Construire une boussole pour explorer le cryptoverse

Table des matières

1. Introduction : comprendre le phénomène crypto au-delà des apparences

Depuis leur création en 2009, avec le bitcoin, les cryptomonnaies ont connu une histoire mouvementée, mais riche. Leur adoption a explosé, passant de quelques milliers à près de 900 millions d'utilisateurs (Statista). Pourtant, malgré cette croissance spectaculaire, elles restent mal comprises du grand public.

a. Un écosystème divisé : entre mèmes et utilité réelle

Cette méconnaissance s'explique en partie par l'image dominante des cryptomonnaies, souvent associées à l'instabilité, à la spéculation pure, au *meme-trading*, ou au regret persistant de ne pas avoir investi dans le Bitcoin à ses débuts

Pourtant derrière cette façade bruyante, un mouvement de fond est en cours. Comme le révèle « The 2025 Geography of Crypto Report », les cryptomonnaies connaissent une adoption croissante, en particulier dans les régions en développement telles que l'Asie, l'Amérique latine, ou l'Afrique.

Cette adoption repose sur des besoins pragmatiques :

- Outil de protection contre l'inflation et la dévaluation monétaire (Nigéria, Turquie par exemple).
- Usage quotidien comme moyen de paiement ou pour les transferts internationaux (Vietnam, Nigéria par exemple).
- À l'inverse, c'est dans les économies les plus développées (Europe, Amérique du Nord) que le rôle spéculatif des cryptos demeure le plus marqué.

Parallèlement, des innovations structurelles redéfinissent l'écosystème :

- Adoption Institutionnelle : Approbation des ETF, adoption institutionnelle aux États-Unis, ou Tokenisation des Actifs du Monde Réel (RWA).
- Infrastructure Décentralisée (DePIN) : De nouvelles solutions de stockage (comme Filecoin), de calcul décentralisé (Akash pour la location de GPU, Render Network utilisé dans le *rendering* de projets 3D), d'entraînement d'IA décentralisé avec Bittensor, ou de location de bande passante avec GRASS.

Face à cette dualité, spéculation d'un côté, et adoption concrète de l'autre, il devient obligatoire d'adopter une grille d'analyse objective. C'est pour naviguer avec une direction claire dans cet univers complexe que j'ai initié ce projet de data science, afin de me doter d'un outil d'exploration et d'analyse rigoureux.

b. Premier pas : de l’intuition à la méthode

Mon approche de l’écosystème repose avant tout sur un besoin d’établir une méthode rigoureuse.

En tant qu’ancien Chef de projet / Business Developer, habitué à l’optimisation multicritères, et aujourd’hui Data Scientist spécialisé dans l’évaluation des modèles d’intelligence artificielle et la qualité des données, il m’est impossible d’aborder l’univers crypto par son angle purement psychologique.

Je parle de psychologie car c’est une composante importante en finance. Mais, dans les crypto, elle est amplifiée par les réseaux sociaux, source de FUD/FOMO (Fear, Uncertainty, Doubt / Fear of Missing Out).

En science, on commence souvent, dans un premier temps, par simplifier le problème. Ici, cela signifie se concentrer sur le prix seul, pour voir s’il contient un signal exploitable.

J’ai donc naturellement commencé à porter mon intérêt sur le bitcoin, à la fois point de départ et véritable poumon de l’écosystème.

Pour cette raison j’ai choisi de construire une démarche analytique objective, basée sur les données, pour m’interroger sur ce qui fait fondamentalement la valeur d’un projet. Surtout qu’une des promesses des cryptomonnaies est celle de la transparence et de l’accessibilité de ces données.

c. Un journal de bord comme laboratoire de réflexion

Pour mener à bien cette exploration, j'ai choisi de structurer l'ensemble de ma démarche en une série de cinq articles, sous différents formats, une sorte de journal de bord qui se veut transparent et méthodologique.

