Modèle d'apprentissage machine pour la prédiction de marché financier

Antoine Boubée
Mathéo Ruanlt

Printemps 2025

1 Introduction

1.1 Finance

Définition de la finance La finance renvoie à un domaine d'activité, celui du financement, aujourd'hui mondialisé, qui consiste à fournir ou à obtenir les capitaux ou les produits financiers nécessaires à la réalisation d'une opération économique. Elle permet de faire transiter des capitaux des agents économiques excédentaires (disposant d'une épargne à faire fructifier) vers les agents économiques déficitaires (ayant besoin de financement pour investir, croître, etc.)

Limites de la prédiction financière La prédiction de l'évolution des marchés est extrêmement difficile. Il est inconcevable de prédire précisément comment le sentiment humain réagira à une nouvelle géopolitique, économique ou politique influant sur les marchés. Les crises financières illustrent bien la complexité et l'imprévisibilité du système.

Cependant, malgré ces incertitudes, il est parfois possible d'avoir une intuition sur la tendance d'un actif. Par exemple, à l'annonce de la sortie d'un nouveau produit par Apple (comme le Vision Pro), on peut supposer que l'action Apple augmentera, en raison de l'anticipation de profits et de l'enthousiasme des investisseurs. Mais cette intuition reste hypothétique et sujette à de nombreuses variables.

De plus, les données financières sont abondantes : prix, volumes, volatilité, corrélations entre actifs, ETF, actualités économiques et politiques, etc.

L'information comme levier stratégique Les fonds de financement et de trading ont pour objectif d'obtenir un PnL (Profit and Loss) positif en fin d'année, et idéalement de battre le marché. L'information constitue l'un des leviers les plus puissants pour y parvenir. Certains fonds paient pour obtenir des données confidentielles sur les entreprises dans lesquelles ils investissent, ce qui leur permet parfois de prendre l'avantage sur le marché. En tant que particulier, il est simplement impossible d'avoir accès à de telles données, difficile est d'avoir les compétences pour les interpréter et très compliqué d'avoir les financements nécessaires pour réaliser des profits additionnels conséquent.

1.2 Machine Learning

Évolution des modèles prédictifs en finance Depuis des décennies, les modèles de régression de séries temporelles ont constitué une pierre angulaire dans le développement des méthodes d'évaluation financière. Cette approche est essentielle tant dans les modèles financiers traditionnels que dans l'intelligence artificielle pour la prévision financière, un domaine marqué par la complexité et l'imprévisibilité des tendances du marché.

L'analyse traditionnelle des marchés financiers repose notamment sur le modèle à trois facteurs de Fama-French (FFM) et la théorie de l'évaluation par

arbitrage (APT) de Chen, Roll et Ross. Ces approches sont cruciales dans la détermination des prix des actifs et utilisent la régression linéaire pour analyser les rendements. Toutefois, elles ne permettent pas de capturer les fluctuations spécifiques du marché. Leur forte dépendance aux données historiques limite leur efficacité pour anticiper les changements futurs ou les événements sans précédent, comme les crises financières.

Cependant, les techniques émergentes d'apprentissage automatique (machine learning, ML) se sont révélées prometteuses pour surmonter ces limitations. Des études antérieures démontrent l'efficacité des approches ML comparées aux modèles traditionnels. En particulier, Billah et Bhuiyan ont souligné la supériorité de l'intégration conjointe du prix des actions et du sentiment issu des actualités dans les techniques d'apprentissage profond (deep learning, DL) pour la prédiction du marché boursier. Ces méthodes utilisent des architectures telles que les réseaux de neurones à mémoire à long terme (LSTM), les réseaux neuronaux récurrents (RNN) et les méthodes d'apprentissage par renforcement (reinforcement learning, RL), qui ont démontré une amélioration substantielle dans la prévision des mouvements du marché, là où les modèles traditionnels échouaient souvent.

