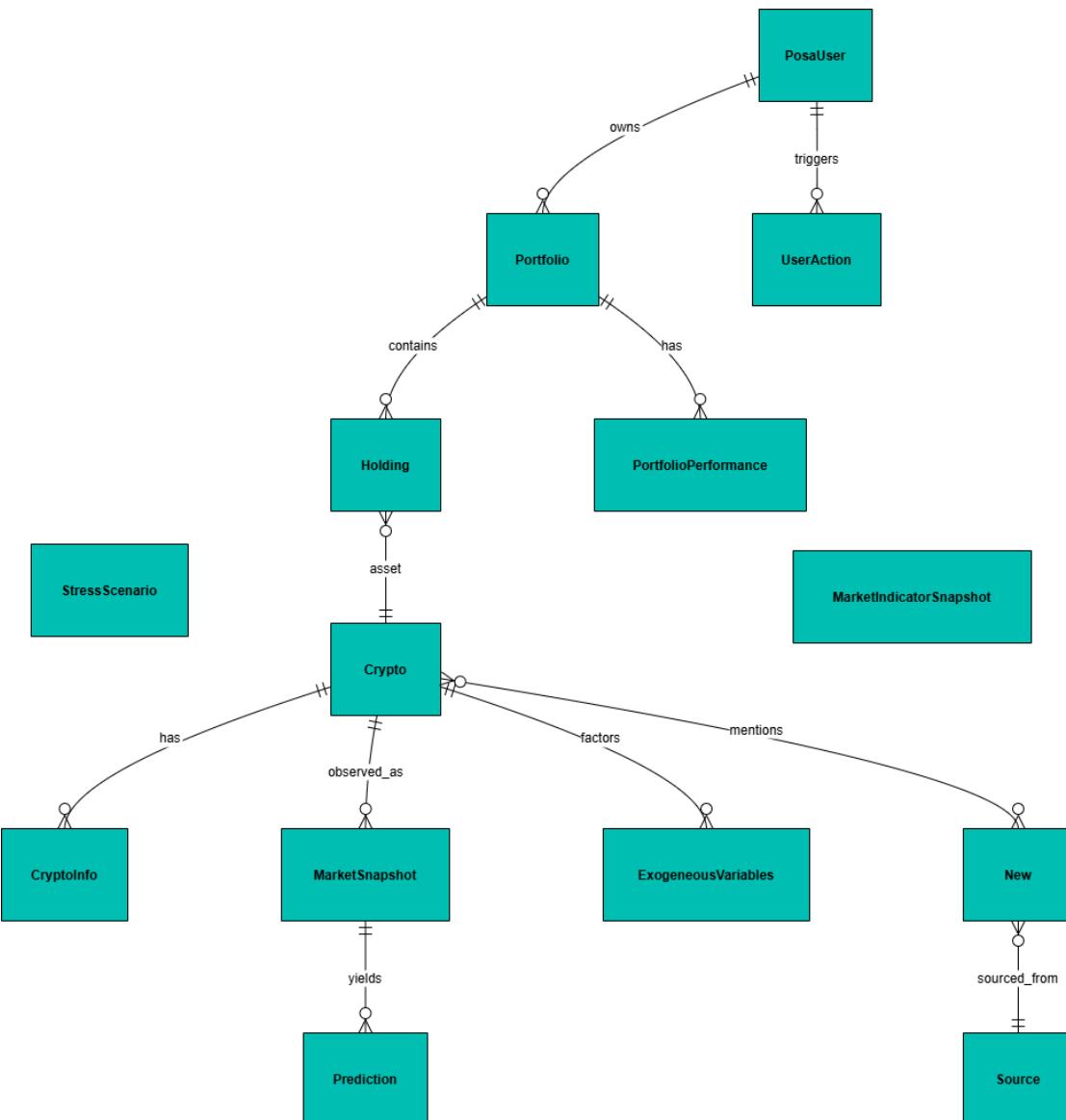


FIGURE D.1 – Relations entre entités (vue synthétique).

