

Data Challenge 2024 - Cryptomarkets

Forecasting and Tactical Allocation for Crypto-assets Portfolios

Federal Finance Gestion & CREM

I - Les grandes lignes du challenge : contexte, enjeux professionnels et académiques

Le sujet et les enjeux

Le challenge porte sur la prévision à court terme des prix et des rendements des principales crypto-monnaies, précisément les onze(11) plus importantes crypto-monnaies cotées sur les marchés, avec en tête le célèbre Bitcoin.

Le challenge présente des enjeux à la fois pratiques pour les investisseurs financiers et en particulier les sociétés de gestion de portefeuille telle que Federal Finance Gestion et le groupe Arkéa Investment Services (AIS), mais aussi des enjeux plus académiques sur la dynamique des prix d'actifs financiers et la question centrale de l'efficience des marchés financiers, recentrée ici sur la classe des crypto-actifs.

Le contexte et les intérêts professionnels et académiques

Cette dualité des enjeux explique le fait que le challenge soit proposé à la fois par la société de gestion Federal Finance Gestion et le CREM, laboratoire universitaire regroupant des enseignants chercheurs de Rennes et de Caen. Ces deux entités collaborent actuellement sur un projet de recherche sur les actifs numériques dans la cadre d'une convention CIFRE (Convention Industrielle de Formation par la Recherche) associée à une thèse de doctorat en sciences économiques. Les participants au Data Challenge 2024 – Cryptomarkets sont en quelque sorte invités à participer à ce programme de recherche.

Pour Federal Finance Gestion

Federal Finance Gestion s'interroge sur le lancement de fonds d'investissement investis sur les marchés des crypto-actifs. A ce titre, la question de la prévision à court terme des prix et rendements des crypto actifs est essentielle pour savoir s'il est possible de construire des portefeuilles complètement investis, quand le marché est en phase haussière et moins exposés quand le marché est baissier. De bonnes prévisions pourraient permettre de proposer à la clientèle des fonds dédiés aux crypto actifs une

surperformance par rapport à une gestion passive de type « buy-and-hold » (acheter et garder), mais aussi, une plus faible volatilité.

En outre, Federal Finance Gestion a développé une solide expertise en finance quantitative, s'appuyant sur des modèles avancés pour anticiper les rendements futurs des actifs financiers. Cette compétence, initialement centrée sur le marché des actions, présente un potentiel significatif de transposition aux autres catégories d'actifs. Le data challenge, proposé aux étudiants de Master, offre ainsi une occasion unique de tester et d'affiner ces modèles prédictifs dans un environnement contrôlé. Les résultats obtenus pourraient non seulement enrichir notre compréhension du marché des actifs numériques, mais également ouvrir la voie à des applications similaires pour d'autres classes d'actifs financiers. Ce faisant, Federal Finance Gestion confirme son engagement dans l'innovation et la recherche de solutions d'investissement à la pointe de la technologie.

Pour les académiques

Les chercheurs du CREM voient dans la question de la prédictibilité des prix des crypto-monnaies un questionnement indirect sur l'efficacité informationnelle de ces nouveaux marchés d'actifs. La question de l'efficacité des marchés financiers est sans doute encore en 2024 la question centrale pour la finance académique.

Selon les principes de l'efficacité de marchés, toute nouvelle information importante sur les revenus futurs distribués par un actif financier, un dividende pour une action, est instantanément intégrée dans le prix de l'action. De cette manière les prix observés sont toujours très proches de leur valeur fondamentale : la valeur actualisée des revenus futurs (dividendes futurs) correctement anticipés. Les évolutions de prix à court terme (d'une heure à une autre, d'un jour à l'autre...) sont alors imputables à l'arrivée d'informations nouvelles qui viennent modifier les anticipations des investisseurs sur les revenus futurs. Mais par définition ces informations nouvelles sont totalement imprévisibles à l'avance. Sous ces hypothèses, les variations de prix et les rendements des actifs financiers sont également imprévisibles. L'exercice de prévision est en quelque sorte vain, sans objet.

La littérature académique récente portant sur les crypto-actifs a remis en cause ces principes d'efficacité informationnelle et montré que la prévision des prix à court terme restait possible sur ces marchés. Cela s'explique sans doute par le fait que pour la plupart de ces actifs ne distribuent pas de revenus. La question de la valeur fondamentale des actifs est de ce fait plus délicate et incertaine. Certaines analyses considèrent même ces actifs comme des valeurs refuges dont la valeur fondamentale serait liée à la défiance des investisseurs vis-à-vis des autres classes d'actifs. Dans ce contexte, la diffusion d'informations nouvelles peut légitimement prendre plus temps à être interprétée et intégrée dans les prix des crypto-actifs. Cette diffusion non instantanée des nouvelles informations rend l'exercice de prévision à court terme de nouveau légitime.

Les méthodes de prévision et la base de données

Toutes les méthodes de prévision sont envisageables : les modèles économétriques incluant les modèles de séries temporelles, uni et multivariés; les méthodes de Machine Learning supervisées ou non supervisées y compris les réseaux de neurones; ou d'autres types de modèles comme les algorithmes génétiques.

La base de données accessible et détaillée dans la partie 2 du document comprend des données de fréquence quotidienne sur la période août 2017-avril 2023. Elle contient des informations sur les 11 crypto-monnaies suivantes :

Bitcoin(BTC), Ethereum(ETH), Binance Coin(BNB), Litecoin(LTC), Dogecoin(DOGE), Bitcoin Cash(BCH), Ripple(XRP), Polygon MATIC(MATIC), Cardano ADA(ADA), Polka Dot(DOT), Solana(SOL)

On trouve de manière cruciale les prix de clôture (23h59 UTC) sur chacune des 11 crypto-monnaies. Ces prix de clôture vont permettre de définir les variables cibles de la prévision à court-terme. La prévision se fera avec un horizon de 1 jour.