La théorie moderne du portefeuille de Harry Markowitz met en évidence l'importance de la corrélation entre actifs. Des recherches récentes ont confirmé une corrélation positive significative entre les informations de sentiment — issues des actualités, des blogs ou des réseaux sociaux — et les tendances du marché boursier. L'avènement de modèles de langage de grande taille (large language models, LLM) tels que ChatGPT ou GPT-4, développés par OpenAI, a amélioré la précision de l'analyse des sentiments dans ce contexte. Alors que les recherches de Lopez-Lira suggéraient que les versions précédentes comme GPT-3 peinaient à prévoir correctement les rendements du marché, les modèles les plus récents, comme GPT-4, ont atteint des ratios de Sharpe parmi les plus élevés, démontrant une fiabilité nettement accrue.

Limites des ensembles de données actuels et perspectives Cependant, le manque d'ensembles de données complets et intégrés a considérablement freiné l'avancement de la recherche, notamment dans la mise en œuvre de modèles sophistiqués fondés sur l'architecture des transformateurs. Ces modèles, s'ils sont correctement alimentés, pourraient améliorer de manière significative l'analyse prédictive en finance.

Pour combler cette lacune, plusieurs ensembles de données antérieurs ont été mobilisés, tels que les actualités de Philips issues de Bloomberg et Reuters, ou celles de Yutkin extraites de Lenta, ainsi que les contributions de sources spécialisées comme Benzinga. Toutefois, ces jeux de données souffrent souvent d'un volume insuffisant pour l'entraînement de modèles de grande taille et n'incluent pas systématiquement les prix des actions associés.

Dans ce contexte, l'introduction d'un ensemble de données vaste et diversifié, intégrant à la fois les prix des actions et les informations de sentiment issues de la presse financière, comme le *Financial News and Stock Price Integration*

Dataset (FNSPID), apparaît comme une avancée cruciale. Il constitue un levier important pour faire progresser la modélisation prédictive et l'analyse de marché assistée par l'intelligence artificielle.

2 Projet

Initialement, nous avions pour objectif d'exploiter le jeu de données du FNSPID afin d'entraîner un modèle de prédiction. Toutefois, la taille particulièrement importante de ce dataset a rapidement révélé des limitations matérielles, rendant son traitement complet impraticable sur nos infrastructures locales. En conséquence, nous avons opté pour une approche alternative visant à capter la tendance sentimentale du marché en nous appuyant sur les données issues de Google Trends. Plus précisément, nous analysons les volumes de recherche relatifs à des mots-clés pertinents associés à l'action étudiée, considérés comme un proxy indirect du sentiment des investisseurs.

Objectif du projet

Ce projet propose une analyse approfondie ainsi qu'une stratégie de trading algorithmique basée sur l'apprentissage automatique, visant à prédire les mouvements du cours de l'action Apple Inc. (AAPL). L'objectif principal est de concevoir un modèle prédictif robuste capable de générer des signaux de trading rentables, tout en intégrant une gestion rigoureuse des risques.

```
projet/
 aaplAnalysis_wo_sentiment.ipynb   # Notebook d'analyse
 dataset/
    full_history/AAPL.csv          # Donnes historiques d'Apple
    googleTrend.csv                # Donnes de tendances Google
 README.md                         # Documentation
```

Données d'entrée

Les données financières utilisées dans ce projet proviennent du fichier AAPL.csv, qui contient les informations de marché classiques : prix d'ouverture, de clôture, plus haut et plus bas de la journée, ainsi que le volume échangé. Ce fichier couvre une période allant de 1980 à 2023. À partir de ces variables de base, nous avons construit un ensemble d'indicateurs techniques afin d'enrichir notre jeu de données. Parmi ceux-ci figurent notamment des moyennes mobiles, des indicateurs de momentum (tels que le RSI ou le MACD), ainsi que des mesures de volatilité (comme l'ATR ou les bandes de Bollinger). Au total, notre base de données regroupe environ quarante indicateurs techniques, qui serviront de variables explicatives pour l'entraînement du modèle prédictif. On vient également calculer les variations quotidiennes.