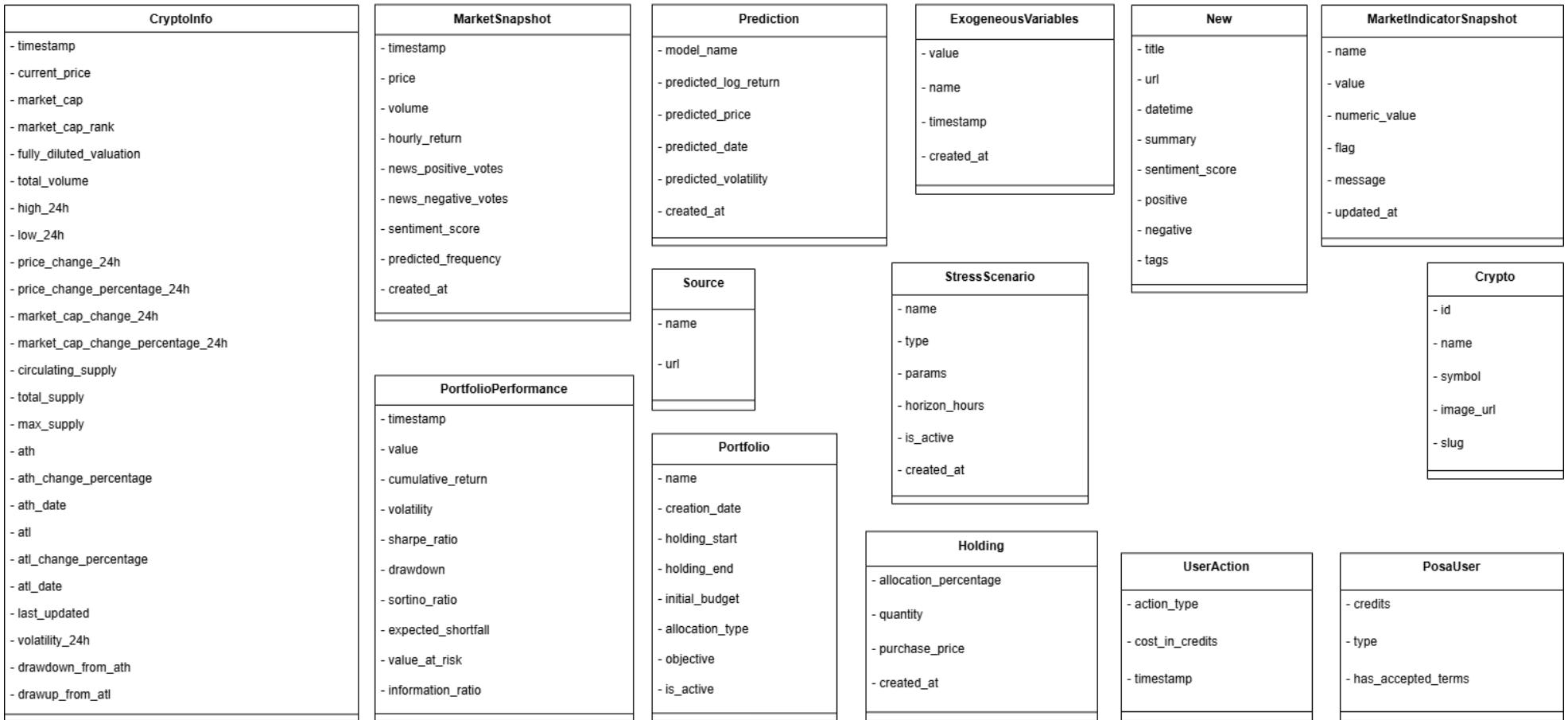


FIGURE D.2 – Entités et attributs (sans relations) — vue élargie.

## Description

La figure D.1 offre une lecture rapide de la structure logique du domaine. Le cœur utilisateur s'appuie sur `PosaUser`, étendu par un journal `UserAction` qui trace les actions et les crédits consommés. Chaque utilisateur peut créer plusieurs portefeuilles modélisés par `Portfolio`. Un portefeuille est défini par une période de détention, un budget initial et un mode d'allocation (manuel ou automatique), ce qui prépare les calculs ultérieurs sans imposer de choix d'implémentation dans l'interface.

Un portefeuille regroupe des positions `Holding` qui rattachent un actif `Crypto` à une allocation et à une quantité. L'évolution dans le temps est matérialisée par `PortfolioPerformance`, qui conserve les séries de valeur, de rendement cumulé et les principaux indicateurs de risque. Cette séparation entre positions et performance permet de recomposer l'historique sans dupliquer l'information financière de référence.