Les variables explicatives incluses dans la base de données ont trait aux trois registres suivants.

Des données de trading (Trading data)

Les volumes quotidiens échangés sur chacune des 11 crypto-monnaies (timezone UTC). Notons ici que les séries de prix de chaque crypto-monnaie peuvent servir à prévoir la dynamique de prix des autres crypto-monnaies, si on soupçonne des interdépendances dynamiques entre les prix des crypto-monnaies.

Des données sur les prix ou rendements d'autres classes d'actifs et sur l'incertitude macroéconomique (Macrofi and uncertainty)

Dans cette rubrique, on trouve en particulier des données sur les taux d'intérêt à court et long terme sur le marché américain, sur les indices boursiers US, sur le prix de l'or, du pétrole, sur les prix des Credit Default Swap (CDS). On trouve également des indicateurs ayant à l'incertitude macroéconomique comme le VIX.

Ces séries macro-financières sont présentées avec plus de précision dans la partie II du document. Elles doivent permettre au final de décrire l'environnement macro-financier dans lequel s'effectuent les choix des investisseurs sur les différentes classes d'actifs. Elles peuvent donc orienter des prévisions sur les crypto-monnaies intégrant de possibles interdépendances avec les prix et rendements des autres catégories d'actifs.

Des données de sentiment et d'attention sur les crypto-monnaies (Sentiment and attention-Twitter)

Les données d'attention sur chaque crypto-monnaie correspondent aux nombre de Tweets (timezone [UTC](#)) mentionnant chaque jour la crypto-monnaie ;

Les données de sentiment sont obtenues à partir de méthodes de type NLP (Natural Language Processing) appliquées au contenu des Tweets. Pour chaque crypto-monnaie, on dispose un indicateur quotidien compris en -1 et 1 traduisant le degré d'optimisme des participants au réseau Twitter sur la progression des prix : plus l'indicateur est élevé plus les messages sont analysés comme optimistes (haussiers).

Précisons enfin que les équipes ne sont pas tenues de proposer un exercice de prévision sur l'ensemble des 11 crypto-monnaies. Elles pourront, à leur discrétion, selon leur progression et la pertinence des modèles proposés, se concentrer sur un panier plus restreint mais incluant au minimum 5 crypto-monnaies.

Dans la suite du document se trouve la deuxième partie qui détaille les points techniques et précise les attentes. Un annexe complète le document en apportant des informations utiles sur les packages Rstudio et Python, les indicateurs de performances prédictives.

Rappel !!!

Le jeu de données est confidentiel et est réservé strictement à l'usage des participants du concours. Ce sont des données exploitées dans le cadre d'une thèse. Il est demandé de ne pas conserver non plus de copie après la fin du challenge.

Tous les éléments nécessaires seront déposés sous un git accessible pour faciliter le téléchargement et l'utilisation par tous:


git: <https://github.com/rennesdatascience/datachallenge2024>

II-Les aspects techniques du challenge et quelques points de vigilance

Après une présentation de la base de données à la première section, il est précisé les attentes dans la deuxième section, suivies des critères qui permettront une évaluation objective des différentes équipes.

II.1.- Base de données

Le jeu de **données de modélisation** se trouve ici:

 `Rennes_DataChallenge2024_Cryptomarkets_dataset`

La description des différentes colonnes se trouve dans un fichier description ici:

[+ Rennes_DataChallenge2024_Cryptomarkets_database_description](#)

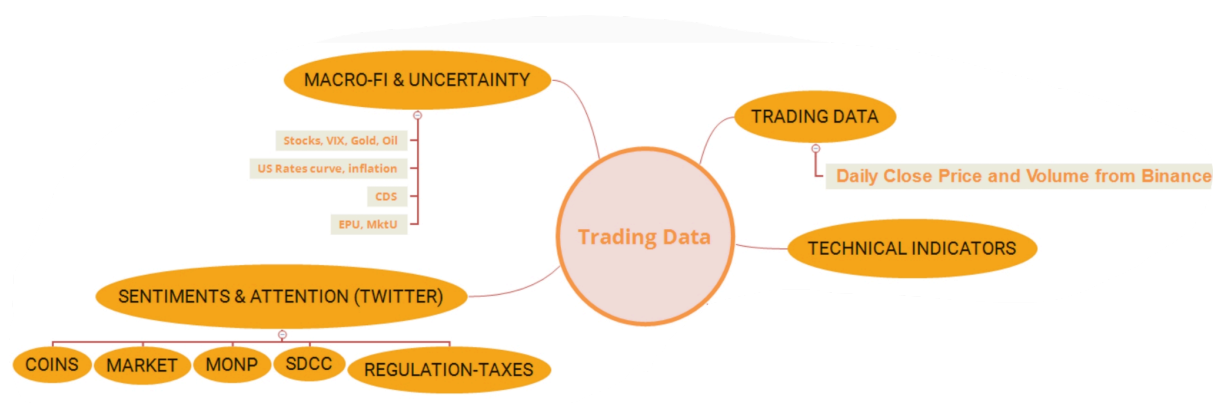
On s'intéresse tout particulièrement à l'étude du pouvoir prédictif des indicateurs de sentiments (polarité des tweets) et d'attention (volume de tweets) par rapport au marché des crypto-monnaies. La base de données a été construite en conséquence. Des données macroéconomiques liées à l'économie américaine, des données financières concernant d'autres classes d'actifs (Equity US, CDS...) ont été ajoutées, en plus des données de prix et de volume des principales crypto-monnaies.

Ce sont des données temporelles allant de 08-2017 à 04/2023 dont :

- la période de **08-2017 à 08-2022** est réservée à la modélisation;
- la période de **09-2022 à 04-2023** sera utilisée pour le test (prévisions hors échantillon) afin d'évaluer la performance des modèles prédictifs de chaque équipe.

