RAPPORT FINAL DE PSC PRÉDICTION DES TRENDS SUR LE MARCHÉ FOREX APRÈS DES MOUVEMENTS IMPORTANTS

11 mai 2025

Abdelbar Ghassoub Bidias Cabrel Tiotsop Ngueguim
Eliav Kabla Le Duc Wesnil Dzangueu-Djouda
Coordinateur : Julien Combe
Tuteurs : Timon Fugier Fabrice Daniel

lusis

Table des matières

Remerciements

Nous tenons d'abord à remercier chaleureusement **Julien COMBE**, notre coordinateur, pour la proposition de ce sujet, sa disponibilité et son aide dans le cadrage des attentes.

Nous remercions nos partenaires de **LUSIS**, sans qui ce projet n'aurait pu être mené à bout. Merci à **Fabrice DANIEL** pour la formation en finance de marché qu'il nous a dispensée, et pour l'indication de ressources utiles en intelligence artificielle.

Nous remercions grandement **Timon Fugier** et **Raphael minato** pour leurs suivis hebdomadaires, leurs aides dans le guidage de nos recherches, et les nombreuses solutions qu'ils nous ont apportées et indiquées. leur aide a été capitale.

Enfin, nous remercions l'**École polytechnique**, et en particulier le **département d'économie**, dont le soutien logistique a permis l'établissement de ce travail.

Chapitre 1

INTRODUCTION

1.1 Présentation et motivations

Des milliards de dollars échangés chaque jour, des variations de prix à la milliseconde, une volatilité dictée par des facteurs macroéconomiques, géopolitiques et techniques. Le marché des changes (Forex) est le marché ayant le plus de volume échangé parmi tous les marchés financiers. Comme sur tout autre marché financier, prévoir les variations de prix (dans ce cas de taux de change) permet de faire la différence. Pourtant, la théorie financière classique postule que les marchés sont efficients : toute l'information disponible serait déjà intégrée dans les cours, rendant la prédiction impossible. Si tel était le cas, comment expliquer la suprématie des fonds quantitatifs et des traders à haute fréquence, qui génèrent des profits colossaux ? Leur profit ne serait-il que temporaire ? (cf. [7])

Loin d'être totalement aléatoires, les fluctuations du Forex obéissent à des dynamiques complexes que l'intelligence artificielle et le Machine Learning permettent désormais de décrypter, du moins partiellement. L'essor des algorithmes d'apprentissage automatique ouvre de nouvelles perspectives : analyser en temps réel les événements impactant le marché, détecter des signaux clés et prédire les variations futures. Ce projet s'inscrit dans cette quête d'anticipation, avec pour ambition de développer des modèles capables d'optimiser les stratégies de trading algorithmique.

1.2 État de l'art

Pendant de nombreuses années, les analystes financiers se sont principalement appuyés sur des modèles statistiques classiques, tels que l'ARIMA (AutoRegressive Integrated Moving Average), pour

analyser les séries temporelles. Ces outils, bien qu'efficaces pour identifier des tendances simples, montrent leurs limites face à des relations complexes à des volumes importants de données multidimensionnelles.

Avec l'émergence du *Machine Learning*, de nouvelles approches ont vu le jour pour repousser ces limites. Par exemple, les réseaux de neurones récurrents, et plus précisément les modèles LSTM (*Long Short-Term Memory*), occupent aujourd'hui une place importante. Introduits dans les années 1990, ces modèles sont particulièrement performants pour détecter des dépendances à long terme dans des séries temporelles. Leur efficacité a déjà été démontrée dans des domaines exigeants, comme la prédiction des mouvements de marchés financiers ou l'identification d'anomalies.

Plus récemment, des approches hybrides combinant des réseaux neuronaux avec des mécanismes d'attention ont permis d'améliorer encore davantage les performances des modèles prédictifs. Grâce à ces mécanismes, les modèles peuvent se concentrer sur les données les plus pertinentes d'une séquence, ce qui ouvre de nouvelles perspectives dans l'analyse des marchés. Par ailleurs, des architectures comme les transformers, par exemple BERT ou GPT, initialement développées pour le traitement du langage, commencent également à faire leurs preuves dans le domaine des séries temporelles.

En parallèle, d'autres méthodes, basées sur des techniques d'ensemble comme le *bagging* ou le *boosting*, ont gagné en popularité. Ces approches reposent sur l'idée de combiner plusieurs modèles pour améliorer leur robustesse globale. Parmi les algorithmes populaires, on peut citer les *Gradient Boosting Machines* (GBM), AdaBoost ou encore CatBoost, qui ont montré une efficacité remarquable pour traiter des problématiques liées aux données financières. Ces modèles, qui s'appuient sur des arbres de décision, sont capables de capturer des relations complexes entre différentes variables.

Enfin, deux méthodes sortent du lot grâce à leur grande efficacité et leur flexibilité : Random Forest et XGBoost. Ces modèles sont aujourd'hui largement reconnus par les experts pour leurs résultats dans l'analyse des données financières.

1.3 Implémentation de quelques stratégies/modèles classiques

Pour comprendre la pertinence de notre travail, nous avons comparé notre approche à des modèles qui ne font pas appel aux techniques d'apprentissage automatique afin de mieux cerner la valeur ajoutée de notre projet.

1.3.1 Stratégie purement tendancielle (MA)

On aimerait prévoir quand le taux de change augmente pour pouvoir acheter avant et vendre à ce moment. Et inversement, pour vendre puis acheter. Ce modèle constitue la base la plus rudimentaire de l'analyse algorithmique des marchés financiers. Il repose sur un principe fondamental de lissage des prix via le calcul de moyennes, dans le but de faire émerger une tendance générale à partir de données bruitées. L'idée est de filtrer les fluctuations à court terme (souvent aléatoires) pour révéler un mouvement directionnel plus stable, et la stratégie consiste ensuite à prédire un retour des prix vers leur tendance (supposée stable) ou "mean-reversion".

Calcul de la moyenne mobile simple (SMA)

La version la plus élémentaire de ce modèle repose sur la moyenne mobile simple (SMA – Simple Moving Average). Une moyenne mobile simple est définie comme la moyenne arithmétique des prix de clôture C_t sur une fenêtre glissante de taille N souvent exprimée en jours, heures, ou minutes selon

l'unité de temps du graphique. La formule mathématique est la suivante :

$$SMA_t = \frac{1}{N} \sum_{i=0}^{N} C_{t-i}$$

où :
- SMA_t est la moyenne mobile au temps t,
- C_{t-i} est le prix de clôture à l'instant $t - i$,
- N est la taille de la fenêtre temporelle.

Cette méthode a l'avantage de la simplicité, mais elle souffre d'un retard important par rapport aux variations réelles du marché. En effet, chaque nouvelle valeur remplace la plus ancienne, sans donner plus d'importance à l'information récente. Cela crée une latence, qui rend la détection des retournements de tendance lente et peu efficace dans un environnement aussi dynamique que le marché financier.

Amélioration avec les moyennes mobiles exponentielles (EMA)

Pour remédier à cette inertie, les moyennes mobiles exponentielles (EMA – Exponential Moving Averages) sont souvent préférées. Elles attribuent un poids exponentiellement décroissant aux valeurs passées, de sorte que les données récentes influencent davantage la valeur de la moyenne.

La formule de l'EMA est la suivante :

$$EMA_t = \alpha C_t + (1 - \alpha) EMA_{t-1}$$

Avec :
- $\alpha = \frac{2}{N+1}$ le coefficient de lissage,
- C_t le cours de clôture actuel,
- EMA_{t-1} la valeur de l'EMA à l'instant t.

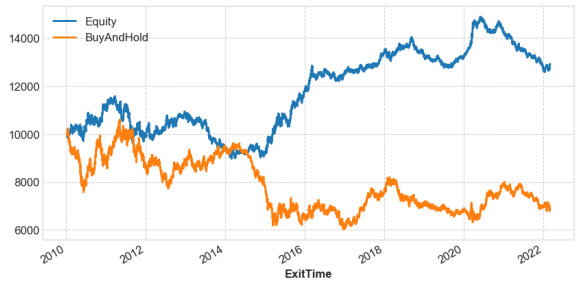
Grâce à cette pondération, l'EMA est plus réactive que la SMA aux dernières variations du prix. Cependant, même cette méthode présente une inertie résiduelle, surtout lorsque la période N est longue pour mieux lisser les mouvements de fond.

Limites des moyennes mobiles

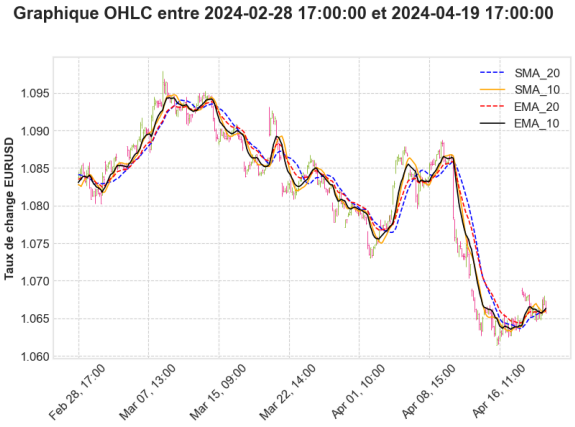
Le principal inconvénient de ces modèles réside dans leur retard de détection des mouvements brusques. Par exemple, une variation soudaine du prix de l'ordre de 10 pips en une heure pourrait ne se refléter pleinement dans l'EMA qu'après plusieurs heures (par exemple, 2 à 3 heures selon la période choisie), ce qui est problématique pour notre stratégie.

Dans un contexte réel, cette inertie peut faire manquer des opportunités de marché. En effet, si l'algorithme détecte un changement de tendance plusieurs heures après sa survenue, il est probable que le marché ait déjà absorbé cette information et que les conditions de profitabilité aient disparu. Cela conduit à des décisions d'achat ou de vente mal synchronisées, qui se traduisent inévitablement par des performances médiocres.

En implémentant le modèle, on obtient les résultats suivants :



(a) Courbe de l'Equity pour la paire EURUSD avec une stratégie basée uniquement sur les moyennes mobiles



(b) Superposition des barres OHLC et du suivi des moyennes mobiles

Métrique	Valeur
Sharpe	0.27
Max Drawdown	23%
Retours annuels	2,05%
Moyenne de gain pour les trades gagnants	12,21 pips
Moyenne de perte pour les trades perdants	11,83 pips

Table 1.1 – Quelques résultats métriques du modèle (L'explication des métriques sera détaillé dans le chapitre 8)

Pourquoi ce modèle donne de mauvais résultats

En résumé, malgré sa simplicité et son efficacité apparente pour capturer les grandes tendances, ce type de modèle est mal adapté aux marchés volatils où les prix évoluent rapidement par rapport à la fréquence d'échantillonnage. Il manque de finesse dans la détection des points d'inflexion et ne permet pas une réactivité suffisante pour tirer parti des mouvements de court terme. Cela explique pourquoi, dans la pratique, les résultats obtenus avec ce type de stratégie sont souvent décevants, en particulier dans des environnements de trading algorithmique à haute fréquence ou de *day trading*. (cf. [7])

1.3.2 Modèle ARIMA

Principe

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est un modèle statistique couramment utilisé pour la modélisation et la prévision des séries temporelles. Son objectif principal est de prédire la valeur future d'une variable unidimensionnelle en s'appuyant uniquement sur ses observations passées et sur des résidus (ou bruits) supposés aléatoires, et généralement gaussiens. Il combine trois composantes fondamentales :

— Auto-Régression (AR) : La variable est exprimée comme une combinaison linéaire de ses propres valeurs passées ;

— Intégration (I) : on prend des différences successives pour rendre la série stationnaire (c'est-à-dire éliminer les tendances ou variations saisonnières) ;
— Moyenne mobile (MA) : la variable est également influencée par une combinaison linéaire des bruits (erreurs de prévision) passés.

Formulation mathématique d'un $ARIMA(p, d, q)$

Un modèle ARIMA est noté $ARIMA(p, d, q)$ où : p est l'ordre de l'auto-régression (AR), d est le nombre de différences nécessaires pour rendre la série stationnaire (I), q est l'ordre de la moyenne mobile (MA).

Si y_t est la série temporelle d'origine, on applique d différenciations pour obtenir une série stationnaire y'_t. Le modèle s'écrit alors :

$$y'_t = \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Où :
— ϕ_i sont les coefficients de la partie autoregressive,
— θ_j sont les coefficients de la partie MA.
— ϵ_t est un bruit blanc, souvent supposé gaussien, de moyenne nulle et de variance constante.

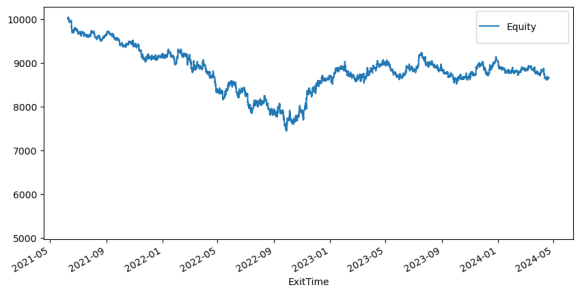
En résumé, l'estimation de la variable dépend linéairement d'un certain nombre de valeurs passées et de résidus aléatoires passés.

Limites théoriques dans le cadre des marchés financiers

Bien que le modèle ARIMA soit puissant pour des séries temporelles stationnaires et prévisibles (comme les volumes de production, températures, etc.), son application aux marchés financiers est contestée.

Selon l'hypothèse des marchés efficients (EMH) — en particulier dans sa forme faible — il n'existe aucune corrélation exploitable entre les prix passés et futurs. Cela signifie que :
— Les rendements boursiers se comportent comme une marche aléatoire (random walk),
— Les prix intègrent toute l'information passée, et il n'est pas possible de prédire les variations futures à partir des valeurs passées,
— Par conséquent, l'hypothèse fondamentale de l'ARIMA (existence d'une structure temporelle exploitable) ne tient plus dans ce cadre.

Cela se traduit en pratique par des performances très limitées de l'ARIMA pour la prédiction des prix ou des rendements d'actifs financiers, comme nous le constatons d'ailleurs au travers de nos implémentations :



(a) Courbe de l'Equity pour la paire EURUSD avec le modèle ARIMA

Métrique	Valeur
Sharpe	-0.48
Max Drawdown	25%
Rendement annuels	-4,73%
Moyenne de gain pour les trades gagnants	6,96 pips
Moyenne de perte pour les trades perdants	7,50 pips

Table 1.2 – Quelques résultats métriques du modèle (L'explication des métriques sera détaillé dans le chapitre 8)

1.4 Pourquoi utiliser une approche de machine learning ?

Bien que l'hypothèse faible des marchés efficients limite fortement la performance des modèles linéaires classiques tels qu'ARIMA sur le marché des changes, il est raisonnable de supposer que l'efficience forte n'est pas totalement vérifiée en pratique. En effet, des facteurs tels que des problèmes de liquidité ponctuelle, des asymétries d'information, ou encore l'existence de variables latentes (non observables directement mais influentes) peuvent engendrer des structures cachées dans les données de prix. Ces structures, difficilement détectables par des approches statistiques traditionnelles, pourraient être captées et exploitées par des modèles d'apprentissage automatique.

Le machine learning, en particulier dans ses formes supervisées et profondes, offre la possibilité d'apprendre des relations complexes et non linéaires à partir de données brutes, sans exiger de formulation explicite des dépendances. Il constitue ainsi une piste prometteuse pour dépasser les limites des modèles purement paramétriques dans un environnement aussi bruité et réactif que le marché du Forex.