Le référentiel `Crypto` constitue l'axe central des données de marché. Les états récupérés auprès des sources externes sont archivés dans `CryptoInfo`, tandis que `MarketSnapshot` concentre les instantanés alignés et prêts à l'usage des modèles. Les résultats de prédiction sont consignés dans `Prediction`, de manière à distinguer clairement le temps de l'observation du temps de la prévision, ce qui facilite les évaluations rétrospectives. Les variables exogènes `ExogeneousVariables` enrichissent, lorsque nécessaire, les descriptions d'actifs par des facteurs complémentaires.

Le contexte informationnel est modélisé par `New`, qui relie des contenus à des actifs, avec une provenance explicite via `Source`. Cette articulation permet d'exploiter aussi bien des signaux quantitatifs que des signaux d'actualité, sans les confondre dans les séries de prix. À un niveau agrégé, `MarketIndicatorSnapshot` stocke des indicateurs de marché calculés et mémoisés pour l'interface, afin d'éviter des recalculs coûteux et de garantir des temps de réponse stables.

Enfin, `StressScenario` catalogue les chocs utilisés par les tests de résistance. Les paramètres et le type de scénario y sont décrits de façon autonome, ce qui autorise leur réutilisation dans différents contextes d'analyse sans rigidifier le modèle de données.

La figure D.2 détaille les champs portés par chaque entité, conformément au fichier `models.py`. On y retrouve, côté marché, les mesures usuelles (prix, volumes, capitalisation, extrêmes et variations) aux horodatages explicites, ainsi que les colonnes synthétiques nécessaires aux modules de prédiction. Côté portefeuille, les attributs couvrent l'identité, la période de détention, l'allocation et les métriques de suivi. L'ensemble préserve la lisibilité entre données sources, données dérivées et résultats de modèles, tout en restant cohérent avec les besoins fonctionnels exposés dans les chapitres précédents.

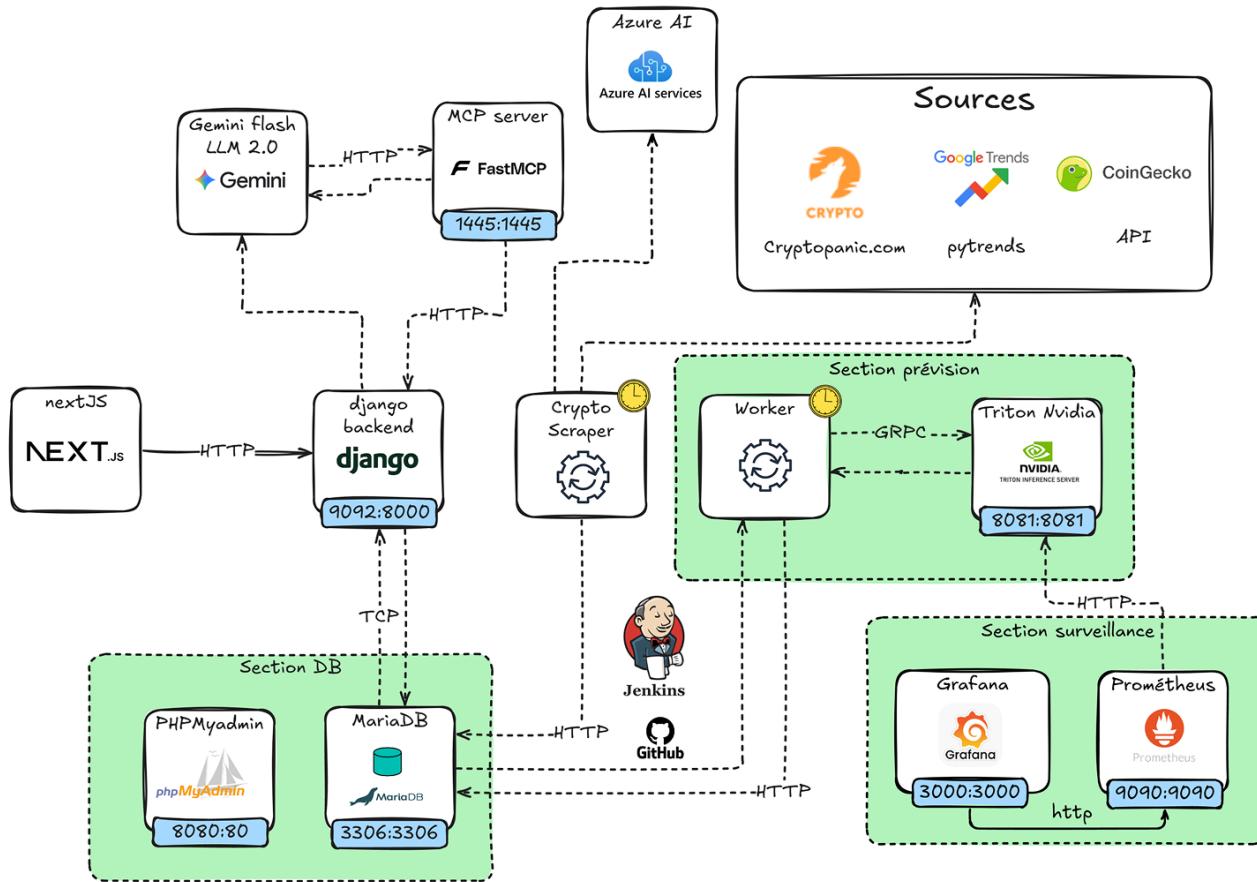
## D.2 Couche applicative (Backend)