Ce cheminement s'articulera de la manière suivante :

- Cet article : Présentation de la démarche et des objectifs.
- Article 2 (Notebook) : Les fondations : maîtriser la source des données.
- Article 3 (Notebook) : Le cœur de l'expérience : la modélisation ML.
- Article 4 : Retour sur expérience (En cours de rédaction)
- Article 5 : Le pivot : quand les résultats imposent une nouvelle direction. (En cours de rédaction)

2. Postulat de départ : Peut-on deviner demain avec hier ?

a. L'hypothèse testée : peut-on prédire le prix avec le prix seul ?

L'idée ici est de savoir si le prix passé seul est suffisant pour déterminer le prix futur. Parce qu'en lisant sur internet, il est possible de trouver les deux réponses à cette question. C'est l'un des débats les plus vifs dans la finance quantitative. Il est donc question de me forger mon propre avis sur la question, en me basant sur mes propres expériences et une méthodologie reproductible, tous mes codes étant publique.

Les séries temporelles en finance sont souvent considérées comme un exercice extrêmement compliqué. Elles sont dites non-stationnaires et potentiellement des marches aléatoires, ce qui signifie que les signaux d'hier ne sont pas de bons prédicteurs pour demain.

La question centrale est : avec le bond technologique récent en Intelligence Artificielle (notamment les réseaux récurrents comme les LSTM et GRU que j'utiliserai), est-il possible de résoudre ce problème ou, du moins, d'extraire un signal exploitable qui contredirait l'hypothèse d'inefficience du marché ?

b. Une démarche en trois temps : données, exploration, modélisation

Le premier pas pour tester cette hypothèse, est d'établir un plan d'action. Pour cela, je me suis renseigné en amont en lisant des articles, des repos sur GitHub, et des notebooks sur Kaggle. Cette revue rapide m'a permis de confirmer une direction claire, et de décider d'un plan en trois étapes :

- Le pipeline de donnée : Tout projet de modélisation sans données fiable et de qualité est voué à l'échec. En plus de garantir l'intégrité des données, chercher à maîtriser l'acquisition d'un système de récupération de données nous permettra d'intégrer plus facilement d'autres cryptos par la suite.
- L'analyse exploratoire (EDA) : Une fois le dataset récupéré, l'étape suivante est de s'assurer de sa qualité, et de comprendre les données soi-même. L'EDA est la base indispensable pour extraire des informations pertinentes et décrypter les dynamiques du marché.
- La modélisation : Enfin il sera temps d'établir des modèles d'intelligence artificielle. Nous commencerons par établir un modèle de base (inspiré de codes en ligne), pour s'assurer de la faisabilité de la démarche, avant de passer à une étape d'optimisation pour obtenir les meilleurs résultats possible.

C'est à l'issue de cette démarche que nous pourrons déterminer si l'hypothèse de prédictibilité par le prix seul est validée ou, au contraire, rejetée.

1. La feuille de route : construire avant de prédire

L'hypothèse de départ étant établie, de même que la nécessité d'avoir une démarche rigoureuse, nous pouvons détailler la feuille de route.

Dans les parties suivantes, nous présenterons avant tout la philosophie, et les enjeux majeurs de chacune des démarches prévues. Les aspects techniques, y compris le code, les analyses détaillées, et les résultats bruts, seront quant à eux disponibles dans les notebook associés.

Enfin, une attention particulière a été portée à la modularité du code, pour plusieurs raisons :

- Pouvoir le réutiliser pour d'autre cryptos facilement, ou dans d'autres projets.
- M'entraîner à rédiger du code propre, lisible et maintenable.
- Améliorer la lisibilité globale du projet.

a. Article 2 – Les fondations : maîtriser la source des données

Comme il a été dit précédemment, la qualité des données est le facteur de succès numéro un en Machine Learning. Il est impossible d'obtenir des résultats fiables lorsque les données sont incomplètes, biaisées ou mal comprises.