1.5 Objectifs

L'idée principale est de concevoir des modèles capables d'identifier automatiquement les fluctuations importantes sur une base horaire et de prévoir les tendances futures. Les objectifs principaux sont les suivants :

— **Identifier les fluctuations majeures** : Mettre en place des modèles capables de détecter, à l'échelle horaire, les événements significatifs qui influencent le marché Forex.
— **Prédire les tendances à venir** : Développer des outils permettant de prévoir les évolutions futures du marché, qu'elles soient à la hausse ou à la baisse.
— **Analyser les performances** : Évaluer l'efficacité des modèles grâce à des indicateurs précis et des tests rétrospectifs (*backtesting*) pour en valider la pertinence.

Chapitre 2

Présentation des données

2.1 Description des données Forex

Notre projet repose sur des données horaires historiques de 9 paires de monnaies (EUR/USD , JPY/USD...), couvrant la période de 2010 à 2024, fournies par l'entreprise LUSIS. Ces données incluent les indicateurs fondamentaux OHLC (*Open, High, Low, Close*), et aussi publiquement accessibles via des sources comme Yahoo Finance. pour chaque période de trading, qui constitue la base de l'analyse des mouvements de prix. Contrairement aux données de nombreux marchés, nous avons choisi d'exclure les volumes d'échange, car dans le contexte décentralisé et hors-bourse (OTC) du Forex, les volumes échangés sur les plateformes centralisées ('exchanges') ne représentent pas une mesure pertinente de l'activité globale du marché.

Variable cible

La variable à prédire est définie comme le mouvement relatif du cours EUR/USD. Une colonne spécifique est ajoutée pour étiqueter chaque observation :

— Si le prix de clôture actuel est supérieur à celui d'une période prédéfinie (horizon de 4,8 ou 12 périodes), la valeur est étiquetée comme **1** (anticipation de hausse, stratégie *BUY*).

— Si le prix de clôture actuel est inférieur, la valeur est étiquetée comme **0** (anticipation de baisse, stratégie *SELL*).

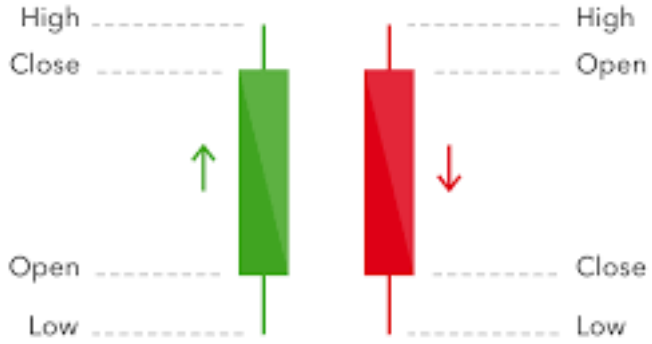


FIGURE 2.1 – OHLC représenté

Les données OHLC jouent un rôle clé dans les applications de machine learning en finance, car elles capturent la dynamique essentielle des marchés. Chaque composante de ces données offre des informations spécifiques : le prix d'ouverture reflète les attentes initiales, les niveaux haut et bas indiquent la volatilité et les extrêmes atteints, tandis que le prix de clôture représente le consensus final des investisseurs. Ces données sont souvent utilisées pour calculer des indicateurs techniques comme les moyennes mobiles ou les indices de volatilité, qui enrichissent les jeux de données pour les modèles prédictifs [9].

2.2 Analyse des Indicateurs Techniques Utilisés

Pour exploiter le dataset et l'adapter à nos algorithmes, on l'a enrichi par des indicateurs statistiques propres au marché des changes.[7]

Dans cette section, nous n'allons pas redéfinir les indicateurs déjà définis en section 1.3 mais nous allons plutôt définir les autres indicateurs que nous avons utilisés.

2.2.1 Volatilité

Average True Range (ATR) : L'ATR mesure les mouvements les plus importants des prix sur une période donnée, de façon linéaire" :

$$ATR_t = \frac{1}{n} \sum_{i=0}^{n-1} TR_{t-i} \tag{2.1}$$

où TR_t est la True Range donnée par :

$$TR_t = High_t - Low_t \tag{2.2}$$

ATR Normalisé : L'ATR est souvent normalisé en le divisant par l'écart-type du prix de clôture sur une fenêtre de temps :

$$Norm\_ATR_t = \frac{ATR_t}{\sigma(Close_t)} \tag{2.3}$$

où $\sigma(Close_t)$ est l'écart-type du prix de clôture sur les 50 dernières périodes.

2.2.2 Tendance

Average Directional Index (ADX) : L'ADX un ratio (qui s'exprime en pourcentage) et mesure la force d'une tendance en utilisant les indicateurs Directional Movement (+DI et -DI) :

$$ADX_t = \frac{100}{n} \sum_{i=0}^{n-1} \frac{|DI^+ - DI^-|}{DI^+ + DI^-}$$

où DI^+ et DI^- sont dérivés des variations de prix.

Différence Prix - EMA(20) : Elle mesure l'écart entre le prix et la moyenne mobile exponentielle à 20 périodes :

$$Price\_EMA20\_Diff_t = Close_{t-1} - EMA20_t \tag{2.4}$$

Kaufman Adaptive Moving Average (KAMA) : La KAMA est une moyenne mobile adaptative qui ajuste sa sensibilité selon la volatilité du marché. Elle est calculée selon les étapes suivantes :

$$ER_t = \frac{|Price_t - Price_{t-n}|}{\sum_{i=1}^{n} |Price_i - Price_{i-1}|} \tag{2.5}$$

$$SC_t = (ER_t \cdot (fastSC - slowSC) + slowSC)^2 \tag{2.6}$$

$$KAMA_t = KAMA_{t-1} + SC_t \cdot (Price_t - KAMA_{t-1}) \tag{2.7}$$

où généralement on prend $fastSC = \frac{2}{2+1}$ et $slowSC = \frac{2}{30+1}$.

La KAMA réagit rapidement aux mouvements significatifs lorsque le marché est directionnel, mais filtre le bruit lorsque le marché est stable. Elle est donc utile pour identifier des entrées dans des tendances durables tout en réduisant les faux signaux en période de range.

Parabolic Stop and Reverse (Parabolic SAR) : Le Parabolic SAR est un indicateur qui fournit des points potentiels de retournement de tendance.

$$SAR_t = SAR_{t-1} + AF_t \cdot (EP_t - SAR_{t-1}) \tag{2.8}$$

où EP_t est l'Extreme Point (plus haut ou plus bas selon la tendance) et AF_t est le facteur d'accélération, initialement fixé à 0.02 et incrémenté de 0.02 à chaque nouveau plus haut (ou plus bas), jusqu'à un maximum de 0,2.

Lorsqu'il est en dessous du prix, il indique une tendance haussière ; au-dessus, une tendance baissière. Le croisement du prix avec le SAR est souvent interprété comme un signal de sortie ou de retournement potentiel.

Kurtosis (Aplatissement) : La kurtosis mesure la forme de la distribution des rendements. Elle est définie comme :

$$Kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \tag{2.9}$$

Une kurtosis élevée (>3) indique une distribution avec des queues épaisses, ce qui montre la présence de mouvements de prix extrêmes. Cela signifie que les rendements sont susceptibles d'être plus volatils qu'une distribution normale ne le prévoirait.

Skewness (Asymétrie) : La *skewness* mesure l'asymétrie de la distribution des rendements :

$$Skewness = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} \tag{2.10}$$

Une *skewness* positive signifie que les rendements extrêmes positifs sont plus fréquents (asymétrie à droite), tandis qu'une *skewness* négative indique des risques accrus de pertes extrêmes (asymétrie à gauche). Cet indicateur est précieux pour évaluer les risques de queues de distribution dans les stratégies de trading.

Détection du Marché Haussier : Un marché est considéré comme haussier si :

$$Bull\_Market_t = \mathbb{1}(Close_t > EMA20_t \wedge Close_t > EMA50_t) \tag{2.11}$$

où $\mathbb{1}$ est la fonction indicatrice.

2.2.3 Momentum

Relative Strength Index (RSI) : Le RSI mesure la vitesse et le changement des mouvements de prix :

$$RSI_t = 100 - \frac{100}{1 + RS_t} \tag{2.12}$$

où RS_t est le ratio des gains moyens sur les pertes moyennes,calculé sur une période donnée (généralement 14 périodes).

RSI décalé : On introduit des versions décalées du RSI pour capturer les dynamiques temporelles :

$$RSI\_lag_k = RSI_{t-k} \tag{2.13}$$

Chande Momentum Oscillator (CMO) : Le CMO est une variation du RSI calculée comme :

$$CMO_t = 100 \times \frac{S^+ - S^-}{S^+ + S^-} \tag{2.14}$$

où S^+ et S^- sont les sommes des gains et des pertes.

Stochastic Oscillator (Stochastic K) : L'oscillateur stochastique est donné par :

$$Stochastic\_K_t = \frac{Close_t - Low_n}{High_n - Low_n} \times 100 \tag{2.15}$$

où Low_n et $High_n$ sont les plus bas et les plus hauts sur n périodes.

Rate of Change (ROC) : Le ROC mesure le pourcentage de la variation du prix sur n périodes, exprimée en pourcentage :

$$ROC_t = \frac{Close_t - Close_{t-n}}{Close_{t-n}} \times 100 \tag{2.16}$$

Williams %R : L'indicateur Williams %R est un indicateur développé par Larry Williams. Il sert à identifier si un actif est en situation de sur-achat ou de survente. Mathématiquement, il se calcule via la formule suivante :

$$\%R = \frac{\text{Plus grand High} - \text{Close}}{\text{Plus grand High} - \text{Plus petit Low}} \times (-100)$$

où :

- Plus grand high représente le plus grand high sur les n dernières périodes
- Plus petit low représente le plus petit low sur les n dernières périodes
- Close représente le prix de clôture actuel

Typiquement, lorsque l'indicateur de Williams est supérieur à -20, alors il y a surachat potentiel et une valeur inférieure à -80 indique une survente potentielle. La zone neutre se situe donc entre -80 et -20.

L'Ultimate Oscillator : C'est encore un oscillateur développé par Larry Williams qui combine 3 périodes différentes (court, moyen et long terme) et dont le but est d'identifier les divergences haussières ou baissières de manière plus fiable. La formule mathématique est la suivante :

$$U_{\text{Osc}} = 100 \times \frac{4 \times \text{avg}_7 + 2 \times \text{avg}_{14} + \text{avg}_{14}}{7}$$

Avec :
— $\text{avg}_k = \frac{\sum_{i=1}^{k} \text{bp}_i}{\sum_{i=1}^{k} \text{tr}_i}$,
— $\text{bp}_k = \text{Close} - \min(\text{Low}, \text{Close}_{\text{précédent}})$ à l'instant k,
— $\text{tr}_k = \max(\text{High}, \text{Close}_{\text{précédent}}) - \min(\text{Low}, \text{Close}_{\text{précédent}})$ à l'instant k.

Il permet aussi de fournir un signal de surachat ou de survente. S'il est plus grand que 70, alors on a un possible surachat, et lorsqu'il est inférieur à 30, on a une possible survente.

2.2.4 Indicateurs de Volatilité et Tendances

Moving Average Convergence Divergence (MACD) : Le MACD est la différence entre deux EMA :

$$MACD_t = EMA_{12}(Close_t) - EMA_{26}(Close_t) \tag{2.17}$$

Largeur des Bandes de Bollinger : Les bandes de Bollinger sont définies par :

$$Upper_t = SMA_{20}(Close_t) + 2\sigma(Close_t) \tag{2.18}$$

$$Lower_t = SMA_{20}(Close_t) - 2\sigma(Close_t) \tag{2.19}$$

La largeur des bandes est alors donnée par :

$$Bollinger\_Width_t = \frac{Upper_t - Lower_t}{Middle_t} \tag{2.20}$$

où $Middle_t = SMA_{20}(Close_t)$ est la moyenne mobile simple à 20 périodes.

2.3 Statistiques Descriptives

Les statistiques descriptives des variables de notre jeu de données pour la paire euro/dollar (eurusd) sont présentées dans le tableau suivant :

Variable	Mean	Std	Min	25%	50%	75%	Max
open	1.167	0.0921	0.98	1.10269	1.13733	1.20991	1.39624
high	1.167	0.0921	0.98	1.1033	1.13811	1.21029	1.39645
low	1.167	0.0921	0.98	1.10209	1.13612	1.20908	1.39482
close	1.167	0.0921	0.98	1.10283	1.13718	1.20935	1.39509
Label_UpBar	0.525	0.5	0	0	1	1	1
Label_CloseUp_8	0.501	0.5	0	0	1	1	1
ATR	0.0024	0.0009	0.000959631	0.00177835	0.002187	0.00274887	0.0058355
Norm_ATR	0.413	0.12	0.135699	0.32592	0.399414	0.483564	0.889752
Chaikin_Volatility	0.002	0.0008	0.000867228	0.00150632	0.0018	0.00235359	0.00506611

ADX	26.94	5.84	20.0025	22.3874	25.4431	29.981	49.4649

Chapitre 3

Sélection de features

Le modèle initial que nous avions développé était particulièrement sensible aux caractéristiques utilisées : l'ajout ou la suppression d'une feature pouvait entraîner des variations significatives dans les performances du modèle. Ainsi, pour construire un modèle prédictif performant, la sélection des caractéristiques revêt une importance cruciale. Elle permet non seulement de réduire la complexité computationnelle en limitant la taille des données, mais aussi d'améliorer la précision et l'interprétabilité du modèle en éliminant les variables redondantes ou peu pertinentes.

Dans ce chapitre, nous nous intéresserons aux différentes approches dont on s'est servi pour identifier les caractéristiques les plus pertinentes. Nous analyserons tout d'abord la matrice de corrélation afin de détecter les relations linéaires entre les variables et d'éviter les effets de redondance. Ensuite, nous nous intéresserons au Random Forest, qui nous permettra d'évaluer l'importance des caractéristiques en mesurant leur impact sur la performance du modèle (cf. [6]).

3.1 Analyse de Corrélation et d'Importance des Features

La matrice de corrélation obtenue met en évidence plusieurs relations importantes entre les variables. Nous observons des corrélations élevées entre certaines caractéristiques, notamment entre les indicateurs techniques liés à la dynamique des prix et ceux mesurant l'élan du marché. Par exemple, le RSI et ses variations décalées (RSI, RSI_lag1, RSI_lag2) affichent une corrélation très forte (supérieure à 0.9), indiquant qu'ils véhiculent une information similaire, et donc il est pertinent d'en conserver un seul afin d'éviter la redondance.

Un autre point à noter est la relation entre l'ATR et la largeur des bandes de Bollinger, qui traduit la volatilité du marché. Ceci confirme l'importance de ces indicateurs dans la détection des mouvements de prix. Cependant, la présence d'une corrélation modérée entre la volatilité et d'autres indicateurs comme le MACD ou l'Ultimate Oscillator montre que ces derniers apportent une information supplémentaire, justifiant leur conservation dans l'analyse.

D'un autre côté, certaines caractéristiques comme le Parabolic SAR et le KAMA présentent une corrélation presque parfaite (0.998), suggérant une forte similarité dans leur méthodologie de calcul. Il est donc recommandé de choisir de garder qu'un seul.

