ML regressor

tilian bourachot

April 2025

1 Introduction

Dans cette partie, notre objectif est de **prédire les cinq prochaines valeurs horaires du prix de l'USDC**, en s'appuyant uniquement sur les observations passées. Ce type de prédiction à court terme peut s'avérer particulièrement utile dans des contextes de surveillance de marché ou de détection précoce d'instabilité.

Nous avons choisi de nous concentrer sur l'**USDC** car, contrairement à d'autres stablecoins, ses prix présentent une série relativement **stationnaire**, ce qui rend son comportement plus prévisible. Cette propriété, confirmée par une analyse préliminaire via des modèles ARMA, nous permet d'utiliser des modèles de régression sur les prix sans transformations complexes.

2 Variables utilisées

À partir des prix journaliers de l'USDC (données issues de Binance), nous construisons un ensemble de **variables explicatives** destinées à alimenter les modèles de prédiction. Ces variables sont extraites à partir d'une série temporelle de prix en utilisant des transformations classiques.

Les types de variables utilisées sont les suivants :

- **Retards (lags)** : les 10 dernières valeurs connues du prix sont intégrées sous la forme de variables `lag_1`, `lag_2`, ..., `lag_10`, afin de fournir au modèle un historique immédiat de l'évolution du marché.
- **Statistiques roulantes** :
  - Moyenne mobile sur 3 pas (`roll_mean_3`) et sur 5 pas (`roll_mean_5`), qui permettent de lisser le signal et d'identifier des tendances locales ;
  - Écart-type mobile sur 3 pas (`roll_std_3`) et sur 5 pas (`roll_std_5`), donnant une mesure de la volatilité locale du prix.

Une fois ces variables construites, les observations incomplètes (issues des opérations de décalage et de lissage) sont supprimées. Le jeu de données est ensuite découpé en deux parties :

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	\
10	1.000054	1.00005	1.00005	1.000065	1.000102	1.000061	0.999981	
11	1.000074	1.000054	1.00005	1.00005	1.000065	1.000102	1.000061	
12	1.000097	1.000074	1.000054	1.00005	1.00005	1.000065	1.000102	
13	1.000146	1.000097	1.000074	1.000054	1.00005	1.00005	1.000065	
14	1.000135	1.000146	1.000097	1.000074	1.000054	1.00005	1.00005	

	lag_8	lag_9	lag_10	roll_mean_3	roll_std_3	roll_mean_5	\
10	0.999991	1.000003	0.999998	1.000059	0.000013	1.000059	
11	0.999981	0.999991	1.000003	1.000075	0.000022	1.000065	
12	1.000061	0.999981	0.999991	1.000106	0.000037	1.000084	
13	1.000102	1.000061	0.999981	1.000126	0.000026	1.000101	
14	1.000065	1.000102	1.000061	1.000113	0.000048	1.000102	

	roll_std_5	y
10	0.000011	1.000074
11	0.000021	1.000097
12	0.000039	1.000146
13	0.000039	1.000135
14	0.000038	1.000058

Figure 1: Variables utilisés

- un **ensemble d'apprentissage**, utilisé pour entraîner les modèles ;
- les **5 dernières lignes**, utilisées comme données de test, que le modèle devra prédire.

Cette structuration reflète exactement notre objectif : prédire les **5 prochaines valeurs du prix** en utilisant uniquement les observations antérieures.

3 Modèle retenu et analyse des variables

Plusieurs modèles de régression ont été testés pour la prédiction des 5 prochaines valeurs du prix de l'USDC. Après comparaison des performances sur les données de test, c'est le **modèle Random Forest** qui a obtenu les meilleurs résultats, en particulier la plus faible erreur absolue moyenne (MAE).

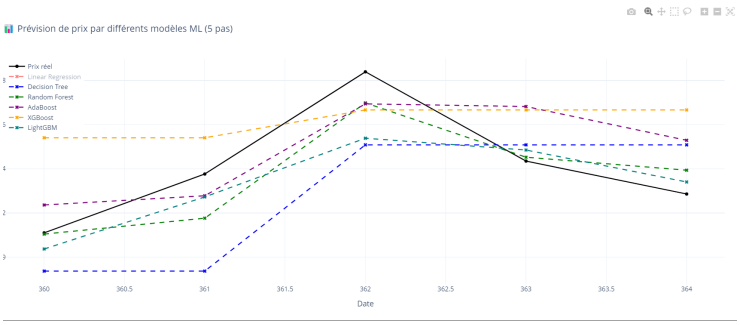

Figure 2: Variables utilisés

La MAE (Mean Absolute Error) mesure l'erreur moyenne entre les prédictions

et les observations réelles. Elle est définie par la formule suivante :

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

où $\hat{y}_i$ est la valeur prédite, y_i la valeur réelle, et n le nombre d'observations.

```
Random Forest          : MAE = 0.00000943
LightGBM               : MAE = 0.00001164
AdaBoost               : MAE = 0.00001720
Decision Tree          : MAE = 0.00002479
XGBoost                : MAE = 0.00002759
```

Figure 3: MAE

Importance des variables

Une fois le modèle entraîné, nous avons examiné l'importance des variables utilisées dans la prédiction. L'analyse révèle que la variable `roll_mean_3`, c'est-à-dire la moyenne mobile du prix sur 3 pas, est de loin la plus déterminante. Elle est suivie par :

- le décalage `lag_2` (prix d'il y a 2 pas),
- l'écart-type mobile `roll_std_3`.

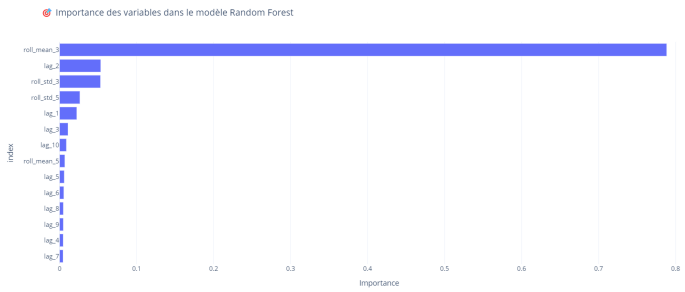

Figure 4: Imporatance des variables pour Random Forest

Cela montre que le modèle s'appuie fortement sur les **tendances récentes et la volatilité locale** pour effectuer ses prédictions. À l'inverse, certaines variables comme `lag_9`, `lag_4` ou `lag_7` ont une influence négligeable.

Ce comportement conforte les observations issues de l'analyse préliminaire par modèles ARMA : le processus semble bien capturé par un modèle autorégressif sur quelques pas récents, ce qui valide à la fois notre choix d'actif (USDC) et notre stratégie de modélisation.

4 Conclusion

Cette étude a montré qu'il est possible de prédire de manière raisonnablement précise les prochaines valeurs du prix de l'USDC à partir de ses observations passées, en utilisant un modèle de régression de type `Random Forest`. Le modèle tire principalement parti des tendances récentes (moyenne mobile) et de la volatilité locale pour effectuer ses prédictions, ce qui est cohérent avec la structure stationnaire de la série temporelle.

Pour aller plus loin, il serait pertinent d'effectuer une **optimisation fine des hyperparamètres** du modèle retenu. Cela pourrait se faire à l'aide d'une recherche sur grille (`GridSearchCV`) ou aléatoire (`RandomizedSearchCV`), en explorant notamment la profondeur maximale des arbres, le nombre d'estimateurs, ou encore les critères de division. Une telle phase d'optimisation permettrait d'améliorer la performance du modèle tout en maîtrisant le risque de surapprentissage.