Concernant les données de Trading de crypto-monnaies provenant de Binance, on considère les données Crypto/USDT (au lieu de Crypto/USD ou Crypto/EUR) car ces transactions font l'objet de plus de volume donc de suffisamment de liquidité.

a. Composition de la base de données



On n'impose aucune restriction quant au choix des variables. Chaque équipe décide d'intégrer ou pas des données de sentiments et d'attention, des données macro-financières, des données de trading d'autres crypto-monnaies. Chaque équipe est libre de créer de nouvelles variables, ou de transformer les données selon les besoins de son approche.

Les données de trading incluent les séries de prix de clôture et volume pour les crypto-monnaies suivantes:

No	Code	Crypto-monnaie	Début training set
1	BTC	Bitcoin	08-2017
2	ETH	Ethereum	08-2017
3	BNB	Binance Coin	08-2017
4	LTC	Litecoin	08-2017
5	DOGE	Dogecoin	11-2019
6	BCH	Bitcoin Cash	11-2019
7	XRP	Ripple	05-2018
8	MATIC	Polygon MATIC	05-2019
9	ADA	Cardano ADA	03-2020
10	DOT	Polka Dot	08-2020
11	SOL	Solana	08-2020

b. Définition et sens

Pour une description plus détaillée de la base de données, on se réfère à ce fichier de description:

 [Rennes_DataChallenge2024_Cryptomarkets_database_description](#)

Ci-dessous quelques-uns des indicateurs:

No	Catégorie	Données/Indicateur	Définition	Interprétation
1	MacroFi	VIX	Volatility Index of US Equity SP500	Mesure d'incertitudes ou de risques sur les marchés actions américains. C'est la volatilité implicite associée au prix des options du SP500.
2	MacroFi	USEPUINDEXD	US Economic Policy Uncertainty	Indicateur d'incertitude de politique économique aux Etats-Unis.
3	MacroFi	FedSF_NewsSentiment	Federal Reserve of San-Francisco Daily Sentiment	

				<i>Indicateur de confiance sur l'économie américaine construit à partir d'articles de presse économique.</i>
4	MacroFi	ECRPUS1YIndex	<i>Probabilité de Récession à l'horizon d'un an de l'économie américaine</i>	<i>Probabilité de Récession à l'horizon d'un an de l'économie américaine</i>
5	MacroFi	CDS	<i>Credit Default Swap</i>	<i>Prix des contrats de couverture contre le risque de défaut des obligations souveraines. C'est la perception du risque de défaut par les marchés.</i>
6	MacroFi	SP500, Nasdaq	<i>US Equity Index</i>	<i>Indice de valorisation des marchés actions.</i>
7	Sentometrics	btc_nlp_sentiment_mean	<i>Valeur moyenne de polarité (calculée par nltk) des tweets relatifs au Bitcoin (prix, volume, fluctuation, prédiction, incertitude) en anglais, retweet exclu, postés par une liste de personnalités influentes.</i>	
8	Sentometrics	btc_tweet_count/btc_volume_tweets	<i>Volume de Tweets relatifs à Bitcoin: en anglais, retweet exclu, posté par des profils spécifiques.</i>	
9	Sentometrics	cryptomkt_uncertainty_attention	<i>Volume de tweets sur le marché global des crypto-monnaies avec les termes incertitude, crash, baisse, forte volatilité, crise; par des comptes vérifiés.</i>	
10	Sentometrics	sdcc_vader_sentiment_composite_max	<i>Valeur maximale sur une journée de polarité composite calculé par VADER à partir des tweets relatifs à la dette souveraine ou dévaluation des devises-monnaies</i>	

“La polarité dans l'[analyse des sentiments](#) fait référence à l'identification de l'orientation des sentiments (positif, neutre et négatif) dans un langage écrit ou parlé. ”, DataFranca.org

c. Valeurs manquantes

Le modélisateur se doit de faire attention à la structure des données, notamment les données manquantes :

a. Données manquantes sur des séries de prix et de volume:

Ne pas faire d'imputation !!!

Les différentes crypto-monnaies n'ont pas été créées à la même date. La date de début peut-être impactée par d'autres paramètres liés à la création de Binance et de ses services aussi. Il serait judicieux de créer des groupes (pool de crypto-monnaies ayant la même date de début) pour faciliter la manipulation des données.

b. Données manquantes sur des séries macroéconomiques et financières:

Il y a peu de données manquantes sur ces séries. Les valeurs manquantes sont dues essentiellement aux jours non ouvrés des bourses des classes d'actifs traditionnels (jour de fermeture de la bourse US). Des imputations sont possibles.

c. Données manquantes sur des séries de sentiment et d'attention:

- séries de données d'attention (volume de tweets): pas de données manquantes
- séries de données de sentiment (polarité des tweets): des données manquantes selon la crypto-monnaies.

Ces données manquantes sont du fait de la construction de la requête d'extraction. En particulier, sur Bitcoin on a limité les tweets seulement à une liste fermée d'utilisateurs. Quand rien n'a été posté par ces utilisateurs en lien au marché (prix, volume, tendance, etc) de la crypto-monnaies, cela conduit à des valeurs manquantes. Pour d'autres crypto-monnaies où il n'y a pas eu de restriction sur l'auteur du tweet, il n'y a pas de valeurs manquantes.

Des imputations sont nécessaires !!!

Attention à ne pas s'efforcer à gérer des valeurs manquantes là où il n'y a guère le besoin : pour des dates où l'on ne dispose pas des données de Trading.