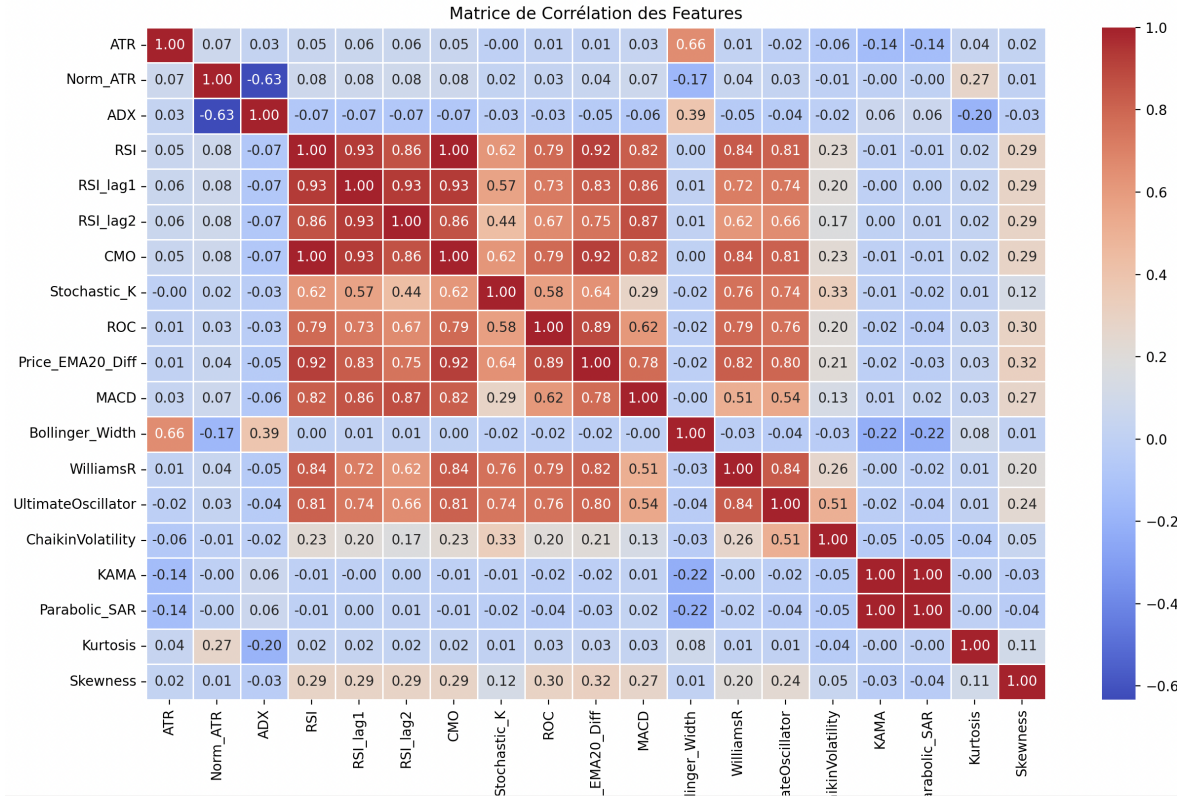


Figure 3.1 – Matrice de corrélation des features

3.2 Importance des features par catégorie

Pour classer les variables, nous avons utilisé le Random Forest, une méthode bien adaptée à cette tâche en raison de sa capacité à évaluer l'importance des caractéristiques de manière robuste et interprétable. Afin de structurer cette analyse, nous avons regroupé les caractéristiques en plusieurs catégories : les indicateurs de volatilité (tels que l'ATR et la volatilité de Chaikin), les oscillateurs de momentum (comme le RSI, le ROC et l'Ultimate Oscillator), les indicateurs de tendance (tels que le MACD, le Parabolic SAR et le KAMA), ainsi que des indicateurs combinés de volatilité et de tendance (comme le Skewness, l'ADX et le Kurtosis). (Pour les définitions des indicateurs, bien vouloir se référer à la section 2.2)

Nous avons pu par la suite mettre en lumière les indicateurs jouant un rôle déterminant dans l'explication des mouvements de prix. Par exemple, des indicateurs comme l'ATR et la volatilité de Chaikin se sont révélés particulièrement influents par rapport aux autres indicateurs de volatilité. Toutefois, nous avons également observé que certaines caractéristiques fortement corrélées, bien que pertinentes individuellement, risquaient d'introduire une redondance inutile.

3.3 Discussion des Résultats

L'analyse de la matrice de corrélation et des tableaux d'importance des caractéristiques a permis d'établir un cadre rigoureux pour la sélection des variables les plus pertinentes. Nous avons constaté

que certaines caractéristiques, comme l'ATR, la volatilité de Chaikin et la largeur des bandes de Bollinger, sont parmi les plus influentes. Leur forte importance indique qu'elles jouent un rôle clé dans la dynamique du marché.

De plus, les indicateurs comme le MACD, le ROC et l'Ultimate Oscillator, bien que présentant des corrélations modérées avec d'autres variables, apportent une information complémentaire utile à la prédiction. En revanche, certaines caractéristiques fortement corrélées, telles que le RSI et ses variations décalées, ainsi que le Parabolic SAR et le KAMA, ont été éliminées afin de réduire la redondance.

Indicateurs retenus : ATR , Norm_ATR , volatilité de Chaikin , Bollinger , MACD , ROC , Ultimate Oscillator , Price_EMA20_Diff , Stochastic_K.

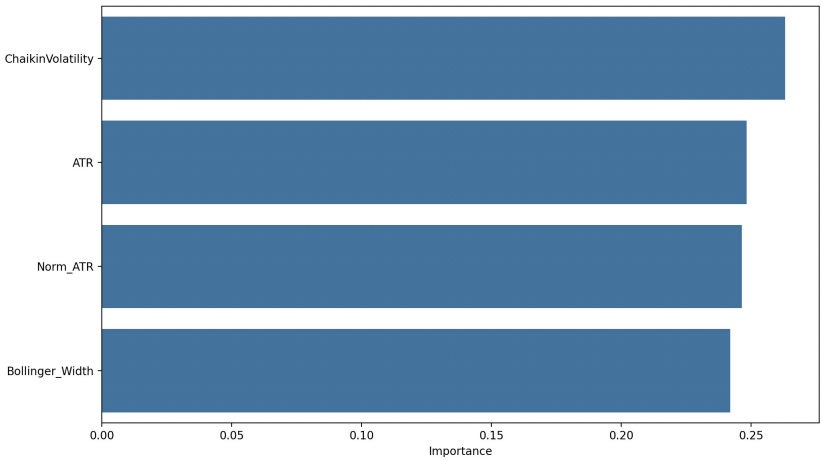


Figure 3.2 – Importance des indicateurs de volatilté

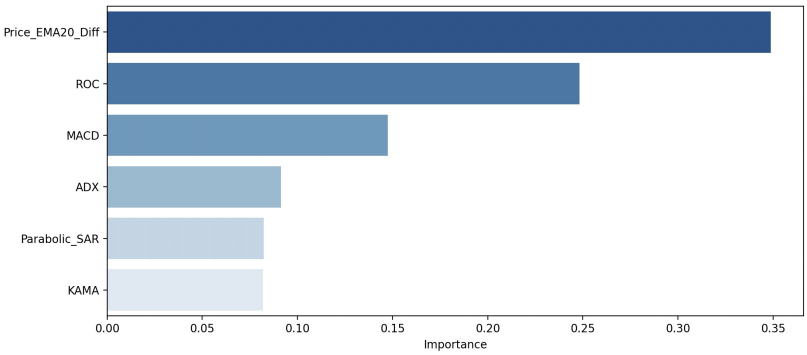


Figure 3.3 – Importance des indicateurs de tendance

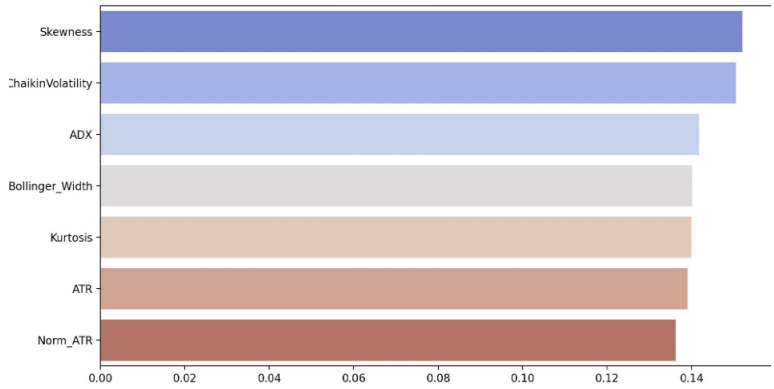


Figure 3.4 – Importance des indicateurs de volatilité et tendance

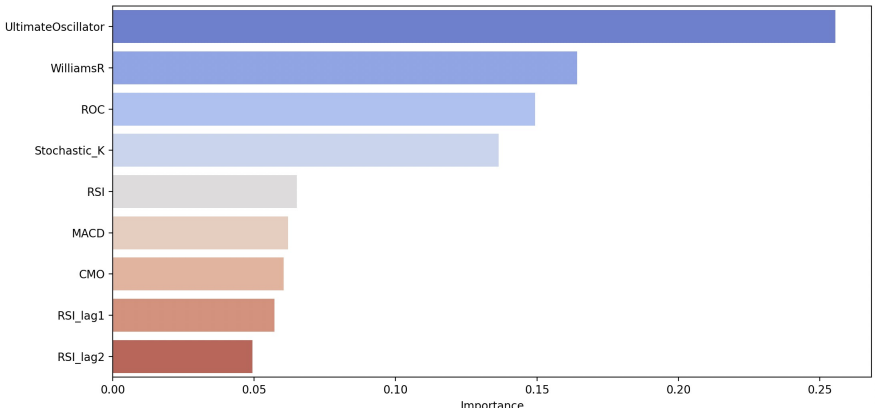


Figure 3.5 – Importance des oscillateurs stationnaires

Chapitre 4

Modèle de Classification - Random Forest

4.1 Introduction au Random Forest

Le Random Forest est un algorithme d'apprentissage supervisé qui repose sur un ensemble d'arbres de décision. Conçu pour améliorer la précision et la robustesse des modèles prédictifs, il fonctionne en combinant plusieurs arbres, chacun étant construit à partir d'échantillons aléatoires des données et de sous-ensembles de caractéristiques. [2]

Principaux atouts et caractéristiques [?] :

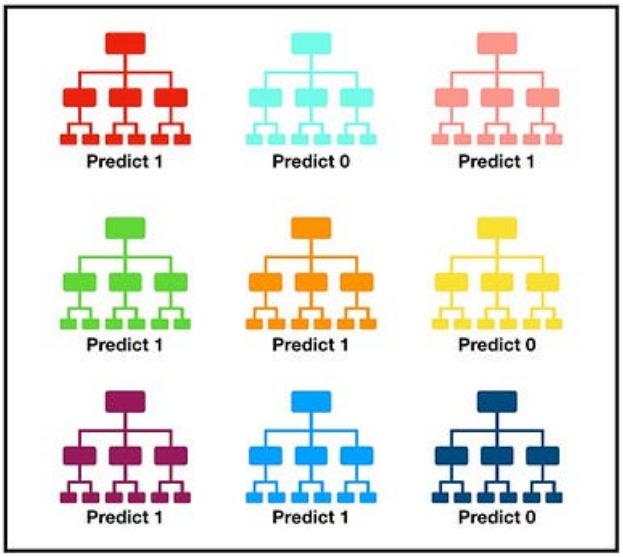
- **Résistance au surapprentissage :** En agrégeant les prédictions de plusieurs arbres, le Random Forest limite le risque de surapprentissage souvent rencontré avec un arbre de décision unique.
- **Robustesse face aux données bruitées :** Grâce à la technique du bootstrap, il devient moins sensible aux valeurs aberrantes et au bruit.
- **Capacité à gérer des données complexes :** Il est capable de modéliser efficacement les relations non linéaires et les interactions complexes entre les variables.
- **Facilité d'interprétation :** Des indicateurs comme l'importance des variables permettent d'identifier les caractéristiques les plus influentes dans le processus de décision.
- **Adaptabilité aux données déséquilibrées :** En ajustant les poids des classes ou en utilisant des techniques d'échantillonnage, il s'adapte aux jeux de données où certaines classes sont sous-représentées.

4.2 Théorie du Random Forest

Soit un ensemble d'entraînement $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$, où $X_i \in \mathbb{R}^d$ représente un vecteur de caractéristiques et Y_i la variable cible.

Construction des arbres :

Pour chaque arbre T^b, un sous-ensemble de données $\mathcal{D}^b$ est généré par bootstrap, c'est-à-dire en tirant aléatoirement, avec remise, des observations de $\mathcal{D}$. À chaque nœud de l'arbre, au lieu de considérer toutes les variables, un sous-ensemble aléatoire $\mathcal{S}^b \subset \{1, \ldots, d\}$ de m variables est sélectionné ($m \ll d$). Le critère de division est choisi de manière à minimiser une fonction d'impureté I, telle que :



- En classification (ex. Gini) :

$$I(S) = \sum_{k=1}^{K} p_k(1 - p_k)$$

où p_k est la proportion d'observations de classe k dans l'ensemble S.

- En régression (ex. variance) :

$$I(S) = \frac{1}{|S|} \sum_{i \in S} (Y_i - \bar{Y}_S)^2$$

où $\bar{Y}_S$ est la moyenne des Y_i dans S.

Prédiction finale :

Une fois B arbres construits, la prédiction du modèle pour une nouvelle instance X est donnée par :

- En classification (vote majoritaire) :

$$\hat{Y} = \arg\max_k \sum_{b=1}^{B} \mathbf{1}(T^b(X) = k)$$

- En régression (moyenne des prédictions) :

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} T^b(X).$$

Le *Random Forest* réduit la variance du modèle en agrégeant plusieurs arbres peu corrélés, offrant ainsi une meilleure généralisation et une plus grande robustesse au bruit.

4.3 Optimisation des hyperparamètres

Pour s'adapter aux spécificités de notre marché, on lance un grid-search sur un tableau de paramètres pour trouver la combinaison la plus adéquate [8] :

- **Nombre d'arbres (n_estimators) :** Nombre total d'arbres de décision construits dans la forêt.
- **Profondeur maximale (max_depth) :** Profondeur maximale autorisée pour chaque arbre de décision.
- **Critère de séparation (gini, entropy) :** Fonction utilisée pour évaluer la qualité d'une division à chaque nœud de l'arbre.
  - **Gini :** La fonction d'impureté de Gini mesure la probabilité qu'un échantillon choisi au hasard soit mal classé si on l'étiquette selon la distribution des classes dans le nœud. Elle est définie par la formule :

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

    où p_i est la proportion d'exemples appartenant à la classe i, et C le nombre total de classes.
  - **Entropie :** L'entropie est une mesure de l'incertitude ou de l'hétérogénéité des classes dans un nœud. Plus l'entropie est élevée, plus les classes sont mélangées. Elle est définie par :

$$Entropie = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

    où p_i est la proportion d'exemples de la classe i. L'entropie est nulle lorsque toutes les observations appartiennent à une seule classe.
- **Fraction d'échantillons (max_samples) :** Proportion ou nombre absolu d'échantillons tirés (avec remplacement) du jeu de données d'origine pour entraîner chaque arbre.