Construire un pipeline de donnée

Il est possible de télécharger des *dataset* « tout faits », mais ceux-ci présentent plusieurs inconvénients :

- Obsolescence : Ils sont figés, et ne bénéficie donc pas de mise à jour en temps réel.
- Qualité : Selon leur provenance, leur méthode d'acquisition et de nettoyage leur qualité n'est pas assurée.
- Flexibilité : Ils ne sont pas forcément disponibles pour toutes les cryptomonnaies.

Une autre option serait l'analyse *on-chain,* c'est-à-dire interroger la *blockchain* directement pour obtenir des informations. Mais une fois encore, cela présente différents inconvénients techniques :

- Modularité réduite : Il faut développer une méthode pour chaque blockchain, ce qui augmente la complexité, et réduit la modularité.
- Limitation des données : Les protocoles des blockchains enregistrent des informations dont les transactions, mais pas les prix des montant échangés.

Pour obtenir ces données de manière efficace, il est donc préférable de passer par une API de marché. Bien que ces services soient souvent payants pour les données historiques, la plateforme Binance offre la possibilité de les obtenir gratuitement, au prix d'un peu de travail technique. En effet, elle ne permet de récupérer les données que par paquets de 1000 lignes.

Malgré ce petit défi technique, il est possible de développer une solution automatisée pour pallier cette limitation.

L'analyse exploratoire des données (EDA)

Maintenant qu'une méthode fiable de récupération des données a été développée il faut s'assurer que les données soient fiables et utilisables.

Dans un premier temps, on utilise une approche classique : on vérifie la présence de valeurs manquantes ou aberrantes, grâce aux fonctionnalités fournies par des bibliothèques de data science comme Pandas. Le tracé de graphiques est également essentiel pour visualiser la continuité des données.

Dans l'entraînement du modèle, nous n'utiliserons que les prix de clôture. Cependant, les données téléchargées contiennent bien plus d'informations (Volume en BTC, Nombre de Transactions, etc.). Il est donc pertinent d'en extraire des informations.

C'est ici que j'ai choisi de m'écarter de l'approche traditionnelle. Plutôt que de me tourner vers des indicateurs techniques classiques de la finance (RSI, Bandes de Bollinger ou autres), qui sont largement disponibles en ligne et sur lesquels je n'aurais pas de valeur ajoutée, j'ai cherché à apporter ma propre analyse en me concentrant sur :

- Le Comportement du Marché : Analyser la distribution de la taille moyenne des transactions pour identifier si le marché Binance est dominé par le retail ou les institutions.
- La Dynamique Interne : Étudier la pression acheteur/vendeur et les volumes pour comprendre la dynamique interne des échanges.
- Les Anomalies : Identifier les ruptures de tendance ou de cycles qui pourraient signaler un changement profond dans le marché (comme celles observées en 2025).

C'est à partir de cette analyse exploratoire que la décision de limiter la période d'entrainement du modèle à l'année 2025 uniquement a été prise.

Pourquoi ? Parce qu'on a pu mettre en évidence plusieurs phases distinctes du marché, avec des comportements différents des acteurs. Entraîner le modèle sur une période où les dynamiques de marché et les comportements sont hétérogènes fausserait les résultats obtenus.

Ainsi, l'EDA révèle son importance : elle ne doit pas être perçue comme une étape optionnelle, mais comme une base essentielle de la modélisation.

b. Article 3 : Le cœur de l'expérience : la modélisation ML.

Afin d'établir une référence initiale et de valider la faisabilité du projet sur les données du Bitcoin, nous avons commencé par une implémentation simple, inspirée des tutoriels existants. Cette démarche a rapidement mis en évidence une limitation méthodologique majeure.

La majorité des implémentations de séries temporelles trouvées en ligne utilisent une approche dite "*Single-step*" (ou prédiction pas à pas), souvent directement fournie par les bibliothèques de machine learning.

Le principe est simple, le modèle prend une fenêtre de données historiques comme entrée, puis prédit une seule valeur future. Ensuite, la fenêtre glisse d'un pas dans le temps, en remplaçant l'ancienne donnée la plus ancienne par la valeur réelle observée, et non par la valeur prédite. Ce schéma est illustré dans la figure suivante.