Sur RStudio, les fonctions `zoo::na.locf()` ou `DMwR2::knnImputation()`, prêtent facilement à la gestion des valeurs manquantes. Sur python la librairie [scikit-learn](https://scikit-learn.org/) propose des méthodes équivalentes avec `sklearn.impute::KNNImputer()` ou `sklearn.impute::SimpleImputer()`.

```
'''
>>> import numpy as np
>>> from sklearn.impute import KNNImputer
>>> nan = np.nan
>>> X = [[1, 2, nan], [3, 4, 3], [nan, 6, 5], [8, 8, 7]]
>>> imputer = KNNImputer(n_neighbors=2, weights="uniform")
>>> imputer.fit_transform(X)
array([[1. , 2. , 4. ],
       [3. , 4. , 3. ],
       [5.5, 6. , 5. ],
       [8. , 8. , 7. ]])

'''

'''
>>> import numpy as np
>>> from sklearn.impute import SimpleImputer
>>> imp = SimpleImputer(missing_values=np.nan, strategy='mean')
```



```

>>> imp.fit([[1, 2], [np.nan, 3], [7, 6]])
SimpleImputer()
>>> X = [[np.nan, 2], [6, np.nan], [7, 6]]
>>> print(imp.transform(X))
[[4.         2.         ]
 [6.         3.666...]
 [7.         6.         ]]
'''

```

d. Transformation & Factor Engineering

La transformation des variables ou la construction de nouvelles ne constituent pas une nécessité absolue pour la construction de modèles de prévision. Cependant, manquer à la transformation alors que c'est nécessaire peut réduire la précision du modèle. Par ailleurs, le cadre théorique de certaines approches peut l'imposer. Parfois, il peut s'agir d'une bonne pratique tout simplement.

Ci-dessous quelques exemples en guise d'aide aux non-initiés à la finance empirique.

Transformation

1) Transformation pour besoin de stationnarité

Le cadre théorique de modélisation économétrique en finance est fondé sur des données stationnaires. En général, on modélise et prédit le rendement d'un actif (au lieu de son prix directement); puis on revient facilement au prix si besoin.

Si P_t est le prix d'un actif (niveau d'un indice, prix du Bitcoin/USDT, niveau du SP500).

Alors on calcul le rendement qui est stationnaire par la formule:

$$\text{rendement}_t = \ln\left(\frac{p_t}{p_{t-1}}\right) = \ln(p_t) - \ln(p_{t-1})$$

Quand il ne s'agit pas de prix d'actifs, ou si les valeurs de la série x_t ne sont pas toujours positives alors on peut stationnariser par le calcul du pourcentage de variation:

$$\text{Si } \min(x)_{1 \leq t \leq T} < 0, \text{ alors } \% \text{variation}(x_t) = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1.$$

Si la série peut prendre des valeurs nulles, alors la stationnarité s'obtient par simple différence comme

$$\Delta x_t = x_t - x_{t-1}$$

En principe, une série de taux se stationnarise par simple différence première.

2) Certains algorithmes performant mieux avec des données standardisées (normalisées ou uniformisées)

Ce serait le cas pour le Random Forest quand il est appliqué à des données temporelles de nature macroéconomique. Ainsi que certaines approches de réduction de dimension comme les ACP.

Normalisation:

$$Z_t = \frac{x_t - \text{moyenne}(x)}{\sigma(x)}$$

Uniformisation:

$$Z_t = \frac{x_t - \min(x)}{\max(x) - \min(x)}$$

Certains algorithmes offrent la possibilité de spécifier via un argument si les données doivent être standardisées.

Construction de nouvelles variables

1) Variables retardées (lag, shift)

Une des caractéristiques des séries de prix vient du fait de la corrélation du prix courant au prix passé (le prix de demain = prix d'aujourd'hui + une composante stochastique ou aléatoire). En conséquence, on peut avoir une auto-corrélation de la série des rendements avec les prix passés qui permettent de prédire le prix courant.

L'enjeu est de déterminer le nombre maximal de retards jugé pertinent pour la prédiction.

Variable retardée de p date = $z_t(p) = x_{t-p}$

2) Variables d'écart (delta)

On peut aussi créer des variables d'écart. Entre autres pour deux(2) raisons.

- Réduire le nombre de prédicteurs (simplifier le modèle) sans perdre le contenu informationnel.
- La variable d'écart a un sens à elle, une interprétation propre. C'est le cas des spreads de taux (Taux 1 an - taux 1 mois, Taux 5 ans - Taux 2 ans; Taux 10 ans - Taux 2 ans, à titre d'exemple).

Variable d'écart = $z_t = x_t - y_t$

3) Variables de trend : moyenne mobile (moving average)

En bourse, l'analyse chartiste consiste à simplement utiliser des indicateurs techniques de moyenne glissant sur p dates pour prédire les prix des actifs.

- *Simple Moving Average (SMA)*
- *Exponentially Weighted Moving Average (EMA)*
- *Momentum*
- *RSI, etc.*

Chacun est libre de se documenter en ligne sur les indicateurs techniques pressentis comme pertinent pour les marchés de crypto-monnaies.

Attention !!!

Ne pas trop se focaliser sur la création de nouvelle variable inutilement. Ni de choisir une fenêtre large, car les séries ne sont pas très longues.

II.2.- Attendus

a. Prévvision de Rendement à Court Terme

On attend une prévision à très court terme des prix/rendements des crypto-monnaies, avec un accent sur les tendances du marché. Il est demandé aux participants de réaliser une prévision sur un horizon de 1 jour.

Le nombre de crypto-monnaies considérées pour la prévision sera de 5 au minimum.

b. Période d'Observation et d'Apprentissage

Il faut utiliser les données historiques de 08-2017 à 08-2022 pour l'apprentissage. Les prévisions se feront de 09-2022 jusqu'en 04-2023.

Les critères quantitatifs d'évaluation de la performance du modèle prédictif sont détaillés dans l'Annexe A.d.

c. Construction et Performance d'un Portefeuille d'Investissement

Nous proposons de manière optionnelle la réalisation d'un backtest financier consistant à simuler une gestion active de portefeuille fondée sur les modèles de prévision développés.