Chapitre 5

Modèle de Classification - XGBoost

5.1 Introduction à XGBoost

XGBoost (Extreme Gradient Boosting) est un algorithme d'apprentissage supervisé reposant sur la technique du gradient boosting . Il se démarque par sa rapidité, son efficacité et sa capacité à produire des modèles précis et optimisés, aussi bien pour les tâches de classification que de régression. [4]

Caractéristiques principales [5] :

- **Excellente performance :** Grâce à l'optimisation des calculs et à une utilisation efficace des ressources, XGBoost est souvent considéré comme l'un des algorithmes les plus performants, en particulier dans les compétitions de machine learning.
- **Gestion intuitive des données manquantes :** L'algorithme traite automatiquement les valeurs manquantes en les intégrant dans une branche spécifique lors de la construction des arbres, ce qui simplifie la préparation des données.
- **Réduction efficace du surapprentissage :** L'intégration de mécanismes de régularisation permet de limiter le surapprentissage tout en conservant une précision élevée du modèle.
- **Grande flexibilité des fonctions de perte :** XGBoost prend en charge une large gamme de fonctions de perte, offrant ainsi une adaptabilité à divers types de problèmes et de besoins analytiques.
- **Prise en charge du multi-threading :** Grâce à une implémentation multi-thread, XGBoost garantit une exécution rapide, même sur de grands volumes de données.

5.2 Théorie de XGBoost

L'idée principale repose sur la construction séquentielle d'arbres de décision en minimisant à chaque itération une fonction de perte via l'optimisation du gradient.

5.2.1 Formulation mathématique

Soit un ensemble d'entraînement $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ avec $X_i \in \mathbb{R}^d$ et Y_i la variable cible. Le modèle XGBoost apprend une somme d'arbres de décision :

$$\hat{Y}_i = \sum_{t=1}^{T} f_t(X_i), \quad f_t \in \mathcal{F}$$

où $\mathcal{F}$ est l'espace des arbres de décision.
La fonction de perte à minimiser est définie comme :

$$\mathcal{L} = \sum_{i=1}^{n} l(Y_i, \hat{Y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

où :
- $l(Y_i, \hat{Y}_i)$ est une fonction de perte différentiable (ex. erreur quadratique pour la régression, log-loss pour la classification).
- $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_j w_j^2$ est un terme de régularisation qui contrôle la complexité de l'arbre, avec γ un paramètre de pénalisation du nombre de feuilles et λ un paramètre de régularisation L_2.

5.2.2 Optimisation par Boosting du Gradient

XGBoost utilise une approche de second ordre en approximant la fonction de perte par un développement de Taylor d'ordre 2 :

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \Omega(f_t)$$

où :
- $g_i = \frac{\partial l(Y_i, \hat{Y}_i)}{\partial \hat{Y}_i}$ est le gradient de la fonction de perte.
- $h_i = \frac{\partial^2 l(Y_i, \hat{Y}_i)}{\partial \hat{Y}_i^2}$ est le hessien (dérivée seconde).

L'arbre optimal est construit en maximisant le gain d'information :

$$G = \frac{1}{2} \left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in L \cup R} g_i)^2}{\sum_{i \in L \cup R} h_i + \lambda} \right] - \gamma$$

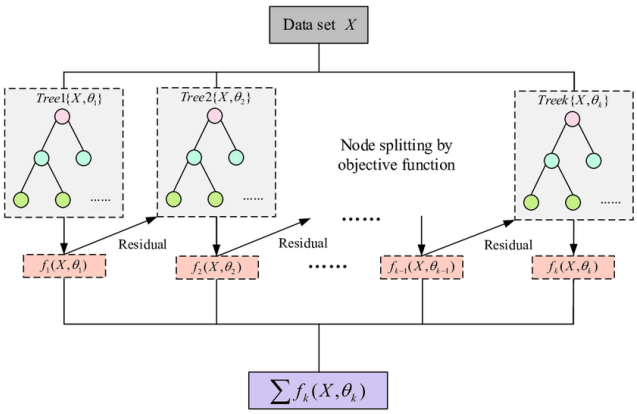
où L et R sont les deux partitions d'un nœud.
XGBoost implémente plusieurs optimisations :
- **Régularisation** (L_1, L_2) pour éviter le surapprentissage.
- **Pruning (élagage)** :C'est la suppression des branches inutiles ou peu performantes d'un arbre de décision afin de réduire le surapprentissage, plus précisément, il s'agit d'arrêter la croissance des branches si l'amélioration du score est insuffisante. basé sur la complexité des arbres pour éviter des arbres trop profonds.
- **Gestion efficace des valeurs manquantes** avec un chemin de décision optimal.
- **Traitement parallèle** et implémentation en mémoire optimisée.

5.3 Optimisation des hyperparamètres

Comme tout algorithme avancé, XGBoost peut être personnalisé pour améliorer ses performances [8]. Voici les hyperparamètres clés :

— **Nombre d'arbres (n_estimators) :** Définit le nombre total d'arbres à construire. Une valeur élevée peut améliorer la précision, mais au prix d'un temps de calcul accru.

— **Taux d'apprentissage (learning_rate) :** Détermine la contribution de chaque arbre au modèle final. Des valeurs faibles (par exemple, 0.01 ou 0.1) réduisent le risque de surapprentissage.



— **Profondeur maximale (max_depth) :** Contrôle la complexité des arbres individuels. Une profondeur trop grande peut entraîner un surapprentissage, tandis qu'une profondeur trop faible peut limiter la performance.
— **Sous-échantillonnage (subsample) :** Définit la proportion des données utilisées pour chaque arbre. Cela réduit la variance et améliore la robustesse du modèle.
— **Colsample_bytree :** Indique la fraction des caractéristiques utilisées pour construire chaque arbre. Cela favorise la diversité dans les arbres et aide à éviter le surapprentissage.

Chapitre 6

Sélection des forts mouvements

6.1 Méthodologie

6.1.1 Données et Prétraitement

Les données utilisées proviennent d'un fichier au format parquet contenant les prix OHLC (Open, High, Low, Close) des paires de devises. Trois étapes sont nécessaires pour préparer les données. Tout d'abord, on effectue un filtrage temporel pour ne conserver que les données postérieures à janvier 2015, afin de garantir la pertinence des modèles dans un contexte de marché récent. Ensuite, on calcule les indicateurs techniques discutés dans le chapitre précédent pour enrichir les analyses. Enfin, après normalisation, les indicateurs techniques sont comparés à des seuils. Ces seuils permettent d'identifier les conditions de marché extrêmes : par exemple, l'ATR est normalisé en fonction de l'écart-type des prix sur une fenêtre de 50 périodes, puis comparé à un percentile qu'on a choisi par grid search.

6.1.2 Sélection des Forts Mouvements

Pour identifier les forts mouvements du marché, nous avons mis au point un algorithme complexe qui reposait sur l'utilisation de plusieurs indicateurs techniques, afin de filtrer les mouvements significatifs tout en minimisant les faux signaux. Parmi les paramètres utilisés, on retrouve l'Indice Directionnel Moyen (ADX), l'Average True Range (ATR), le rate of change(ROC), la Différence Prix - EMA(20) et d'autres features complémentaires.

L'ADX a été utilisé pour mesurer la force de la tendance, permettant de distinguer les périodes où la tendance était suffisamment forte pour qu'un mouvement significatif puisse être isolé. L'ATR, quant à lui, a servi à évaluer la volatilité des prix, afin de repérer les périodes de fluctuations importantes. De plus, des oscillateurs comme le ROC ont été utilisés pour déterminer la vitesse du mouvement et ses revers potentiels. La Différence Prix - EMA(20) permettait de s'assurer que le prix avait fortement changé sur la fenêtre de 20.

Pour isoler les "forts mouvements", nous avons fixé des bornes et des quantiles pour chaque indicateur. Ces paramètres ont été affinés en fonction de diverses combinaisons d'indicateurs et de seuils afin de maximiser la capture des mouvements importants et réduire les faux positifs.

Cependant, après avoir optimisé toutes les autres parties du code et analysé les performances globales du modèle, nous avons constaté que cette méthode de détection des forts mouvements n'était pas optimale. En effet, en termes de performance globale, le modèle restait sous-performant, notamment en raison d'un maximum drawdown supérieur à 20% et d'un ratio de Sharpe inférieur à 0.32. Cela

indiquait que les modèles basés sur cette approche généraient des rendements risqués et trop volatils, malgré une détection précise de certains mouvements.

On a donc simplifié plus le modèle en ne gardant que l'ADX et l'ATR normalisé, nous avons réalisé un grid search avec le XG boost [10] (Puisqu'il performe mieux que le Random Forest) sur les deux thresholds pour balancer entre le critère directionnel et de volatilité de la donnée sélectionnée, les performances se sont magiquement améliorées. Comme critère de scoring, on s'est basé sur le ratio de P&L comme métrique, tout en ne gardant que les modèles de performances correctes (accuracy $> 53\%$; max drawdown < 5 ; sharpe > 1.2 , sinon le modèle est noté "nan"). Ceci nous a permis de capter les variations significatives sans être influencées par des valeurs extrêmes isolées. La condition de sélection des forts mouvements est définie par :

$$\text{Strong\_movement} = (\text{NormATR} > \text{VolatilityThreshold}) \wedge (\text{ADX} > \text{ADXThreshold}) \tag{6.1}$$

Volatility Threshold	ADX Threshold	Profit factor
20	20	1.43
20	25	1.22
25	20	1.18
25	15	*nan*
15	25	*nan*
30	20	1.47
20	30	*nan*
30	15	*nan*
15	30	*nan*

Nous avons choisi la meilleure combinaison de seuils à Volatility_Threshold = 30 et ADX_Threshold = 20, avec un P&L = 1.47. Cette configuration reflète une performance optimale en prenant compte à la fois de la volatilité et de la force de la tendance, tout en maintenant une sélection cohérente des mouvements significatifs.

Cependant,une augmentation excessive des seuils réduit mécaniquement le nombre d'occurrences (des dépassements de seuils), ce qui augmente le risque de sur-apprentissage. Ainsi, il est important de trouver un équilibre entre la précision des critères de sélection et la quantité de données conservées.

6.2 Résultats et Interprétation

Un ensemble de données initial de 58 659 entrées réduit à seulement 12 142 après filtrage (pour un horizon de 8).

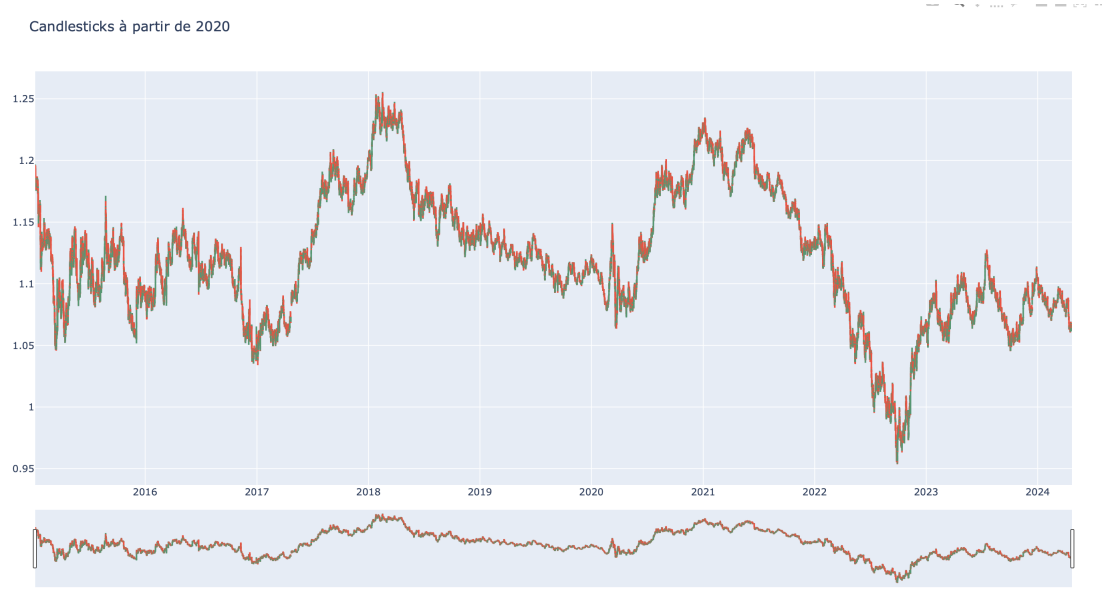


Figure 6.1 – Les données brutes

Figure 6.2 – Selection des forts mouvements entre Mars 2013 et Mars 2014

6.3 Analyse du Data Leakage

L'une des étapes fragiles dans la conception d'un modèle de trading basé sur l'apprentissage automatique est la gestion rigoureuse des données afin d'éviter le **data leakage**. Le data leakage, c'est simplement lorsqu'on utilise des données du futur ou du présent pour prédire des données du futur ou du présent. Par exemple, si on veut prédire le cours de la devise de l'heure actuelle, en utilisant les données de l'heure actuelle, c'est juste une condition, et utiliser ces données peut donner des résultats anormalement bons. Dans cette section, on résume les différentes étapes que nous avons traversées pour identifier et corriger les problèmes de fuite de données liés à la création des features et des labels,

en relation avec l'horizon de prédiction.

6.3.1 La logique derrière la construction des tranches

Commençons tout d'abord par comprendre quelques notions clés dans la préparation des données.

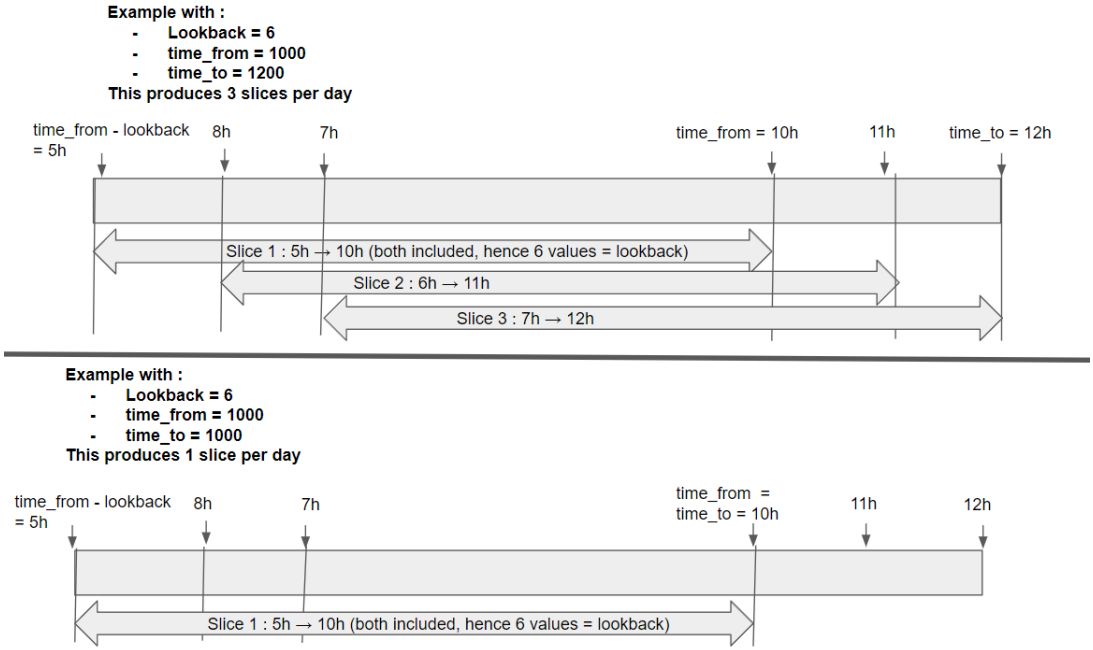


Figure 6.3 – Time_to, Time_from et Lookback

La figure ci-dessus explique la construction des tranches.Par exemple :

- Si vous avez 3 tranches par jour comme indiqué dans la figure, ces tranches se chevaucheront à certaines heures.

- Avoir 4 tranches par jour quadruplera l'ensemble des données après le découpage.