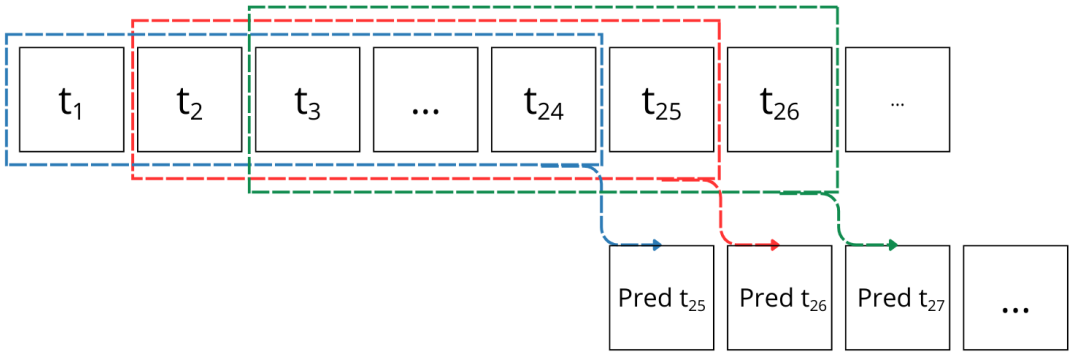

Figure 1. Schéma explicatif de l'approche single-step

Bien que cette méthode génère souvent de belles métriques (le modèle étant constamment "corrigé" par la réalité), elle ne présente que peu d'intérêt pratique. Si l'on se demande quel sera le prix du bitcoin dans une heure (le pas de temps employé), on ne prend guère de risque à répondre qu'il sera encore à peu près le même. Aucun besoin d'intelligence artificielle pour deviner cela.

En outre, cette approche de faire est inapplicable en production :

- Elle suppose un accès continu et immédiat à la valeur réelle au moment de la prédiction, ce qui exige un pipeline de données maintenu à jours en permanence, ce qui peut entrainer un surcoût important.
- Pire encore : si la granularité temporelle est trop courte (à la seconde voire plus bas en trading haute fréquence), et que le temps nécessaire pour récupérer les données, préparer l'entrée du modèle et produire la prédiction est supérieur à ce pas de temps, alors le résultat est périmé avant même d'être calculé.

C’est pour tenter de coller au plus près à la réalité opérationnelle qu’une méthode alternative a été mise en place, inspirée de la documentation officielle de TensorFlow.

Il s’agit d’une approche « multi-step », dans laquelle les prédictions générées sont réinjectées comme entrée pour alimenter les prédictions suivantes. Ainsi, le modèle ne dépend plus des valeurs réelles futures, uniquement de ses propres prédictions comme illustré dans la figure suivante :

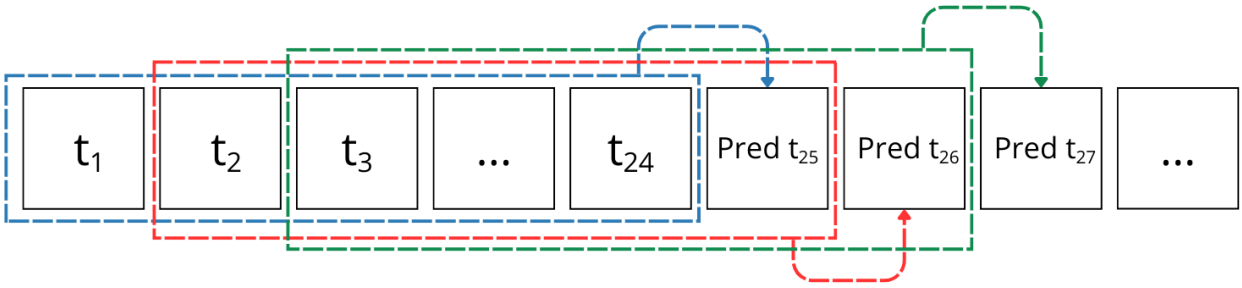

Figure 2. Schéma explicatif de l'approche multi-step / Walk forward

C’est là que réside la véritable valeur de ce travail : tester les modèles dans des conditions proches de la réalité opérationnelle.