Construire un portefeuille basé sur des prévisions hors échantillon, avec des stratégies d'investissement incluant des positions longues et neutres.

Démarrer avec une base 100 ou un portefeuille de 10 000€.

d. Stratégie d'Investissement

Il s'agit d'une stratégie simple qui tire profit de la capacité du modèle à prédire le directionnel du marché appliquée à un portefeuille simple de deux (2) actifs: une crypto-monnaie et un actif sans-risque. L'actif sans risque ne distribue pas de rendement ($r_f = 0$). La stratégie est implémentée/répliquée pour chaque crypto-monnaie considérée dans le panier. Les règles d'allocation sont simples:

- A la date de départ (31/08/2022), 0% du portefeuille est investi sur la crypto-monnaie et 100% sur l'actif sans risque.
- En cas de prévision haussière ($\hat{r}_{t+1} > 0.5 * \sigma$): on augmente l'exposition sur la crypto-monnaie à 100%. Sigma représente l'écart-type des rendements quotidiens de la crypto-monnaie sur l'échantillon d'apprentissage.
- En cas de prévision baissière ($\hat{r}_{t+1} < -0.5 * \sigma$): on réduit l'exposition sur la crypto-monnaie de 50%. Si elle était à 100%, l'exposition passe à 50% de 100% c'est-à-dire 50% de la valeur du portefeuille. Si elle était de 50%, elle passe à 25% et ainsi de suite.
- En cas de prévision neutre ($|\hat{r}_{t+1}| \leq 0.5 * \sigma$): on ne modifie pas la composition du portefeuille.

Les frais de transaction sont nuls par hypothèse. Pas de vente à découvert, donc pas d'emprunt.

Une illustration de backtest de stratégie d'allocation est donnée dans le fichier ci-dessous:

 rennes_datachallenge2024_output_format.xlsx

En toute date t de la période de backtest, on détermine l'exposition α la part de l'actif investie dans la crypto-monnaie en fonction de la prévision de rendement \hat{r}_t . La valeur du portefeuille en fin de journée se calcule facilement en fonction du rendement observé r_t sur les marchés par :

$$\text{Valeur portefeuille}(VL)_t = VL_{t-1} \times [1 + \alpha_t \times r_t]$$

e. Sorties et résultats attendus

On précise que chaque équipe doit fournir:

1. Leurs codes pour vérification, si besoin.
2. La présentation dont le support démontre le travail produit et la compréhension de la problématique.
3. Un fichier de sortie récapitulant les principaux résultats dont:
 - a. les prévisions hors échantillon à 1 jour, pour le panier choisi de crypto-monnaies;
 - b. les indicateurs de performance des prévisions économétriques et de la stratégie d'allocation, sur le même panier de crypto-monnaies;
 - c. pour chaque crypto-monnaie, le graphique de comparaison des rendements observés et prédits;

- d. *optionnellement, le graphique de comparaison en termes de prix;*
- e. *la composition de chacun des portefeuilles mono-crypto, pour vérification si nécessaire;*
- f. *le graphique de la Valeur du portefeuille contre la série de prix base 100 de la cryptomonnaie.*

Une équipe peut choisir de travailler avec les prix ou les rendements selon sa stratégie empirique de modélisation. Si la prévision porte sur le prix, les prévisions de prix seront converties en prévisions de rendements.

Un exemple de fichier de sortie est proposé sous ce format:

 `rennes_datachallenge2024_output_format.xlsx`

II.3.- Critères d'évaluation

Comme chaque équipe aura à prédire les rendements de plus d'une crypto-monnaie. La meilleure performance sera considérée pour chaque équipe. En d'autres termes, les équipes seront évaluées par rapport à leur meilleur modèle.

L'évaluation se fait selon trois niveaux dont : (a) la présentation orale de la solution; (b) la performance prédictive des modèles empiriques de prévision; (c) l'exercice de backtest de stratégie d'allocation.

Les indicateurs tels les RMSE et le taux de bonne prévision sur le sens (signe du rendement, hausse/baisse) d'évolution de la cryptomonnaie seront considérés pour départager les équipes sur une même crypto-monnaie. S'il est besoin de comparer des modèles sur deux(2) crypto-monnaies différentes, le MPE se prête à une telle utilisation.

Le Ratio de Sharpe sera retenu comme mesure de performance du backtest de stratégie d'allocation. A Ratio de Sharpe égal, le drawdown permettra de différencier les équipes.

Ce fichier de script, en langage R, illustre le calcul des indicateurs d'évaluation de prévisions hors échantillon "[rennes_datachallenge2024_cryptomarkets_evalpred.html](#)"

[script R: évaluation de performance hors échantillon](#)

III-Annexe

Annexe A.- Modélisation Prédictive en finance

Soit y la variable dépendant et z (un sous-ensemble de x) les variables explicatives.

Le modèle (1): $y_t = g(f(z_t)) + e_t$ est un modèle explicatif.

Le modèle (2): $y_t = g(f(z_{t-1})) + e_t$ est un modèle prédictif.

Dans le modèle prédictif, toute l'information nécessaire à la prédiction pour la période d'après est disponible à la période courante. On parle de "modèle prédictif réalisable" car on peut l'utiliser pour la prévision.

Le modèle (1) n'est pas réalisable car il faudrait prédire chaque variable de z à chaque horizon, avant de pouvoir prédire la valeur de la cible y . C'est ce qui confère leur popularité en macroéconométrie aux modèles ARIMA(p,d,q):

$y_t = a + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t + \dots + \psi_{t-q} \epsilon_{t-q}$ où la valeur de la cible à l'instant t dépend uniquement des valeurs passées.

On s'intéresse à un modèle prédictif !!!