### D.2.1 Endpoints principaux

Chemin (pattern)	Type	Rôle (résumé)	Permissions / Notes
/cryptos/ ; /cryptos/{id}/	ViewSet	Référentiel des actifs (annotate la dernière prédition par modèle pour chaque crypto). Endpoints REST : list, retrieve, create, update, destroy.	Lecture publique ; écriture = authentifié (IsAuthenticatedOrReadOnly).
/crypto-infos/ ; /crypto-infos/{id}/	ViewSet	Observations marché (prix, market cap, etc.). Action dédiée : GET /crypto-infos/top/ (top variations récentes).	Lecture publique ; écriture = authentifié. Action top = AllowAny.
/portfolios/ ; /portfolios/{id}/	ViewSet	Portefeuilles de l'utilisateur (filtrage name, crypto, start, end). Actions : POST /portfolios/{id}/simulate/ (lance la simulation, 202), GET /portfolios/{id}/crypto-returns/.	Authentifié (propriétaire). Recherche activée ; actions lancées en thread.
/holdings/ ; /holdings/{id}/	ViewSet	Positions d'un portefeuille (scopées à l'utilisateur courant). Endpoints REST standard.	Authentifié.
/news/ ; /news/{id}/	ViewSet	Articles d'actualité reliés aux cryptos. Action : GET /news/latest/ (5 dernières).	ViewSet = Authentifié ; /news/latest/ = AllowAny.
/users/ ; /users/{id}/	ViewSet	Gestion des utilisateurs (administration).	Admin uniquement (IsAdminUser).
/market-snapshots/ ; /market-snapshots/{id}/	ViewSet	Instantanés horaires alignés « prêts modèle » (features). Endpoints REST standard.	AllowAny.
/predictions/ ; /predictions/{id}/	ViewSet	Prédictions par snapshot / modèle (persistance des résultats).	Lecture publique ; écriture = authentifié (IsAuthenticatedOrReadOnly).
/cryptos/latest-info/	GET	Liste des cryptos avec dernières métriques (prix, var 24h, market cap).	Public (permission commentée).
/market/indicators/	GET	Indicateurs agrégés marché (tableau d'indicateurs calculés).	Authentifié.
/market/history/	GET	Historique d'un symbole sur une période (range : 1d/7d/30d), liste horaire prix.	Authentifié ; params : symbol, range.
/crypto-relations/	GET	Matrice de relations : Spearman ou causalité de Granger ; renvoie matrix + cryptos.	Authentifié ; params : type (spearman/granger), period (ex. 30d), lag, k.
/crypto-map/	GET	Cartographie marché (UMAP + DBSCAN) avec métriques (return, volatility, volume).	Authentifié ; params : days, min_points, k.
/risk/simulate/	GET	Simulation prévisionnelle (NGARCH + Monte Carlo) : renvoie chemins, vol, métriques.	Public (permission commentée). Params : symbol, horizon_hours, n_sims.
/risk/option/price/	POST	Tarification d'options (MC) à partir de l'historique horaire (6 mois).	Authentifié ; payload validé par OptionPricingInputSerializer.
/risk/stress/scenarios/	GET	Liste des scénarios de stress actifs (catalogue).	Permissions par défaut (non précisées dans la vue).
/risk/stress/apply/	POST	Applique un scénario de stress à un portefeuille donné.	Permissions par défaut ; payload : scenario, portfolio_id.
/auth/user/	GET	Récupère le profil de l'utilisateur courant.	Authentifié.
/auth/token/	POST	JWT (obtain pair).	DRF SimpleJWT.
/auth/token/refresh/	POST	Rafraîchit le token d'accès.	DRF SimpleJWT.
/auth/register/	POST	Inscription ; retourne aussi les tokens (refresh/access).	Public (CreateAPIView).