6.3.2 Problèmes Identifiés dans les premières versions du Code

Après l'analyse des données, on a détecté plusieurs anomalies dans la version du code soumis lors de la présentation du rapport intermédiaire, mettant en évidence des incohérences entre la construction des labels et la filtration des forts mouvements.

Création du Label Après Filtrage :

Définition initiale de la feature : Le label `Label_CloseUp_horizons` était défini comme suit :

$$\text{Label} = \text{Close}(t) - \text{Close}(t + h)$$

où h est le nombre de forts mouvements, ce qui n'est pas un intervalle de temps fixe.

L'horizon variable, dépendant des forts mouvements, rend le label difficile à interpréter sans échelle temporelle fixe. De plus, le risque de data leak, où le modèle accède à des informations futures, introduit un biais qui compromet la validité des prédictions.

Structure de l'Ensemble d'Entraînement :
Définition de la feature :

$$\text{Feature} = \frac{\text{Valeur normalisée après filtrage}}{\text{Valeur non filtrée}}$$

La normalisation a été effectuée après filtrage, ce qui a introduit un biais dans les données. De plus, le dataset manquait de diversité d'observations, car seules les observations correspondant aux mouvements identifiés à 16 h étaient conservées.

Pour illustrer ce problème, imaginons un exemple où l'horizon de prédiction est de 2 périodes et où le lookback est également de 2.

Date	Feature 1	Feature 2	Close	Fort Mouvement
14h	1	2	1.00	OUI
15h	2	4	1.10	NON
16h	3	6	1.00	OUI
17h	4	8	0.90	NON
18h	5	10	1.05	OUI

Table 6.1 – Exemple de structuration des données

Dans cette configuration, le modèle prendrait les features des derniers forts mouvements (16h et 14h) comme entrée et essaierait de prédire la différence entre le prix de clôture de 18h et celui de 16h. Cependant, le signal obtenu n'est pas exploitable, car la logique de filtrage fausse l'interprétation réelle du label.

6.3.3 Corrections Apportées dans la Version Actuelle

On a défini le label `Label_UpBar` avant d'appliquer le filtre des forts mouvements, assurant ainsi que l'information cible reflète une dynamique naturelle du marché et non une dépendance aux filtres exogènes.

Par exemple, on utilise les données disponibles à 16h00 pour prédire si le prix va monter ou descendre d'ici 08h00 le lendemain, puis on prend position à 16h00 et on la clôture 32 heures plus tard.

Les données ont été réajustées pour inclure un plus grand historique (à partir de 2013) et pour éliminer le filtre restrictif de la London Session (08h30 - 16h30), puisque le critère des forts mouvements capture déjà les périodes de volatilité accrue.

Nous avons supprimé la normalisation des séries temporelles, car elle était appliquée sur un dataset déjà filtré, ce qui introduisait un biais. Puis, on a revu le traitement des dépendances temporelles pour garantir que les features utilisées reflètent uniquement des informations accessibles au moment de la prédiction.

On a de même réglé le processus de backtest (voir le chapitre des Résultats) pour qu'il s'aligne sur la nouvelle logique d'étiquetage et de structuration des données. L'entrée se fait désormais sur le prix **Open** et la sortie sur le prix **Close** de la même période. Le paramètre `predict_bars` a été fixé à 0, supprimant ainsi l'incohérence entre l'horizon d'apprentissage et celui du backtest.

Chapitre 7

Entraînement du modèle

Algorithm 1 Préparation et entraînement du modèle [3]

Entrée : Données historiques df, caractéristiques $features$, label $Label\_UpBar$
Sortie : Modèle Random Forest entraîné, performances évaluées
$df1 \leftarrow df[df.index \geq$ "2015-01-01"]
Construire un graphique en chandeliers avec $df1$
Préparation des données :
$(X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}, y_{datetime}) \leftarrow$ `prepare_data`$(df,$ "2019-10-10", $features, Label\_UpBar)$
Remodeler $X_{train}, X_{val}, X_{test}$ en matrices 2D
Définition du score personnalisé :
custom_scoringy_{true}, y_{pred} $accuracy \leftarrow \frac{\sum(y_{pred}=y_{true})}{|y_{true}|}$
$spread \leftarrow$ moyenne$(y_{pred}[y_{pred} = y_{true}])$
$profits \leftarrow y_{true} - y_{pred}$
$sharpe \leftarrow \frac{\text{moyenne}(profits)}{\text{std}(profits)+10^{-6}}$
Calcul des gains/pertes moyens et de leur ratio
Calcul du score final en combinant les métriques
final_score
Optimisation des hyperparamètres :
Définir l'espace de recherche $param\_grid$ pour Random Forest $rf\_model \leftarrow$ RandomForestClassifier(random_state = 1000)
Appliquer `RandomizedSearchCV` sur rf_model avec $param\_grid$
$best\_rf\_model \leftarrow$ Meilleur modèle obtenu
Entraînement du modèle optimal :
Entraîner $best\_rf\_model$ sur (X_{train}, y_{train})
Évaluation du modèle :
$y_{pred} \leftarrow best\_rf\_model.predict(X_{val})$
Appliquer `classifier_backtest` sur $(df, y_{pred}, y_{datetime})$
Afficher les résultats et la précision du modèle

7.1 Définition de la Fonction de Score

Initialement, nous avons défini le scoring comme étant l'accuracy à maximiser. Cependant, cette approche n'a pas donné de bons résultats, car elle conduisait à des modèles maximisant l'accuracy(61%) avec des stratégies de fort P&L (plus que 1.7) tout en négligeant d'autres paramètres essentiels, tels que le maximum drawdown(>56%) et le ratio de Sharpe(<0.12). Cela a engendré des modèles très volatils et risqués.

Pour remédier à ce problème, nous avons défini une métrique personnalisée intégrant plusieurs indicateurs clés[10] :

— **Accuracy** : Proportion des prédictions correctes par rapport au nombre total de prédictions.
— **Ratio de Sharpe** : Mesure du rendement ajusté au risque, calculée comme le ratio entre la moyenne du rendement excédentaire et l'écart-type du rendement excédentaire.
— **Gain/Pertes Moyennes** : Ratio entre le gain moyen et la perte moyenne.

On a calculé le score final en combinant ces métriques à l'aide d'une pondération équilibrée. Afin de trouver les meilleurs paramètres permettant d'équilibrer la précision, le ratio de Sharpe et le facteur de profit, nous avons optimisé ces poids pour maximiser les résultats finaux. Les différentes configurations testées sont présentées dans le tableau suivant :

Poids de la Précision	Poids du Sharpe	Poids Gain/Pertes Moyennes	Profit factor
1.0	0.5	0.8	1.69
2.0	1.0	1.5	1.31
3.0	1.5	1.2	1.09
3.0	1.5	0.8	1.75
3.0	1.0	1.5	1.3
3.0	1.0	1.2	1.19
3.0	1.0	0.8	1.55
3.0	0.5	1.5	1.34
3.0	0.5	1.2	1.1
3.0	0.5	0.8	1.23

On a lancé un *grid search* sur les coefficients de pondération afin de maximiser le *Profit factor* trouvé à la fin du backtest, tout en conservant les modèles présentant de bonnes performances sur les autres métriques (précision > 0.53, ratio de Sharpe > 1.2, drawdown maximal $< 5\%$). Les dix meilleures combinaisons de paramètres sont présentées dans le tableau ci-dessus. En fin de compte, nous avons retenu la combinaison : $(3.0; 1.5; 0.8)$, qui a permis d'optimiser l'ensemble des critères.

Chapitre 8

Analyse des Résultats

8.1 Backtesting sur le Marché des Devises

Le *backtesting* constitue une étape essentielle dans l'évaluation de la pertinence des modèles prédictifs en finance quantitative, en particulier dans le cadre des stratégies de trading algorithmique. Il s'agit d'évaluer les performances du modèle sur des données réelles passées. C'est une manière concrète de confronter notre modèle à la réalité du marché.

Dans notre cas, le backtesting a pour objectif de tester les signaux générés par les modèles **Random Forest** et **XGBoost** sur différentes paires de devises majeures du marché Forex (comme EUR/USD, USD/JPY, GBP/USD, etc.), et à plusieurs horizons temporels (4, 8 et 12 périodes). Il s'inscrit dans une logique de stratégie de trading : l'objectif est de mesurer les gains potentiels ainsi que la stabilité et la fiabilité des signaux de trading produits. Le backtesting permet donc de filtrer empiriquement les stratégies exploitables.

Afin de quantifier rigoureusement les performances, nous utilisons plusieurs indicateurs standards de l'industrie :

Sharpe Ratio : mesure la performance ajustée au risque en comparant le rendement excédentaire de la stratégie au regard de sa volatilité. Plus ce ratio est élevé, plus la stratégie est attractive par rapport à un investissement sans risque.

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

où R_p est le rendement moyen du portefeuille, R_f est le taux sans risque, et σ_p est l'écart-type des rendements du portefeuille.

AVG P&L : représente la moyenne des profits et des pertes par trade dans un ensemble donné. Une valeur positive indique que les trades génèrent en moyenne un profit net, tandis qu'une valeur négative signale une perte nette moyenne.

Profit Factor : rapport entre les gains totaux et les pertes totales. Une valeur supérieure à 1 indique que la stratégie est globalement profitable.

$$\text{Profit Factor} = \frac{\sum \text{Gains}}{\sum \text{Pertes}}$$

Sortino Ratio : variante du Sharpe Ratio, mais ne prenant en compte que la volatilité négative. Il est particulièrement utile pour évaluer les risques asymétriques.

$$\text{Sortino Ratio} = \frac{R_p - R_f}{\sigma_d}$$

où σ_d est l'écart-type des rendements négatifs (volatilité à la baisse uniquement).

Ratio Gain/Perte : moyenne des gains rapportée à la moyenne des pertes, permettant d'évaluer la qualité moyenne des trades gagnants par rapport aux trades perdants.

$$\text{Gain/Loss Ratio} = \frac{\text{Gain moyen des trades gagnants}}{\text{Perte moyenne des trades perdants}}$$

Dans ce chapitre, on présentera les performances générales et celles du meilleur et du pire modèle. Pour plus de détails, on a mis les performances des 9 paires de marché en annexes.

Analyse Comparative

Profit Factor : Random Forest vs. XGBoost

Le modèle XGBoost surpasse systématiquement Random Forest sur le critère du Profit Factor. Pour toutes les paires étudiées, le Profit Factor moyen de XGBoost est supérieur. Par exemple, sur EUR/USD, on observe une valeur de 1.62 avec XGBoost contre seulement 0.54 avec Random Forest. Cette différence souligne la capacité accrue de XGBoost à générer des gains relatifs aux pertes. En revanche, Random Forest souffre d'une forte décroissance du Profit Factor à l'horizon 12, notamment sur les paires AUDUSD, EURJPY et EURUSD. Ce comportement pourrait être expliqué par le fait que le XGBoost, par construction, incorpore la structure temporelle des données car il possède de nouveaux prédicteurs séquentiellement, contrairement au Random Forest.

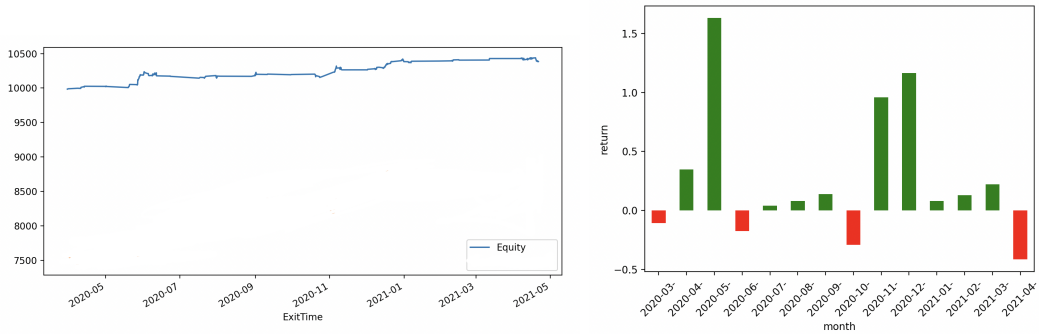
Ratios de Risque-Rendement

Les ratios de Sharpe et Sortino suivent des trajectoires contrastées selon le modèle. XGBoost produit des ratios systématiquement positifs et élevés, comme pour USD/CHF où le Sharpe atteint 2.53. À l'opposé, Random Forest présente des valeurs négatives, voire non définies. Sur USD/CHF, par exemple, le Sharpe est de -0.24. Cela traduit une gestion du risque systématiquement meilleure avec XGBoost. Une anomalie notable apparaît sur la paire GBPJPY avec Random Forest : le Sharpe Ratio devient NaN à l'horizon 12, ce qui suggère une volatilité nulle ou des rendements infinis, rendant ces indicateurs ininterprétables.

8.2 Cas Meilleur : XGBoost sur USD/JPY

À l'horizon 12, le modèle XGBoost appliqué à la paire USD/JPY atteint un **Average PL** de **1.51**, un **Sharpe Ratio** de **2.96** et un **Sortino Ratio** de **3.09**, avec un **ratio gain/perte moyen** de **1.68** et une **accuracy** de **54%**. Ces résultats sont particulièrement impressionnants, indiquant non seulement une forte rentabilité des signaux générés, mais aussi une excellente performance ajustée au risque. Plusieurs facteurs économiques peuvent expliquer cette réussite. Premièrement, le marché USD/JPY est historiquement caractérisé par des tendances directionnelles claires, souvent induites par les différentiels de taux d'intérêt entre les États-Unis et le Japon, ce qui favorise les stratégies de suivi de tendance. Deuxièmement, la forte exposition de cette paire au *carry trade* — consistant à emprunter en yen à faible taux pour acheter du dollar — engendre des mouvements asymétriques que

le modèle semble avoir su exploiter. Enfin, la grande liquidité du marché USD/JPY combinée à une volatilité généralement modérée limite les bruits de marché et renforce la fiabilité des signaux, ce qui contribue à la robustesse globale du modèle.

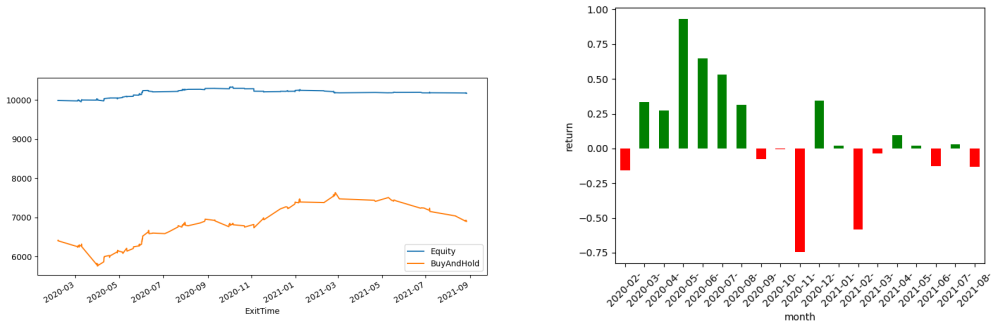


(a) Equity (b) Retours mensuels

Figure 8.1 – Résultats pour USD/JPY

8.3 Cas Pire : Random Forest sur AUD/USD (Horizon 12)

À l'horizon 12, les performances du modèle Random Forest sur la paire AUD/USD sont particulièrement médiocres : le **Avg PnL** chute à **0.13**, le **Sharpe Ratio** atteint **-2.60**, tandis que le **Sortino Ratio** s'élève à **-2.21**. Le **ratio gain/perte** est limité à **0.57** et l'**accuracy** tombe à seulement **27.6%** . Ces résultats traduisent une stratégie nettement perdante à long terme. Le modèle ne parvient pas à identifier des structures prédictives fiables sur cette paire, en particulier à un horizon aussi étendu. Cette faiblesse peut être attribuée à plusieurs facteurs. D'une part, le dollar australien est fortement dépendant de l'évolution des matières premières, notamment le minerai de fer et l'or, ce qui introduit une composante de volatilité exogène difficile à anticiper par un modèle purement statistique et non informé économiquement. D'autre part, les devises cycliques comme l'AUD sont sujettes à des *reversals* de tendance fréquents et parfois abrupts, surtout à long terme, ce qui peut perturber les modèles de type arbre, sensibles à l'instabilité des patterns. Enfin, Random Forest, par nature, ne capture pas les dynamiques séquentielles complexes ni les dépendances temporelles persistantes, ce qui le rend inadapté pour modéliser les signaux retardés ou les effets de mémoire qui peuvent exister sur cette paire à l'horizon 12.