Il est acceptable de proposer un modèle quasi-réalisable où 1 seule variable (v) des variables explicatives n'est pas retardée: cela laisse la porte ouverte à des simulations, des prédictions conditionnelles à un scénario sur v .

a. Cible de prévision

- *Prévisions d'une valeur, du niveau : prévision du rendement du Bitcoin*
- *Prévision de quantile ou d'intervalle de confiance: prévision de la Value-at-Risk à 1 jour*
- *Prévision du régime ou de la tendance, de la direction: tendance haussière ou baissière à $p\%$ de probabilité*
- *Prévision des points de retournement: anticiper le retournement vers un marché haussier ou baissier*

Le modèle prédictif dépend de la cible de prévision. Ainsi que les critères de performance.

b. Sélection des variables pertinentes

Si X (matrice de N variables explicatives) est donnée: l'ensemble d'information.

- Il est à considérer le nombre de prédicteurs potentiels : $M \leq N$.
- Si M est grand: il faut considérer la limite du modèle. En économétrie, on limite le nombre de prédicteurs. En Machine Learning, on se soucie moins de la dimension de M .
- On arbitre en quantité et qualité de X : arbitrage complexité-performance.
- On sélectionne par jugement, l'expérience, ou l'expertise. Dans certains cas, une première régression pénalisée peut aider. Ou des approches de réduction de dimensions comme les ACP.
- Le modèle prédictif qu'on peut établir dépend essentiellement des variables prédictives à disposition.
- On peut procéder par sélection par étape (stepwise selection) en utilisant des fonctions dédiées.
- Il reste que le principe pour rajouter une variable serait de choisir celle qui est la plus corrélée à l'erreur estimée $\hat{\epsilon}$.
- En série temporelle, étant donné le besoin de variables retardées pour un modèle prédictif réalisable, on utilise souvent les ACF, PCF et CCF pour choisir les variables explicatives (en lieu et place des corrélations).

Un plus grand nombre de prédicteurs ajoute à la complexité du modèle !

Le but de cette étape, quand elle a lieu, est de proposer in fine un modèle qui soit répliquable ou vendable: le plus performant possible, le plus simple possible, le plus intuitif possible et répliquable.

Cela rejoint la théorie de l'efficience informationnelle :

1. *efficience faible : il n'est pas possible de prédire y_{t+1} seulement en se basant sur les valeurs passées de y (y_t, y_{t-1}, y_{t-p}). Une stratégie purement chartiste ne peut être tout le temps gagnante.*
2. *efficience semi-forte : il n'est pas possible de prédire y_{t+1} en se basant sur les valeurs passées de y (y_t, y_{t-1}, y_{t-p}) et d'autres variables (z_t, z_{t-1}, z_{t-q}). Une stratégie purement fondamentale ne peut être tout le temps gagnante. Ni la combinaison chartiste+fondamentale, d'ailleurs.*
3. *efficience forte : il n'est pas possible de prédire y_{t+1} en se basant sur une connaissance non publique. On ne peut pas manipuler le marché. Les prix s'ajustent au fur et à mesure de l'arrivée des informations de telle sorte que toute l'information disponible soit intégrée dans le prix. Le prix est le seul élément clé car il reflète toute l'information disponible.*

c. Spécification du modèle

On retient que:

- Tous les modèles sont faux. Aucune pleine et entière certitude sur le modèle exact (Data Generating Process).
- Alors on peut tester plusieurs modèles et choisir, ou les combiner.
- Le modèle le plus performant en intra-échantillon (training) peut ne pas l'être en hors-échantillon (test).
- Alors on évalue en continue tout modèle prédictif. On décide en connaissant les forces et les faiblesses.

En général, on utilise ces critères pour choisir la spécification du modèle

→ R^2

- ◆ *Plus c'est élevé, mieux c'est. Mais on y reste pas forcément attaché en finance empirique. En général, les modèles de prévision des rendements (ou prix) d'actifs ont le $R^2 \sim 5\%$. Obtenir un modèle à plus de 5% de R^2 comme objectif intermédiaire n'est donc pas dénué de sens, tant que ce n'est pas au prix d'un grand nombre de prédicteurs (modèle fallacieux).*

- **Akaike AIC** : maximise la vraisemblance que compte tenu des observations, le modèle estimé est le plus probable.
- Hannan and Quinn :
- **BIC** : maximise la vraisemblance que le modèle g soit exact, c'est-à-dire le vrai modèle qui a généré les données observées.
- Likelihood ou Log-Likelihood: Équivalent à R^2 à sa façon quand c'est l'EMV.
- AUC ou précision: si la cible est la prédiction de la tendance, la direction, ou les retournements.

On s'impose un seul critère d'information pour comparer différents modèles du même type !!!

A propos du R^2 faible en finance empirique, cela corrobore l'hypothèse de rendements non prévisibles et une dynamique des prix conforme à une marche aléatoire avec dérive.

Benchmark Model 1 – Random Walk: $\text{rendement}_t = \text{constant} + w_t$

avec $\text{constant} = 0$ ou la moyenne non conditionnelle, $w_t =$ variable aléatoire centrée

Le premier critère d'un bon modèle prédictif c'est de faire mieux (surperformer) les prévisions d'un modèle de marche aléatoire.

Facteurs déterminants:

En cas de:

- changement structurel
- de non-linéarité
- présence de régime

Il convient d'adapter son approche car ces caractéristiques impactent la stabilité du modèle et par conséquent la précision des prévisions.

Tout le long du processus, on garde en tête que l'erreur de prévision:

$$\begin{aligned} \hat{\epsilon}_{t+h} &= y_{t+h} - \hat{y}_{t+h} = g * (f * (z_t)) - g(f(z_t)) + g(z_t) - \hat{g}(z_t) + \epsilon_{t+h} \\ &+ \text{ Erreur d'approximation : } g * (f * (z)) - g(f(z)) \\ &+ \text{ Erreur d'estimation : } g(z_t) - \hat{g}(z_t) \\ &+ g * : \text{ la vraie relation entre } y \text{ et } f(z), \text{ linéaire, quadratique, logarithmique} \\ &+ f * : \text{ transformation nécessaire de } z^1 \\ f &: \text{ la transformation appliquée en réalité à } z \\ &+ g : \text{ la relation choisie en réalité} \end{aligned}$$

L'erreur d'approximation combine l'erreur de jugement quant à la relation entre y et ses prédicteurs, et les transformations des variables. L'erreur d'estimation vient de l'ensemble z des variables retenues comme prédicteurs.