(suite page suivante)

Chemin (pattern)	Type	Rôle (résumé)	Permissions / Notes
/auth/	(include)	Endpoints <code>dj_rest_auth</code> : login, logout, password reset, etc.	Inclus en bloc.
/auth/registration/	(include)	Endpoints d'inscription ( <code>dj-rest-auth registration</code> ).	Inclus en bloc.
/auth/google/token/	POST	Échange un <code>access_token</code> Google contre des JWT internes.	Public ; payload : <code>access_token</code> .
/llm/portfolios/list/	GET	Liste compacte des portefeuilles de l'utilisateur (id, name) pour l'agent LLM.	Permission custom <code>IsLLMRequest</code> .
/llm/portfolio/{pk}/summary/	GET	Résumé d'un portefeuille (structures et positions) pour l'agent LLM.	<code>IsLLMRequest</code> .
/llm/market-metrics/	GET	Indicateurs marché (via <code>MarketInfoManager</code> ) pour l'agent LLM.	<code>IsLLMRequest</code> .
/assist/brief/	POST	Génère un <i>brief</i> Markdown des news (MCP + LLM), paramétrable (since_hours, limit, lang, risk).	Authentifié ; timeout configurable.
/assist/chat/	POST	Chat assisté (contexte + outils MCP) ; renvoie du Markdown.	Authentifié ; historisation légère côté requête.



Architecture de déploiement de la plateforme POSA : l'interface Next.js communique avec l'API Django ; les services d'ingestion et le worker de prédiction alimentent Triton (gRPC) et la base MariaDB ; l'observabilité est assurée par Prometheus et Grafana ; l'assistant s'appuie sur un serveur MCP connecté au LLM externe ; les données proviennent de sources publiques (CoinGecko, pytrnd, CryptoPanic).

**FIGURE D.3** – Architecture de déploiement de la plateforme POSA.

TABLE D.2 – Définitions des métriques utilisées (FinTSB et conventions financières)