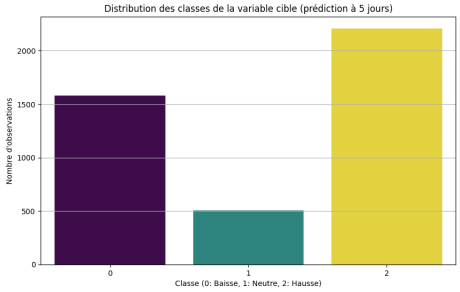

FIGURE 1 – Distribution des classes de la variable cible

En complément de ces indicateurs financiers, nous avons intégré des données issues de Google Trends. Pour chaque mois depuis 2006, nous avons récupéré les volumes de recherche relatifs aux mots-clés “Apple” et “iPhone”. Ces données étant mensuelles, nous avons appliqué une interpolation afin d’estimer des valeurs quotidiennes, compatibles avec la granularité des données de marché.

Après la fusion des deux datasets, on obtient un dataset unifié de 51 features diversifiés.

Création de la Variable Cible

Afin de formuler un problème de classification supervisée, nous avons défini une variable cible basée sur l’évolution future du prix de l’actif étudié. Plus précisément, nous avons calculé les **rendements futurs** à différentes échéances : **1, 3, 5 et 10 jours** après la date d’observation. Ces rendements sont donnés par la formule suivante :

$$R_t^{(n)} = \frac{P_{t+n} - P_t}{P_t}$$

où P_t est le prix à la date t, et P_{t+n} le prix à $t + n$ jours.

Ces rendements ont ensuite été transformés en une **variable de classification tri-classe** :

— **Classe 0 (Baisse)** : si le rendement est inférieur à $-0{,}5\%$

— **Classe 1 (Neutre)** : si le rendement est compris entre $-0{,}5\%$ et $+0{,}5\%$

— **Classe 2 (Hausse)** : si le rendement est supérieur à $+0{,}5\%$

Ce **seuil de significativité de 0,5 %** a été introduit afin d’ignorer les petites variations considérées comme du bruit de marché.

Parmi les différentes échéances, nous avons retenu le **rendement à 5 jours** comme **variable cible principale** pour l’entraînement des modèles. Ce choix permet de capter une dynamique de court/moyen terme tout en conservant une certaine réactivité aux changements de tendance.

Pipeline de Prétraitement pour Données Temporelles

Ce pipeline de prétraitement a été conçu pour optimiser les performances des modèles prédictifs tout en respectant la structure temporelle des données. Il intègre plusieurs étapes successives visant à transformer, normaliser, réduire la dimensionnalité et valider les modèles de manière robuste et sans fuite d'information.

2.1 Étapes du pipeline

1. **Division temporelle (80/20)** : séparation du jeu de données en ensembles d'entraînement et de test selon l'ordre chronologique. Cette méthode garantit l'absence de fuite d'information (data leakage) et simule un contexte réaliste de prédiction.
2. **Transformation de puissance (PowerTransformer - Yeo-Johnson)** : cette transformation permet de rendre les distributions des variables plus symétriques et proches de la normalité, y compris en présence de valeurs nulles ou négatives, ce qui est bénéfique pour la convergence de nombreux modèles.
3. **Standardisation robuste (RobustScaler)** : cette étape applique une standardisation basée sur la médiane et l'écart interquartile, ce qui permet de réduire l'influence des valeurs aberrantes.
4. **Sélection de variables (SelectKBest - information mutuelle)** : sélection automatique des $k = 30$ variables les plus informatives selon une mesure d'information mutuelle avec la variable cible. Cette méthode prend en compte des relations non linéaires.
5. **Validation croisée temporelle (TimeSeriesSplit)** : la performance des modèles est évaluée à l'aide d'une validation croisée spécifique aux séries temporelles. Celle-ci divise les données en 5 blocs successifs en respectant strictement l'ordre temporel.