Savoir quand transformer ou normaliser des variables est capital !!! Cette étape se fait simultanément ou suite au choix des prédicteurs.

On réduit l'erreur d'estimation en choisissant les variables pertinentes. On réduit l'erreur d'approximation par son jugement des relations entre y et z, et par la transformation des variables, si nécessaire.

Pour intégrer la non linéarité, la présence de régimes ou de changements structurels, l'économétrie fait appel à la complexification (ajout de variable, des modèles plus complexes, etc). Le Machine Learning peut s'imposer en surpassant les modèles économétriques classiques. A titre d'exemple:

- *Les Random Forests gèrent naturellement ces comportements sous condition de bonne calibration.*
- *Les réseaux de neurones captent toute forme de non linéarité avec le moins de paramètres à estimer que possible.*

En économétrie, habituellement on se base sur la performance intra-échantillon via les critères d'information (BIC) pour sélectionner les modèles. En Machine Learning, la sélection repose sur la validation croisée, c'est-à-dire en tenant compte de la qualité des prévisions hors échantillon. Ce qui explique la présence des fonctionnalités de validation croisée dans les algorithmes de ML.

d. Evaluation de la performance du modèle prédictif

- **Evaluation sur l'échantillon d'apprentissage (intra-échantillon en séries temporelles)**

Le modèle peut répliquer parfaitement les observations sur l'échantillon d'apprentissage:

- + précision élevée
- + erreur faible

¹ exemple de transformation : en différence, en % de variation, normalisation, discrétisation, moyenne mobile, etc

- **Evaluation sur l'échantillon de validation**

En général, on sépare le jeu de données en jeu d'apprentissage (training set p%) et jeu de validation (validation set 1-p%). Cela permet de valider la capacité prédictive des modèles et de décider lequel choisir en fonction de l'arbitrage biais-variance et d'éviter le surajustement (overfitting).

Sur des données en coupe transversale, on s'habitue à échantillonner (resampling) pour optimiser l'apprentissage d'un modèle de ML. Ceci est plus difficile d'application en série temporelle.

On s'attend à des prévisions:

1) **Optimales:**

- a) *Le modèle ne donne pas systématiquement des sous-prévisions (sur-prévisions). La série prédite fluctue autour de la série observée.*
- b) *Graphique comparatif des deux séries.*
- c) *Aucune classe n'a 0% de chance d'être prédite*
- d) *Matrice de confusion.*

2) **Précises**

- a) *Valeurs prédites proches des valeurs observées. En particulier, le signe de la valeur prédite est le même que celui de la valeur observée, le cas échéant.*
- b) *Des mesures de précisions comme MSFE (Mean Squared Forecast Error), RMSFE, etc.*
- c) *Probabilité prédite soit la plus proche de 1 possible (maximiser la certitude quand on a raison).*
- d) *AUC*

3) **De faible biais**

- a) Statistique U-Theil 1

$$U_1(\text{accuracy}) = \frac{\left[\sum_{h=1}^H (\text{Prévision}_{t+h} - \text{Observation}_{t+h})^2 \right]^{1/2}}{\left[\sum_{h=1}^H (\text{Observation}_{t+h})^2 \right]^{1/2}}$$

- b) Statistique U-Theil 2

$$U_2(\text{quality}) = \frac{\left[\frac{1}{H} \sum_{h=1}^H (\text{Prévision}_{t+h} - \text{Observation}_{t+h})^2 \right]^{1/2}}{\left[\frac{1}{H} \sum_{h=1}^H (\text{Observation}_{t+h})^2 \right]^{1/2} + \left[\frac{1}{H} \sum_{h=1}^H (\text{Prédiction}_{t+h})^2 \right]^{1/2}}$$

- c) Mean Percentage Error

$$MPE = 100 * \left[\frac{1}{H} \sum_{h=1}^H \left(\frac{\text{Obs}_{t+h} - \text{Prev}_{t+h}}{\text{Obs}_{t+h}} \right) \right]$$

- d) Mean Absolute Percentage Error

$$MAPE = 100 * \left[\frac{1}{H} \sum_{h=1}^H \left| \frac{\text{Obs}_{t+h} - \text{Prev}_{t+h}}{\text{Obs}_{t+h}} \right| \right]$$

- e) Si la cible est discrète (tendance ou direction): Quadratic Probability Score

$$QPS = \frac{1}{H} \sum_{h=1}^H \sum_{k=1}^m (D_{k,t+h} - P_{k,t+h})^2, \quad QPS = 0, \text{ modèle parfait}$$

- f) Prédiction du sens/tendance.

En considérant le signe des rendements observés et le signe des rendements prédits. Il s'agit d'évaluer le nombre de signes correctement prédit.

$$\text{Prévision Signe} = \frac{S_{++} + S_{--}}{H}, \text{ nombre de dates de la période de prévision.}$$

$$\text{Prévision Direction Haussière} = \frac{S_{++}}{H_{+}}, \quad H_{+} \text{ nombre de dates de signe positif.}$$

$$\text{Prévision Direction baissière} = \frac{S_{--}}{H_{-}}, \quad H_{-} \text{ nombre de dates de signe négatif.}$$

- 4) ²Sans surajustement donc de variance raisonnable

- a) *La performance sur l'échantillon de validation (hors-échantillon) ne soit pas complètement dérisoire au regard de la performance sur l'échantillon d'apprentissage (intra-échantillon).*

- 5) Meilleures que celles des modèles naïfs (Marche aléatoire ou ARMA simple)

- a) Statistique de test: [Diebold-Marino](#)
b) Utile pour décider le choix entre 2 ou plusieurs modèles prédictifs

U-Theil < 1, le modèle donne de meilleures prévisions que la marche aléatoire. U-Theil = 1, identique. U-Theil > 1, le modèle produit des prévisions pires que la roulette russe ne saurait le faire. On cherche dans l'idéal, s'il existe, un modèle prédictif tel que U-Theil << 1.