Catégorie	Métrique	Description	Formule	Sens
Erreur	<b>MAE</b> (Mean Absolute Error)	Erreur moyenne absolue entre la cible $y_t$ et la prédiction $\hat{y}_t$ . Robuste aux valeurs aberrantes.	$MAE = \frac{1}{T} \sum_{t=1}^T  y_t - \hat{y}_t $	↓
Erreur	<b>MSE</b> (Mean Squared Error)	Erreur quadratique moyenne. Pénalise plus fortement les grandes erreurs.	$MSE = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2$	↓
Classement	<b>MCFD</b> (Mean Correct Forecast Direction)	Taux de bonnes directions prévues : proportion des périodes où le signe de la prévision concorde avec le signe observé (utile pour des rendements).	$MCFD = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{\text{sign}(\hat{r}_t) = \text{sign}(r_t)\}$	↑
Portefeuille	<b>MFTR</b> (Mean Forecasting Trading Return)	Rendement moyen d'une stratégie <i>long/short</i> qui suit le signe prédit : on prend position $s_t = \text{sign}(\hat{r}_t)$ et le rendement de trading par période est $\tilde{r}_t = s_t r_t$ .	$MFTR = \frac{1}{T} \sum_{t=1}^T \text{sign}(\hat{r}_t) r_t$	↑
Portefeuille	<b>ARR</b> (Annualized Return Rate)	Rendement annualisé sur la période étudiée (composition géométrique). $P$ est le nombre de périodes par an (ex. : $P=252$ jours, $P=365\times 24$ heures).	$ARR = \left( \prod_{t=1}^T (1 + r_t) \right)^{P/T} - 1$	↑
Portefeuille	<b>ANV</b> (Annual Volatility)	Volatilité annualisée des rendements simples $r_t$ . $\sigma$ est l'écart-type empirique des $r_t$ .	$ANV = \sigma(r_{1:T}) \sqrt{P}$	↓
Risque	<b>MDD</b> (Maximum Drawdown)	Perte maximale depuis un plus haut sur la courbe de valeur cumulée $C_t = \prod_{i=1}^t (1 + r_i)$ . On reporte souvent l'amplitude $ \text{MDD} $ .	$MDD = \min_{1 \leq t \leq T} \left( \frac{C_t}{\max_{1 \leq s \leq t} C_s} - 1 \right)$	↓
Risque	<b>ASR</b> (Annualized Sharpe Ratio)	Ratio de Sharpe annualisé sur rendements d'excès <sup>1</sup> . Mesure le rendement par unité de risque.	$ASR = \frac{\bar{r} - \bar{r}_f}{\sigma(r)} \sqrt{P}$	↑

**Conventions.**  $T$  : nombre d'observations ;  $r_t$  : rendement simple à la fréquence des données ;  $P$  : facteur d'annualisation (périodes/an) adapté à ta fréquence ;  $\bar{r}$  : moyenne des  $r_t$  ;  $\sigma(r)$  : écart-type des  $r_t$  ;  $\mathbf{1}\{\cdot\}$  : indicatrice ;  $s_t = \text{sign}(\hat{r}_t)$  : signal long/short ;  $\tilde{r}_t = s_t r_t$  : rendement de trading.

---

# BIBLIOGRAPHIE GÉNÉRALE

- AI FOR ALPHA (2025). *Ai for Alpha — AI-powered investment decision*. URL : <https://aiforalpha.com/>.
- ANCRYPTO (2024). *Understanding Ownership on Blockchain*. URL : <https://www.ancrypto.io/what-is-ownership-in-terms-of-blockchain/> (visité le 01/03/2025).
- AUTORITÉ DES MARCHÉS FINANCIERS (2024). *Qu'est-ce qu'une « cryptomonnaie » ?* URL : <https://www.amf-france.org/fr/quest-ce-quune-cryptomonnaie> (visité le 01/03/2025).
- BAO, Wei, Jun YUE et Yulei RAO (2017). “A deep learning framework for financial time series using stacked autoencoders and long-short term memory”. In : *PLOS ONE* 12.7, e0180944. DOI : [10.1371/journal.pone.0180944](https://doi.org/10.1371/journal.pone.0180944).
- BLACK, Fischer et Myron SCHOLES (1973). “The Pricing of Options and Corporate Liabilities”. In : *Journal of Political Economy* 81.3, p. 637-654.
- BOLLERSLEV, Tim (1986). “Generalized Autoregressive Conditional Heteroskedasticity”. In : *Journal of Econometrics*. URL : [https://public.econ.duke.edu/~boller/Published\\_Papers/joe\\_86.pdf](https://public.econ.duke.edu/~boller/Published_Papers/joe_86.pdf).
- BREEDEN, Douglas T. et Robert H. LITZENBERGER (1978). “Prices of State-Contingent Claims Implicit in Option Prices”. In : *Journal of Business* 51.4, p. 621-651.
- CHAINALYSIS (2024). *The 2024 Geography of Crypto Report*. Rapp. tech. October 2024. Chainalysis. URL : <https://www.chainalysis.com>.
- COINBASE (2024). *What is a fork ?* URL : <https://www.coinbase.com/fr/learn/crypto-basics/what-is-a-fork> (visité le 01/04/2025).
- COINHOUSE (2024). *Proof-of-Stake (PoS) : découvrez la preuve d'enjeu*. URL : <https://www.coinhouse.com/fr/academie/blockchain/proof-of-stake> (visité le 01/03/2025).
- CRYPTOAST (2024). *Qu'est-ce que la preuve de travail ou proof-of-work (PoW) ?* URL : <https://cryptoast.fr/qu-est-ce-que-le-pow-proof-of-work/> (visité le 01/05/2025).
- CUNHA, C.R. da et al. (2020). “Relevant Stylized Facts about Bitcoin”. In : *Physica A*. URL : <https://www.sciencedirect.com/science/article/abs/pii/S0378437120300133>.
- DÉVELOPPEMENT DES ACTIFS NUMÉRIQUES (ADAN), Association pour le (2025). *Web3 et crypto en France et en Europe : vers une adoption durable et institutionnelle – Édition 2025*.