Il ne reste plus qu’à implémenter celui-ci, et donc se décider sur l’outil à utiliser parmi la panoplie à disposition du *Data Scientist*.

Il restait alors à choisir l’outil adapté parmi la panoplie disponible au *data scientist*. Les réseaux de neurones récurrents (RNN) s’imposent naturellement : ils sont spécialisés dans le traitement des séquences c’est-à-dire les données ordonnées, qu’il s’agisse de texte ou, dans notre cas, de séries temporelles.

Trois architectures ont été retenues :

- SimpleRNN, le modèle le plus basique, mais limité par sa mémoire à court terme.
- LSTM (*Long Short-Term Memory*), un standard établi en data science et en finance, capable de retenir des dépendances à long terme, au prix d’une plus grande consommation de ressources.
- GRU (*Gated Recurrent Unit*), une alternative plus récente, conçue pour réduire la complexité tout en conservant une capacité d’apprentissage proche de celle du LSTM.

Le SimpleRNN sert de ligne de base, comme référence, le LSTM comme benchmark industriel, et le GRU comme alternative optimisée.

En parallèle, plusieurs hyperparamètres ont été sélectionnés pour l’expérimentation initiale (leurs détails techniques sont réservés aux notebooks)

Pour mettre en place ce processus, deux éléments important ont guidé la conception :

- Code modulaire : pour faciliter la modification des hyperparamètres et la gestion des multiples configurations testées.
- Traçabilité : car le volume croissant d'expériences rend indispensable un suivi précis des résultats. C'est pourquoi MLflow a été intégré dès le départ.
- À la suite de ces expérimentations, il est possible d'en retirer plusieurs résultats :
- Bien qu'il soit possible de maintenir des prédictions relativement stables sur quelques itérations, les erreurs s'accumulent rapidement (phénomène d*'error compounding*), ce qui éloigne rapidement les prédictions de la réalité.
- Indépendamment du niveau d'optimisation des hyperparamètres, aucun modèle ne parvient à capturer la variabilité fine des prix. Ils n'apprennent qu'une tendance générale, souvent incohérente d'une expérience à l'autre.

Afin d'éliminer toute influence perturbatrice de la tendance globale du prix (hausse ou baisse sur la période), nous avons introduit les log returns (rendements logarithmiques). Cette transformation rend la série stationnaire, c'est-à-dire centrée autour d'une moyenne nulle, et se concentre sur les variations relatives plutôt que sur les niveaux absolus.

Le constat est alors sans appel, la courbe prédite par le modèle est une droite à zéro, c'est la démonstration que celui-ci n'apprend rien des données qu'on lui a fournies.

La conclusion la plus évidente est donc que l'hypothèse initiale est fausse : il n'y a pas suffisamment d'informations dans les prix passés seuls pour prédire les prix futurs.

c. Article 4 : Retour d'expérience. (En cours de rédaction – à titre indicatif)

d. Le pivot : quand les résultats imposent une nouvelle direction (En cours de rédaction)

2. Conclusion : une hypothèse fausse, une démarche validée

Cette première étape est terminée. Bien qu'elle puisse paraitre se terminé sur un échec puisque que l'hypothèse testée a été prouvée fausse, et que donc un modèle de machine learning se basant **uniquement** sur les prix passée (j'insiste sur cette partie) apparait comme infaisable. Mais l'objectif initial, à savoir d'utiliser ce projet pour commencer à explorer le cryptoverse est un succès.

J'ai découvert pléthore de ressources qui me seront utiles dans la suite du projets. Elles prennent la forme de sources de données diverses, de guides, d'articles, de publications scientifiques qui m'ont permis de réfléchir sur quel est la prochaine étape de mon parcours.

Donc sans transition, direction la suite : à savoir les articles en version longue, le code publié.