En définitif, on peut combiner plusieurs modèles prédictifs (encompassing=enveloppement). Le bagging, le boosting, le random forest comme toute approche ensembliste intègre nativement l'enveloppement.

- **Évaluation sur l'échantillon de test** : les équipes ne disposent pas de l'échantillon de test avant la fin du temps réglementaire pour la mise en place du modèle prédictif.
 - On va se focaliser sur : (a) l'optimalité, (b) la précision, (c) l'incertitude -QPS/MAPE, (d) la supériorité au modèle naïf.
 - La robustesse

² Incertitude limitée autour de la prévision

Annexe B.- Rstudio - Support

Packages	Utilité	Exemple de fonctions
xlsx, readxl, stargazer	<i>Read-write excel, format model output</i>	
xts, zoo, tseries, tseriesChaos	<i>Manipulate and Modeling of time series</i>	zoo::nz.locf(), zoo::rollmean()
lubridate, glue, car, Hmisc, forcats, DMwR2, ramify, stringr, reshape2	<i>Data manipulation and check</i>	DMWR2::knnImputation()
ggplot2, dygraphs	<i>Advanced Plotting</i>	
forecast, dynamac, vars, tsDyn, urca, dLagM, forecastHybrid	<i>Time series modeling and forecasting</i>	forecast::accuracy(), tsDyn::TVAR(), tsDyn::TVECM(), dynamac::dynardl()
FactoMineR, pracma, flexmix, mclust	<i>Factorial Analysis + Clusterwise Linear Regression + Gaussian Mixture Modeling for Model-Based Clustering</i>	
pROC,	<i>AUC, ROC, multi AUC</i>	
glmnet, randomForest, randtoolbox, ipred, MASS, ada, xgboost, gbm, e1071, caret, ranger, MacroRF	<i>Penalized Regression(Lasso, Ridge, Elastic-Net) + Ensemble methods (Bagging, Boosting, Random Forest, etc)</i>	glmnet::glmnet()
garchx, tvgarch, rugarch,	<i>Econometric Garch, dynamic Garch</i>	
MSTest, MSGARCH, MSwM	<i>Markov Switching Test and Regression</i>	
arules, arulesCBA, arulesViz	<i>Association Rules</i>	
torch, doParallel, foreach, gridextra, nnet	<i>Neural Network + Parallel estimation</i>	torch::nn_gru(), nn_lstm(), nn_gelu()
PerformanceAnalytics, quantmod, TTR	<i>Financial Perf and Risk Analysis</i>	CAPM.dynamic(), TTR::momentum()

Liste non exhaustive des packages de RStudio pour la modélisation et la prévision.

Annexe C.- [Python - Support](#)

Packages	Utilité	Exemple de fonctions
openpyxl, numpy, pandas	<i>Data Manipulation</i>	pd.DataFrame.pct_change()
math	<i>Basic Math Functions</i>	
matplotlib, plotly, ggplot	<i>Plotting</i>	
scipy	<i>Scientific Computing</i>	
statsmodels	<i>Statistics & Econometrics</i>	https://pypi.org/project/statsmodels/
MacroRF	<i>Macroeconomic Random Forest & Financial Backtesting</i>	
scikit-learn	<i>All ML methods</i>	https://scikit-learn.org/stable/user_guide.html
TensorFlow, pytorch	<i>Neural Network</i>	
talib/TA-Lib	<i>Financial Chartist</i>	

Liste non exhaustive des packages python pour la modélisation et la prévision.

Annexe D.- Bon à savoir sur la modélisation prédictive

Disentangling Machine Learning Features : A Preview

Goulet Coulombe, Leroux, Stevanovic and Surprenant (2022, JAE)

- Evaluate marginal effects of \mathcal{G} , $pen()$, τ and L . $f_Z = I_{N_Z}$, f_y : average annual growth rate
 - (Nonparametric) Nonlinearity is the most salient feature
 - Factor modelling outperforms alternative regularization
 - K-fold cross-validation reliable
 - Quadratic loss very resilient

Goulet Coulombe, Leroux, Stevanovic and Surprenant (2021, IJF)

- Evaluate marginal effects of f_Z and f_y with \mathcal{G} and $pen()$. Fix τ and L .
 - (f_Z) Moving average rotations of the data helps
 - Factors and nonlinearity boost the performance
 - (f_y) Path averaging works

I - Les grandes lignes du challenge : contexte, enjeux professionnels et académiques.

1

II-Les aspects techniques du challenge et quelques points de vigilance..... 4

II.1.- Base de données.....	4
a. Composition de la base de données.....	5
b. Définition et sens.....	6
c. Valeurs manquantes.....	7
d. Transformation & Factor Engineering.....	9
Transformation.....	9
Construction de nouvelles variables.....	10
II.2.- Attendus.....	11
II.3.- Critères d'évaluation.....	13

III-Annexe..... 14

Annexe A.- Modélisation Prédictive en finance.....	14
a. Cible de prévision.....	14
b. Sélection des variables pertinentes.....	15
c. Spécification du modèle.....	15
d. Evaluation de la performance du modèle prédictif.....	17
Annexe B.- Rstudio - Support.....	20
Annexe C.- Python - Support.....	21
Annexe D.- Bon à savoir sur la modélisation prédictive.....	22