## BIBLIOGRAPHIE GÉNÉRALE

---

- [https://www.adan.eu/publication/web-3-et-crypto-en-france-et-en-europe-vers-une-adoption-durable-et-institutionnelle-edition-2025/.](https://www.adan.eu/publication/web-3-et-crypto-en-france-et-en-europe-vers-une-adoption-durable-et-institutionnelle-edition-2025/)
- DING, Zhuanxin, Clive W. J. GRANGER et Robert F. ENGLE (1993). “A Long Memory Property of Stock Market Returns and a New Model”. In : *Journal of Empirical Finance* 1.1, p. 83-106.
- DRESCHER, Daniel (2017). *Blockchain Basics : A Non-Technical Introduction in 25 Steps*. Apress. ISBN : 9781484226032.
- DUFOUR, Jean-Marie (avr. 2002). *Tests de causalité*. Rapp. tech. Première version : octobre 2000 ; révisions : avril 2002 ; cette version : 15 avril 2002. Montréal, QC, Canada : Université de Montréal, Département de sciences économiques.
- DUTTA, Aniruddha, Saket KUMAR et Meheli BASU (2020). “A Gated Recurrent Unit Approach to Bitcoin Price Prediction”. In : *Journal of Risk and Financial Management* 13.2, p. 23. DOI : 10.3390/jrfm13020023.
- ENGLE, Robert F. (1982). “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation”. In : *Econometrica*. URL : <https://www.econ.uiuc.edu/~econ536/Papers/engle82.pdf>.
- FINANCIAL STABILITY BOARD (2022). *Assessment of Risks to Financial Stability from Crypto-Assets*. <https://www.fsb.org/uploads/P160222.pdf>.
- FISCHER, Thomas et Christopher KRAUSS (mai 2017). *Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions*. FAU Discussion Papers in Economics 11/2017. Working Paper. Nürnberg : Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics. URL : <https://hdl.handle.net/10419/157808>.
- GAUNTLET (2025). *Institutional-grade vaults for DeFi*. URL : <https://www.gauntlet.xyz/>.
- GHOSH, Pushpendu, Ariel NEUFELD et Jajati Keshari SAHOO (2022). “Forecasting directional movements of stock prices for intraday trading using LSTM and random forests”. In : *Finance Research Letters* 46, p. 102280. ISSN : 1544-6123. DOI : <https://doi.org/10.1016/j.frl.2021.102280>. URL : <https://www.sciencedirect.com/science/article/pii/S1544612321003202>.
- GLASSNODE (2025). *Pricing*. URL : <https://studio.glassnode.com/pricing>.
- GLOSTEN, Lawrence R., Ravi JAGANNATHAN et David E. RUNKLE (1993). “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks”. In : *The Journal of Finance* 48.5, p. 1779-1801.
- HU, Yifan et al. (2025). “FinTSB : A Comprehensive and Practical Benchmark for Financial Time Series Forecasting”. In : arXiv : 2502.18834. URL : <https://arxiv.org/abs/2502.18834>.
- KAIKO (2025). *Pricing and Licenses (enterprise)*. URL : <https://www.kaiko.com/about-kaiko/pricing-and-contracts>.
- KAKADE, K. et al. (2022). “A hybrid ensemble learning GARCH-LSTM based approach”. In : *Energy Policy*. URL : <https://www.sciencedirect.com/science/article/abs/pii/S0301420722003476>.
- MCNEIL, Alexander J. (1999). “Extreme Value Theory for Risk Managers”. In : URL : <https://www.sfu.ca/~rjones/econ811/readings/McNeil%201999.pdf>.