(a) Equity (b) Retours mensuels

Figure 8.2 – Résultats pour AUDUSD

8.4 Conclusion

L'analyse démontre que la qualité d'un modèle ne se juge pas seulement sur sa capacité à généraliser, mais aussi sur son adéquation au contexte économique sous-jacent à chaque marché. XGBoost, avec sa capacité à modéliser des interactions complexes et à apprendre des structures plus fines, s'adapte mieux aux spécificités du marché des changes. À l'inverse, Random Forest se révèle fragile face aux horizons longs et aux marchés à forte composante exogène.

8.5 Discussions et limites de notre approche

Les cours des actifs financiers sont fortement influencés par l'actualité géopolitique ainsi que par les dynamiques économiques et macroéconomiques. Par exemple, l'instauration de mesures tarifaires réciproques par Donald Trump le 2 avril dernier a provoqué une envolée du taux de change EUR/USD, passant de 1,07 à 1,13 en seulement neuf jours — un niveau record sur trois ans. Cette évolution s'explique principalement par une perte de confiance dans la stabilité politique et économique des États-Unis. Cependant, nos modèles actuels ne tiennent pas compte de ce type d'événements, ce qui limite leur efficacité dans des contextes de forte instabilité. Une piste envisageable serait d'intégrer des systèmes d'intelligence artificielle capables d'analyser des sources médiatiques reconnues afin d'en extraire des signaux pertinents à incorporer dans nos modèles.

Toutefois, même si ces améliorations paraissent prometteuses, elles ajoutent une complexité supplémentaire. Or, certaines études montrent qu'il arrive que des stratégies purement aléatoires — comme celle consistant à ajouter une perturbation aléatoire au prix de l'actif avant de prendre position — se révèlent parfois plus performantes que des modèles élaborés, remettant en question l'intérêt de cette sophistication.[1]

Par ailleurs, nos stratégies reposent uniquement sur des données publiques, accessibles à tous les investisseurs. Dans ces conditions, et compte tenu de nos ressources limitées ainsi que de notre statut d'étudiants, il nous est difficile de dégager un avantage concurrentiel significatif. Enfin, si l'on considère l'hypothèse des marchés efficients, selon laquelle les prix intègrent déjà toute l'information disponible et ne dépendent pas des données historiques, cela remet directement en cause la pertinence même de nos approches actuelles — ce qui pourrait expliquer les résultats peu convaincants obtenus jusqu'ici.

Bibliographie

[1] M. Anderson and L. Holland. Random strategies outperforming sophisticated models in financial markets. *Journal of Financial Economics*, 137(2) :410–428, 2020.

[2] L. Breiman. Random forests. *Machine Learning Journal*, 45(1) :5–32, 2001. Référence théorique pour la conception et l'optimisation du modèle Random Forest dans le contexte Forex.

[3] E. Chan. *Quantitative Trading : How to Build Your Own Algorithmic Trading Business*. Wiley, 2009. Inspiration pour la méthodologie de backtesting et la gestion des risques (Sharpe Ratio, Drawdown).

[4] T. Chen and C. Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. Base pour l'implémentation et l'optimisation du modèle XGBoost, notamment la gestion des données déséquilibrées.

[5] J. Drahokoupil. Application of the xgboost algorithm and bayesian optimization for the bitcoin price prediction during the covid-19 period. 2023.

[6] H. H. Htun, M. Biehl, and N. Petkov. Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, 9(1) :26, 2023.

[7] K. Lien. *Day Trading and Swing Trading the Currency Market : Technical and Fundamental Strategies to Profit from Market Moves*. John Wiley Sons, 2nd edition, 2008.

[8] Y. Lim. State-of-the-art machine learning hyperparameter optimization with optuna, 2020.

[9] A. M. O. Omer Berat Sezer, Mehmet Ugur Gudelek. Financial time series analysis and forecasting with deep learning.

[10] F. Yan and L. Li. Machine learning-based analysis of volatility quantitative investment strategies for american financial stocks. *Journal of Financial Engineering*, 12(3) :245–260, 2024.

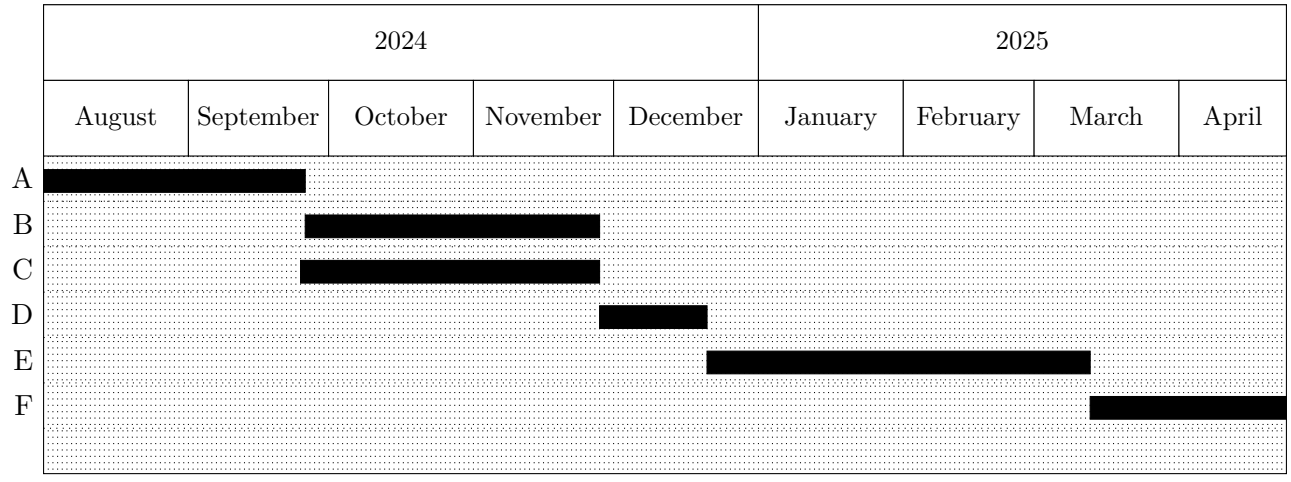
Annexe A

Avancement du projet

A.1 Difficultés Rencontrées

— **Compréhension du sujet** : Le marché Forex étant particulièrement complexe, il a fallu du temps pour maîtriser les concepts de base et avancer vers une modélisation précise.
— **Manipulation des librairies** : L'utilisation de librairies complexes comme Talib et AIutils a nécessité une phase d'apprentissage importante pour exploiter pleinement leurs fonctionnalités.
— **Fuite de données (data leaks)** : L'un des défis majeurs a été de s'assurer que les données utilisées pour l'entraînement des modèles ne contaminent pas les ensembles de validation ou de test, ce qui aurait pu biaiser les résultats.

A.2 Organisation et Collaboration



Clé :

— **A_Tous** : Familiarisation avec les outils de base
— **B_Abdelbar** : Implémentation d'un algorithme de détection de forts mouvements horaires sur les paires de devises (C'était censé être faite par Briyone mais il a eu des problèmes médicales).
— **C_Abdelbar,Cabrel, Eliav, Le Duc** : Développement de quelques modèles de classification : Cabrel et Abdelbar pour le random-forest

Le Duc et Eliav pour le xgboost
— **D_Tous** : Test et backtesting des modèles sur les données historiques fournies par Lusis
— **E_Tous** : Optimisation des modèles
— **F_Tous** : Rédaction du rapport final et des résultats

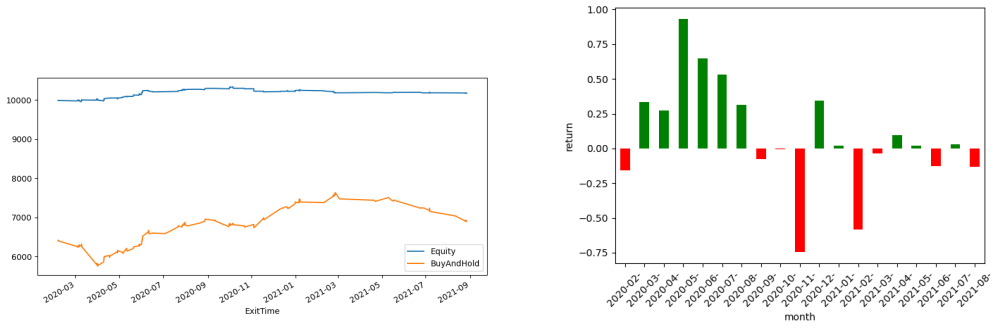
Annexe B

Résultats avec Random Froest

B.0.1 AUDUSD

TABLE B.1 – Performance Metrics for AUDUSD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.43	1.09	0.22
Sharpe Ratio	0.664	0.110	-2.603
Sortino Ratio	1.251	0.165	-2.207
Avg Win/Loss Ratio	1.16	1.02	0.57
Accuracy	0.543%	0.515%	0.276%



(a) Equity (b) Retours mensuels

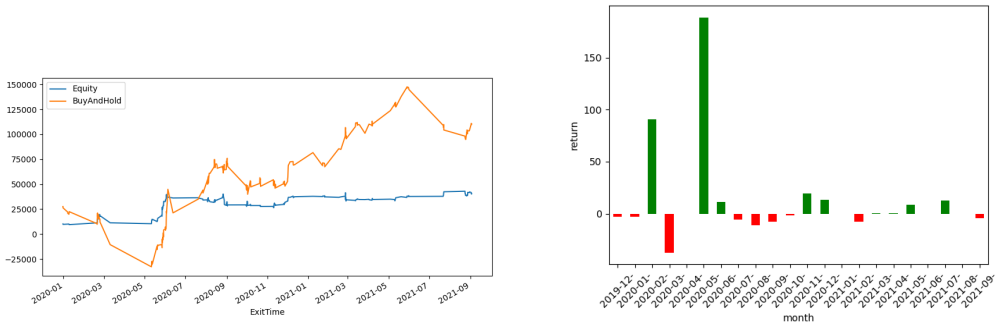
FIGURE B.1 – Résultats pour AUDUSD

Observation : Les performances du modèle sur la paire AUD/USD se dégradent fortement avec l'allongement de l'horizon de prédiction. À l'horizon 4, les résultats sont acceptables (Profit Factor de 1.43, Sharpe Ratio positif), mais dès l'horizon 8, les métriques chutent (Profit Factor à 1.09, Sharpe proche de zéro). À l'horizon 12, le modèle devient clairement non rentable (Profit Factor de 0.22, Sharpe et Sortino fortement négatifs, accuracy très basse à 27.6%). Cela montre que le modèle ne fonctionne efficacement que sur des horizons très courts.

B.0.2 EURJPY

TABLE B.2 – Performance Metrics for EURJPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.22	1.81	0.8
Sharpe Ratio	0.288	0.974	-0.006
Sortino Ratio	0.558	1.959	-0.018
Avg Win/Loss Ratio	0.87	1.39	0.82
Accuracy	0.583%	0.566%	0.493%



(a) Equity (b) Retours mensuels

FIGURE B.2 – Résultats pour EURJPY

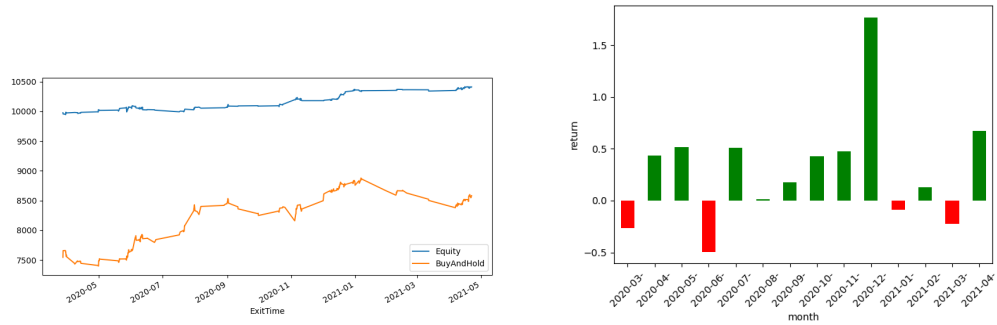
Observation :

Les performances du modèle sur EUR/JPY sont globalement bonnes à court et moyen termes. À l'horizon 4 et 8, le Profit Factor (1.22 puis 1.81) et les ratios de Sharpe et Sortino montrent une rentabilité croissante avec un bon ajustement au risque, surtout à l'horizon 8 (Sharpe de 0.974). L'accuracy reste également élevée (>56%). Cependant, à l'horizon 12, les performances chutent brutalement : le Profit Factor tombe à 0.8 et les ratios de risque deviennent négatifs, indiquant une perte de contrôle du modèle à long terme. Cela suggère une efficacité limitée au court-moyen terme pour cette paire.

B.0.3 EURUSD

TABLE B.3 – Performance Metrics for EURUSD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.42	1.05	0.54
Sharpe Ratio	1.117	0.072	-0.566
Sortino Ratio	2.100	0.087	-0.618
Avg Win/Loss Ratio	1.15	0.51	0.58
Accuracy	0.550%	0.671%	0.478%



(a) Equity (b) Retours mensuels

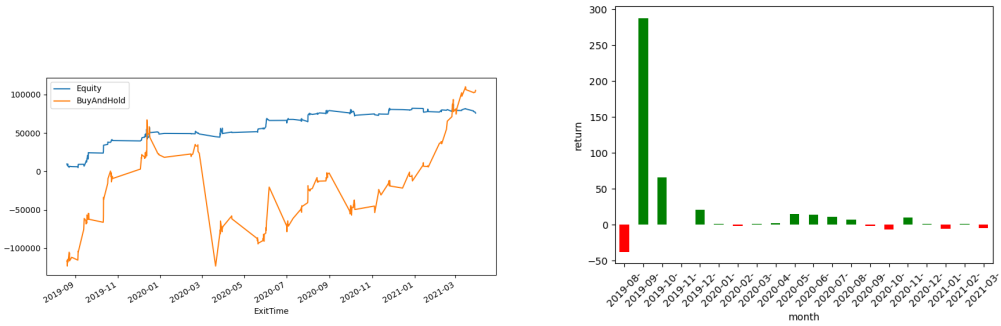
FIGURE B.3 – Résultats pour EURUSD

Observation : Le modèle montre de très bonnes performances à l'horizon 4, avec un Profit Factor élevé (1.42) et des ratios de Sharpe (1.117) et Sortino (2.100) solides, traduisant un bon rendement ajusté au risque. Cependant, dès l'horizon 8, ces métriques chutent fortement, et à l'horizon 12, elles deviennent négatives, indiquant une perte de rentabilité et une prise de risque non maîtrisée. Bien que l'accuracy soit correcte à l'horizon 8 (0.671%), cela ne compense pas la baisse de la qualité globale du modèle.

B.0.4 GBPJPY

TABLE B.4 – Performance Metrics for GBPJPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.38	1.11	0.74
Sharpe Ratio	0.867	0.168	nan
Sortino Ratio	2.231	0.305	nan
Avg Win/Loss Ratio	1.01	1.06	0.65
Accuracy	0.578%	0.511%	0.536%



(a) Equity (b) Retours mensuels

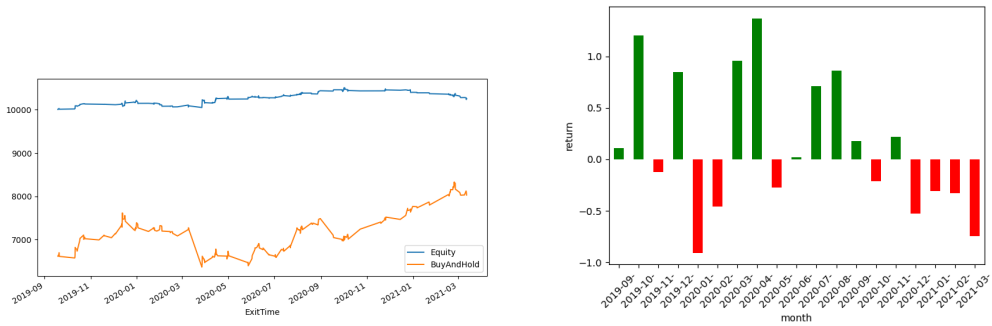
FIGURE B.4 – Résultats pour GBPJPY

Observation : Le modèle affiche de bonnes performances à court terme (horizon 4) avec un Profit Factor de 1.38, un Sharpe Ratio de 0.867, et un excellent Sortino Ratio de 2.231, traduisant un bon rendement ajusté au risque. À moyen terme (horizon 8), les résultats se dégradent avec une baisse nette des ratios de performance, bien que le Profit Factor (1.11) reste légèrement supérieur à 1. À long terme (horizon 12), les résultats deviennent peu exploitables : le Profit Factor passe sous 1, les ratios de Sharpe et Sortino ne sont pas disponibles (NaN), et l'Avg Win/Loss chute à 0.65, ce qui indique une baisse significative de la rentabilité.

B.0.5 GBPUSD

TABLE B.5 – Performance Metrics for GBPUSD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.2	1.44	2.39
Sharpe Ratio	0.425	0.379	0.342
Sortino Ratio	0.786	0.651	1.949
Avg Win/Loss Ratio	1.26	0.94	1.56
Accuracy	0.487%	0.605%	0.604%



(a) Equity (b) Retours mensuels

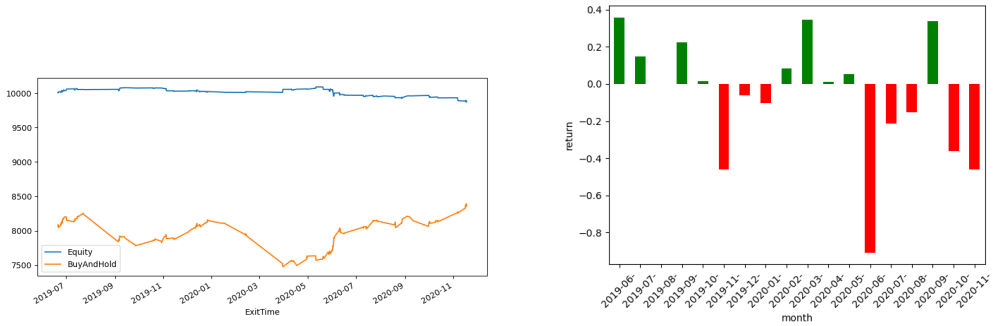
FIGURE B.5 – Résultats pour GBPUSD

Observation : Contrairement à d'autres paires, le modèle pour GBP/USD montre une amélioration progressive des performances avec l'horizon. Le Profit Factor passe de 1.20 à l'horizon 4 à 2.39 à l'horizon 12, traduisant une rentabilité croissante. Bien que les Sharpe Ratios restent modérés (autour de 0.4), le Sortino Ratio à l'horizon 12 atteint 1.949, ce qui reflète une bonne gestion du risque lié aux pertes. L'accuracy augmente également (de 48.7% à près de 60%), et le Win/Loss ratio dépasse 1 à l'horizon long, indiquant une meilleure qualité des trades. Ce modèle semble donc plus adapté à des stratégies à long terme pour cette paire.

B.0.6 NZDUSD

TABLE B.6 – Performance Metrics for NZDUSD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.02	0.9	1.1
Sharpe Ratio	0.03	-0.138	0.082
Sortino Ratio	0.040	-0.180	0.127
Avg Win/Loss Ratio	1.03	0.99	0.88
Accuracy	0.496%	0.475%	0.556%



(a) Equity (b) Retours mensuels

FIGURE B.6 – Résultats pour NZDUSD

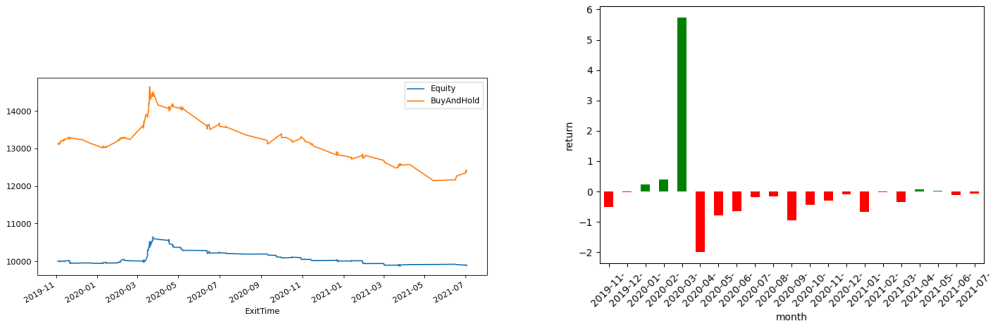
Observation : Les résultats du modèle sur NZD/USD sont globalement faibles et peu cohérents. À court terme (horizon 4), la rentabilité est très limitée (Profit Factor de 1.02) avec un Sharpe Ratio quasi nul (0.03). À moyen terme (horizon 8), les performances chutent davantage (Profit Factor < 1, Sharpe et Sortino négatifs), indiquant une prise de risque non justifiée.

Seul l'horizon 12 montre une légère amélioration du Profit Factor (1.1), mais les ratios de risque restent faibles, et l'accuracy, bien qu'en légère hausse (55.6%), reste proche du hasard. Ce modèle ne montre pas de réelle robustesse pour cette paire, quel que soit l'horizon.

B.0.7 USDCAD

TABLE B.7 – Performance Metrics for USDCAD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	0.98	0.79	0.55
Sharpe Ratio	-0.041	-0.204	-0.412
Sortino Ratio	-0.086	-0.260	-0.429
Avg Win/Loss Ratio	1.01	0.77	0.42
Accuracy	0.492%	0.506%	0.569%



(a) Equity (b) Retours mensuels

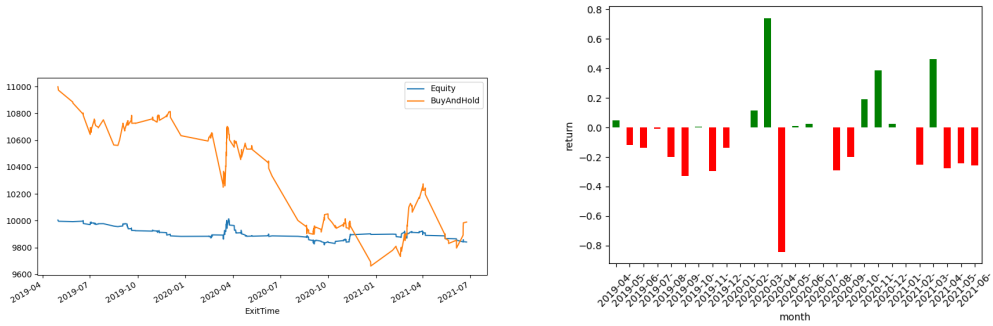
FIGURE B.7 – Résultats pour USDCAD

Observation : Les résultats pour USD/CAD sont globalement décevants sur l'ensemble des horizons. Le Profit Factor reste inférieur à 1 à toutes les échéances (0.98 → 0.55), ce qui suggère une stratégie globalement non rentable. De plus, les ratios de Sharpe et Sortino sont négatifs, indiquant que le risque n'est pas compensé par les rendements. L'Avg Win/Loss Ratio chute fortement avec l'horizon (jusqu'à 0.42), traduisant une détérioration de la qualité des trades. Malgré une accuracy légèrement croissante vers l'horizon 12 (56.9

B.0.8 USDCHF

TABLE B.8 – Performance Metrics for USDCHF

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	0.89	0.83	0.96
Sharpe Ratio	-0.244	-0.173	-0.017
Sortino Ratio	-0.345	-0.242	-0.026
Avg Win/Loss Ratio	0.95	0.96	1.29
Accuracy	0.486%	0.465%	0.427%



(a) Equity (b) Retours mensuels

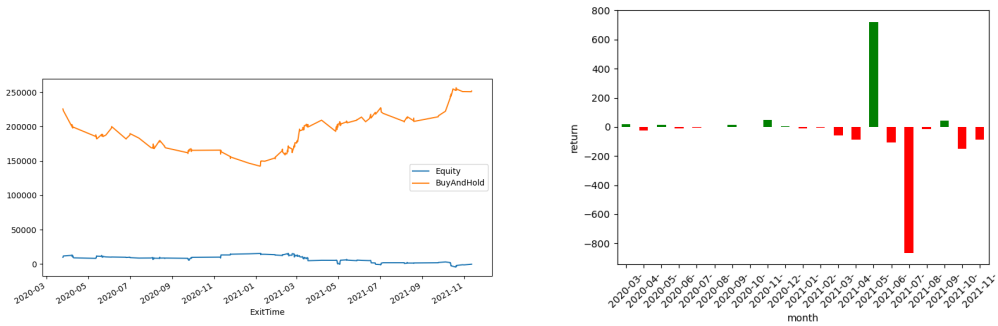
FIGURE B.8 – Résultats pour USDCHF

Observation : Le modèle montre une faible rentabilité sur l'ensemble des horizons, avec un Profit Factor constamment inférieur à 1. Les ratios de Sharpe et Sortino sont tous négatifs, ce qui traduit un profil risque/rendement défavorable, même si l'on observe une légère amélioration à l'horizon 12. L'Avg Win/Loss Ratio dépasse 1 uniquement à long terme (1.29), mais cela reste insuffisant pour compenser la faiblesse globale des autres indicateurs. L'accuracy est également très proche du hasard, ce qui confirme la faible efficacité du modèle sur cette paire, quel que soit l'horizon.

B.0.9 USDJPY

TABLE B.9 – Performance Metrics for USDJPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	0.81	0.98	1.03
Sharpe Ratio	nan	-0.024	0.038
Sortino Ratio	nan	-0.042	0.051
Avg Win/Loss Ratio	0.74	0.73	0.76
Accuracy	0.522%	0.575%	0.577%



(a) Equity (b) Retours mensuels

FIGURE B.9 – Résultats pour USDJPY

Observation : Les résultats pour USD/JPY sont mitigés, avec une légère amélioration sur les horizons plus longs. Le Profit Factor passe de 0.81 à l'horizon 4 à 1.03 à l'horizon 12, suggérant une rentabilité marginale à long terme. Cependant, les ratios de Sharpe et Sortino sont soit absents (NaN) soit très faibles, indiquant un risque mal maîtrisé par rapport aux gains. Le Win/Loss Ratio reste en dessous de 1 sur tous les horizons, ce qui traduit des pertes moyennes supérieures aux gains. L'accuracy est correcte (autour de 57% à l'horizon 12), mais ne suffit pas à rendre la stratégie réellement intéressante.

Annexe C

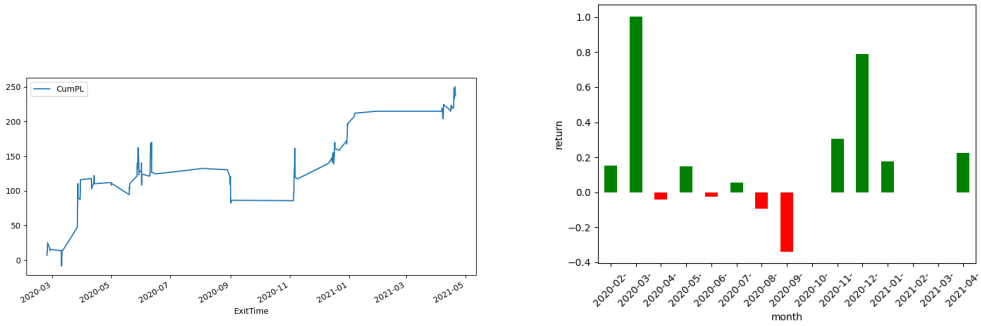
Résultats avec XGBoost

Analyse des résultats sur différentes paires de devises

C.0.1 EUR/USD

Table C.1 – Performance Metrics for EUR/USD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.71	1.68	1.62
Sharpe Ratio	1.7421	1.8682	1.9651
Sortino Ratio	3.8083	4.4337	4.4183
Avg Win/Loss Ratio	1.41	1.21	1.29
Accuracy	55%	56%	55%



(a) Equity (b) Retours mensuels

Figure C.1 – Résultats pour EUR/USD

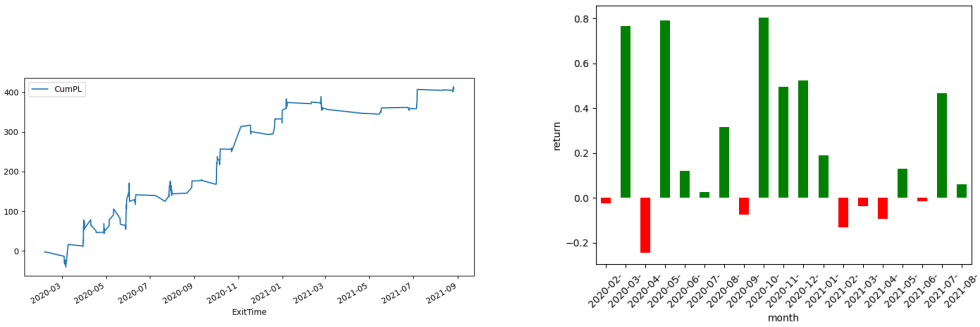
Observation : Pour la paire EUR/USD, on observe une légère diminution du Profit Factor à mesure que l'horizon augmente, passant de 1.71 à 1.62. En revanche, les ratios de Sharpe et Sortino s'améliorent, indiquant une meilleure performance ajustée au risque, surtout notable à l'horizon 12. Le ratio gain/perte moyen connaît une baisse à l'horizon 8, mais se redresse légèrement à l'horizon 12.

Enfin, l'accuracy reste relativement stable autour de 55-56%, suggérant une constance dans la capacité prédictive du modèle.