- MCNEIL, Alexander J. et R. FREY (2000). "Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series". In : *Journal of Empirical Finance*. URL : [https://faculty.washington.edu/ezivot/econ589/EVT\\_Mcneil\\_Frey\\_2000.pdf](https://faculty.washington.edu/ezivot/econ589/EVT_Mcneil_Frey_2000.pdf).
- MERTON, Robert C. (1973). "Theory of Rational Option Pricing". In : *Bell Journal of Economics and Management Science* 4.1, p. 141-183.
- MESSARI (2025). *Pricing / Enterprise / API*. URL : <https://messari.io/pricing>.
- NELSON, Daniel B. (1991). "Conditional Heteroskedasticity in Asset Returns : A New Approach". In : *Econometrica* 59.2, p. 347-370.
- NG, Benny Siu Hon, Christopher R. KNITTEL et Caroline UHLER (déc. 2020). *A Machine Learning Approach to Evaluating Renewable Energy Technology : An Alternative LACE Study on Solar Photo-Voltaic (PV)*. Rapp. tech. Working Paper 2020-021. MIT Center for Energy et Environmental Policy Research (CEEPR). URL : <https://cepr.mit.edu/wp-content/uploads/2021/09/2020-021.pdf>.
- OPENBB (2025a). *Apps Gallery — Portfolio Risk Management*. URL : <https://docs.openbb.co/workspace/gallery>.
- (2025b). *OpenBB Platform (open source)*. URL : <https://github.com/OpenBB-finance/OpenBB>.
- QIU, Jiayu, Bin WANG et Changjun ZHOU (2020). "Forecasting stock prices with long-short term memory neural network based on attention mechanism". In : *PLOS ONE* 15.1, e0227222. DOI : 10.1371/journal.pone.0227222.
- RITCHKEN, Peter et Rob TREVOR (fév. 1999). "Pricing Options under Generalized GARCH and Stochastic Volatility Processes". In : *The Journal of Finance* 54.1, p. 377-402.
- ROSZYK, N. et R. SLEPACZUK (2024). "The Hybrid Forecast of S&P 500 Volatility EnsemblED from VIX, GARCH and LSTM". In : *arXiv*. URL : <https://arxiv.org/abs/2407.16780>.
- SAEF, D. et al. (2024). "Jumps in Cryptocurrencies : Evidence from Tick-by-Tick Data". In : *Digital Finance*. URL : <https://link.springer.com/article/10.1007/s42521-024-00116-1>.
- SHEN, D. et al. (2020). "Forecasting the Volatility of Bitcoin : The Importance of Jumps and Structural Breaks". In : URL : [https://centaur.reading.ac.uk/87932/1/R%26R\\_Forecasting%20the%20Volatility%20of%20Bitcoin\\_complete.pdf](https://centaur.reading.ac.uk/87932/1/R%26R_Forecasting%20the%20Volatility%20of%20Bitcoin_complete.pdf).
- TRUONG, Charles, Laurent OUDRE et Nicolas VAYATIS (2020). "Selective Review of Offline Change Point Detection Methods". In : *Signal Processing*. Review article.
- WELLER, Chris (2025). *Crypto-trading addicts share their stories of losses and struggles*. <https://www.businessinsider.com/crypto-trading-addicts-gambling-therapy-marriage-conflicts-financial-losses-lawsuits-2025-2>. Consulté le 14 août 2025.
- XGBOOST DEVELOPERS (s. d.). *Introduction to Boosted Trees*. <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>. Consulté le 2025-03-12.
- ZAKOIAN, Jean-Michel (1994). "Threshold Heteroskedastic Models". In : *Journal of Economic Dynamics and Control* 18.5, p. 931-955.



## ECOLE SUPÉRIEURE PRIVÉE D'INGÉNIERIE ET DE TECHNOLOGIES

**www.esprit.tn - E-mail : contact@esprit.tn**

**Siège Social : 18 rue de l'Usine - Charguia II - 2035 - Tél. : +216 71 941 541 - Fax. : +216 71 941 889**

**Annexe : Z.I. Chotrana II - B.P. 160 - 2083 - Pôle Technologique - El Ghazala - Tél. : +216 70 685 685 - Fax. : +216 70 685 454**

## BIBLIOGRAPHIE GÉNÉRALE

---