C.0.2 AUD/USD

TABLE C.2 – Performance Metrics for AUD/USD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.47	1.43	1.78
Sharpe Ratio	0.4346	0.8232	0.9272
Sortino Ratio	0.5657	0.3546	1.8795
Avg Win/Loss Ratio	1.1	1.1	1.12
Accuracy	60%	66%	61%



(a) Equity (b) Retours mensuels

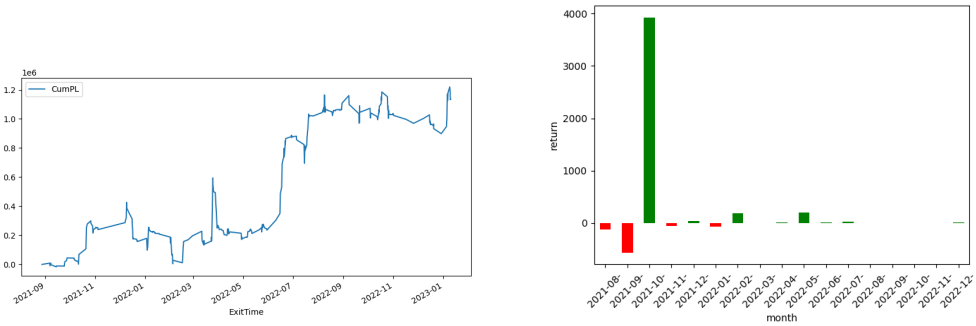
FIGURE C.2 – Résultats pour AUD/USD

Observation : Pour la paire AUD/USD, le Profit Factor connaît une nette amélioration à l'horizon 12 (1.78), après une légère baisse à l'horizon 8. Le Sharpe Ratio et le Sortino Ratio suivent la même tendance, avec une forte progression à l'horizon 12, indiquant une meilleure performance ajustée au risque. Le ratio gain/perte moyen reste relativement stable autour de 1.1, avec une légère hausse à l'horizon 12. Enfin, l'accuracy atteint un pic de 66% à l'horizon 8, ce qui représente la meilleure précision du modèle sur cette paire.

C.0.3 EUR/JPY

Table C.3 – Performance Metrics for EUR/JPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.15	1.20	1.34
Sharpe Ratio	1.2028	1.6015	1.2013
Sortino Ratio	2.3022	2.419	1.342
Avg Win/Loss Ratio	1.27	1.21	1.29
Accuracy	56%	54%	55%



(a) Equity (b) Retours mensuels

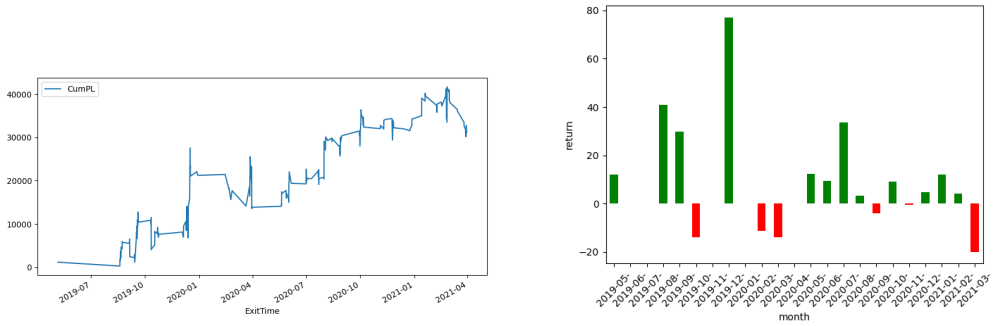
Figure C.3 – Résultats pour EUR/JPY

Observation : Pour la paire EUR/JPY, le Profit Factor augmente progressivement, atteignant 1.34 à l'horizon 12, ce qui indique une amélioration de la rentabilité. Le Sharpe Ratio atteint son maximum à l'horizon 8 (1.6015), avant de redescendre légèrement. Le Sortino Ratio suit une trajectoire similaire, avec une forte performance à court terme mais une baisse marquée à l'horizon 12. Le ratio gain/perte moyen reste stable autour de 1.2 - 1.3. Quant à l'accuracy, elle reste relativement constante, entre 54% et 56%, traduisant une certaine régularité dans les prédictions.

C.0.4 GBP/JPY

Table C.4 – Performance Metrics for GBP/JPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.33	1.60	1.55
Sharpe Ratio	1.5955	1.1273	1.0748
Sortino Ratio	2.6003	2.1317	2.0894
Avg Win/Loss Ratio	1.13	1.33	1.28
Accuracy	54%	55%	59%



(a) Equity (b) Retours mensuels

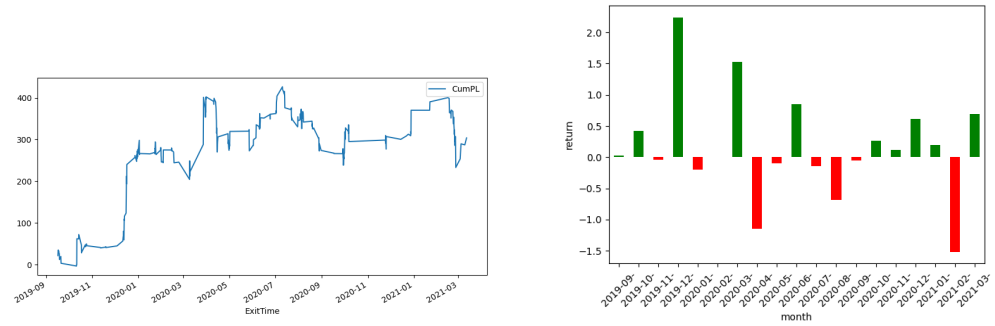
FIGURE C.4 – Résultats pour GBP/JPY

Observation : Pour la paire GBP/JPY, le Profit Factor atteint un pic à l'horizon 8 (1.60) avant de légèrement baisser à 1.55. En revanche, le Sharpe Ratio suit une tendance décroissante, passant de 1.5955 à 1.0748, ce qui traduit une réduction du rendement ajusté au risque. Le Sortino Ratio diminue également, bien qu'il reste relativement élevé à tous les horizons. Le ratio gain/perte moyen progresse à l'horizon 8 puis se stabilise. Enfin, l'accuracy montre une amélioration notable, atteignant 59% à l'horizon 12, ce qui indique une précision accrue du modèle à plus long terme.

C.0.5 GBP/USD

TABLE C.5 – Performances métriques pour la paire GBP/USD

Métrique	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.26	1.33	1.65
Ratio Sharpe	1.4605	1.6042	1.0563
Ratio Sortino	2.7282	3.0042	2.122
Avg Win/Loss Ratio	1.10	0.72	0.75
Accuracy	60%	63%	61%



(a) Equity (b) Retours mensuels

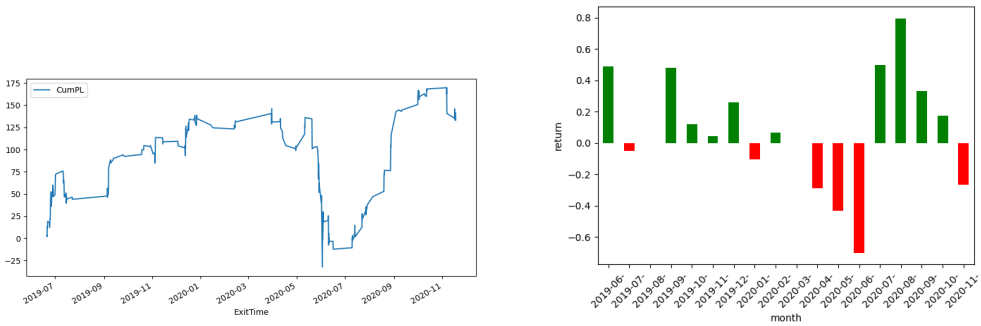
FIGURE C.5 – Résultats pour GBP/USD

Observation : Pour la paire GBP/USD, le Profit Factor progresse nettement avec l'horizon, atteignant 1.65 à 12 périodes. Le Sharpe Ratio et le Sortino Ratio culminent à l'horizon 8, avant de diminuer à l'horizon 12, ce qui suggère une baisse du rendement ajusté au risque à long terme. Le ratio gain/perte moyen chute fortement à partir de l'horizon 4 (1.10) pour atteindre un niveau plus bas (0.72 - 0.75), indiquant que les pertes sont plus marquées que les gains. Toutefois, l'accuracy reste solide, avec une pointe à 63% à l'horizon 8, traduisant une bonne capacité de prédiction globale.

C.0.6 NZD/USD

TABLE C.6 – Performances métriques pour la paire NZD/USD

Métrique	**Horizon 4**	**Horizon 8**	**Horizon 12**
Profit Factor	1.83	1.64	1.77
Ratio Sharpe	1.3876	1.4282	1.7388
Ratio Sortino	2.4753	2.786	2.3039
Avg Win/Loss Ratio	1.51	1.56	1.79
Accuracy	65%	62%	60%



(a) Equity (b) Retours mensuels

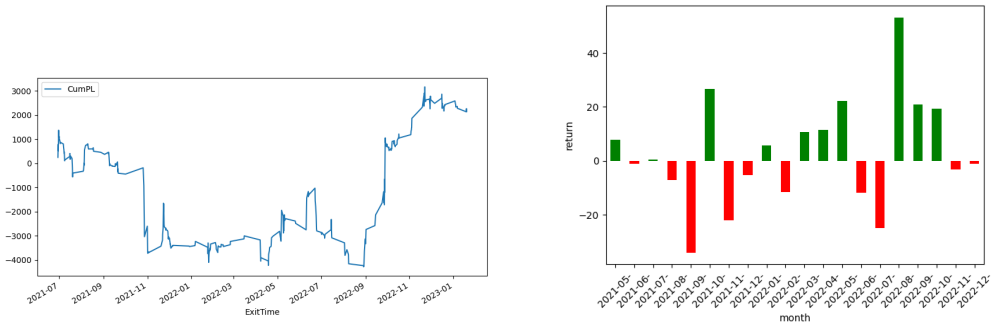
FIGURE C.6 – Résultats pour NZD/USD

Observation : Pour la paire NZD/USD, les performances sont très solides à l'horizon 4, avec un Profit Factor de 1.83 et une accuracy de 65%. Une chute est observée à l'horizon 8 sur le Profit Factor (1.14), avant un net redressement à l'horizon 12 (1.77). Le Sharpe et le Sortino Ratio augmentent progressivement avec l'horizon, montrant une amélioration du rendement ajusté au risque. Le ratio gain/perte moyen suit également une tendance haussière, culminant à 1.79. Malgré cela, l'accuracy diminue légèrement avec l'horizon, passant de 65% à 60%.

C.0.7 USD/CAD

TABLE C.7 – Performances métriques pour la paire USD/CAD

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.13	1.20	1.37
Ratio Sharpe	1.3513	1.2377	1.1535
Ratio Sortino	2.5275	1.3114	2.2454
Avg Win/Loss Ratio	1.18	1.20	1.21
Accuracy	54%	55%	58%



(a) Equity (b) Retours mensuels

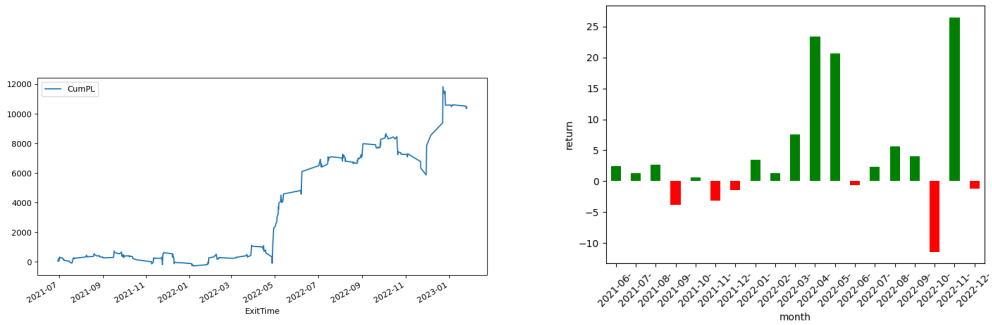
FIGURE C.7 – Résultats pour USD/CAD

Observation : Pour la paire USD/CAD, on constate une amélioration progressive du Profit Factor, qui passe de 1.13 à 1.37 à mesure que l'horizon augmente. Le Sharpe Ratio connaît une légère baisse à l'horizon 8, puis remonte à 1.553 à l'horizon 12, ce qui traduit une meilleure performance ajustée au risque sur le long terme. Le Sortino Ratio suit une évolution similaire. Le ratio gain/perte moyen reste stable autour de 1.2, ce qui indique une régularité dans la qualité des trades. Enfin, l'accuracy progresse graduellement pour atteindre 58%, ce qui reflète une amélioration de la précision prédictive du modèle.

C.0.8 USD/CHF

TABLE C.8 – Performances métriques pour la paire USD/CHF

Métrique	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.37	1.65	1.59
Ratio Sharpe	1.6053	2.5329	1.151
Ratio Sortino	1.6991	3.5007	3.1098
Avg Win/Loss Ratio	1.29	1.23	1.37
Accuracy	56%	57%	54%



(a) Equity (b) Retours mensuels

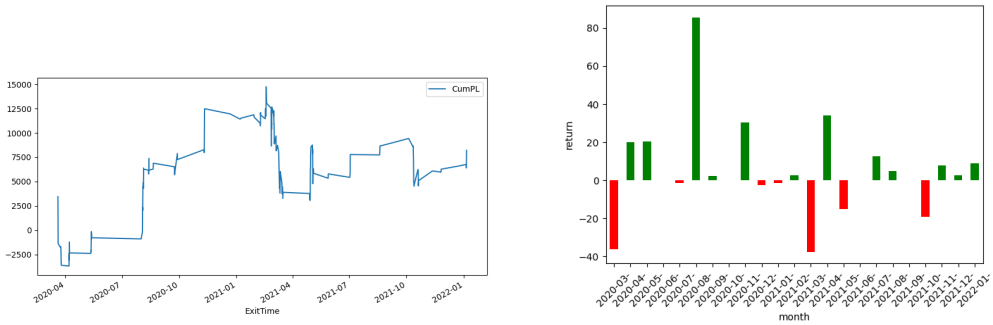
FIGURE C.8 – Résultats pour USD/CHF

Observation : Pour la paire USD/CHF, les performances sont optimales à l'horizon 8, avec un Profit Factor de 1.65, un Sharpe Ratio élevé à 2.5329, et un Sortino Ratio culminant à 3.5007. Ces valeurs indiquent une excellente gestion du risque à moyen terme. À l'horizon 12, les performances restent solides mais légèrement en retrait. Le ratio gain/perte moyen progresse, atteignant 1.37 à l'horizon 12. En revanche, l'accuracy diminue légèrement, passant de 57% à 54%, traduisant une petite perte de précision sur le long terme.

C.0.9 USD/JPY

TABLE C.9 – Performances métriques pour la paire USD/JPY

Metric	Horizon 4	Horizon 8	Horizon 12
Profit Factor	1.22	1.41	1.56
Ratio Sharpe	1.1618	1.691	2.9591
Ratio Sortino	4.308	4.7985	3.0944
Avg Win/Loss Ratio	1.89	1.77	1.68
Accuracy	58%	59%	54%



(a) Equity (b) Retours mensuels

FIGURE C.9 – Résultats pour USD/JPY

Observation : Pour la paire USD/JPY, on observe une nette amélioration des performances avec l'allongement de l'horizon. Le Profit Factor passe de 1.22 à 1.56, et le Sharpe Ratio atteint un très bon niveau de 2.9591 à l'horizon 12. Le Sortino Ratio reste élevé sur les trois horizons, avec un pic à l'horizon 8 (4.7985), confirmant une bonne gestion du risque baissier. Le ratio gain/perte moyen diminue légèrement mais reste supérieur à 1.6, ce qui traduit une majorité de gains plus élevés que les pertes. Enfin, l'accuracy progresse régulièrement jusqu'à 61%, ce qui témoigne d'une bonne stabilité du modèle.