2.2 Fonctionnalités clés

— **Préservation de la structure temporelle** : chaque étape du pipeline est construite de manière à respecter la causalité temporelle, évitant ainsi toute fuite d'information.
— **Réduction automatique de la dimension** : la sélection des variables permet de se concentrer sur les descripteurs les plus pertinents, facilitant ainsi l'interprétation des modèles et réduisant le surapprentissage.
— **Visualisation des découpages temporels** : les splits issus de la validation croisée sont visualisés pour assurer leur cohérence et vérifier la représentativité des différentes phases d'entraînement et de validation.

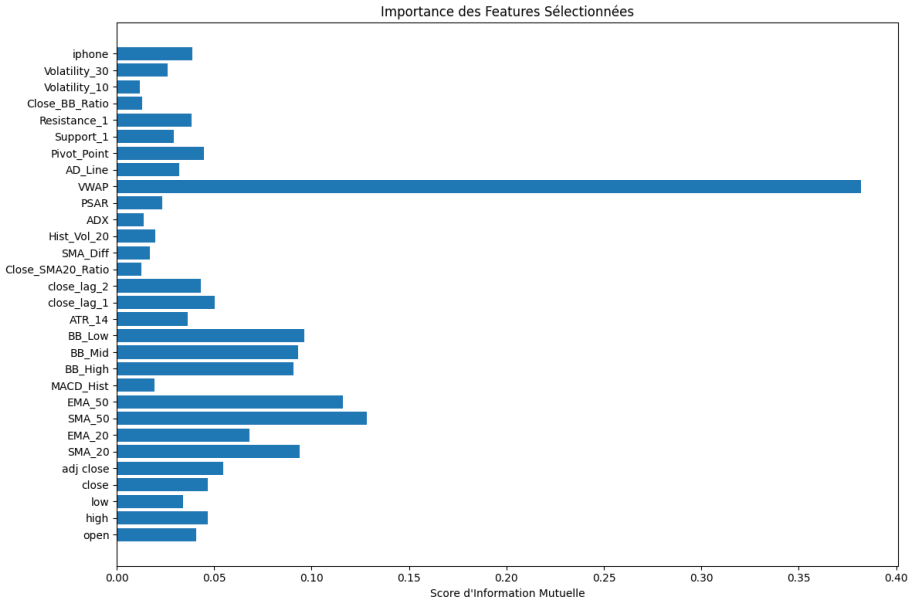

FIGURE 2 – Tableau de l'importance des chaque features

— **Analyse de l'importance des variables** : un graphique d'importance permet de mettre en évidence les variables sélectionnées, offrant une meilleure compréhension des facteurs prédictifs du modèle.

2.3 Équilibrage des classes avec SMOTE

Dans de nombreuses tâches de classification, les données présentent un déséquilibre marqué entre les classes, ce qui peut biaiser l'entraînement du modèle en faveur de la classe majoritaire. Pour remédier à cette situation, la méthode *SMOTE* (Synthetic Minority Over-sampling Technique) a été utilisée.

SMOTE agit en générant de nouvelles instances synthétiques de la classe minoritaire à partir d'interpolations entre échantillons voisins. Contrairement à un simple suréchantillonnage avec répétition, cette approche permet d'élargir la représentation de la classe minoritaire dans l'espace des caractéristiques, tout en réduisant le risque de surapprentissage.

La procédure suivie est la suivante :

— Inspection de la distribution des classes dans l'ensemble d'entraînement.

— Application de SMOTE pour équilibrer les effectifs des classes.

— Vérification de la distribution obtenue après rééchantillonnage.

Un mécanisme de gestion d'erreur a été intégré afin d'assurer la robustesse de la procédure : en cas d'échec de l'application de SMOTE (par exemple, en présence d'un nombre insuffisant d'échantillons minoritaires), l'entraînement se poursuit avec les données originales.

Ce rééchantillonnage a été effectué uniquement sur l'ensemble d'entraînement, afin de préserver l'intégrité de la validation et d'éviter tout biais d'évaluations.

3 Entraînement, Optimisation et Évaluation des Modèles

L'objectif de cette phase est d'entraîner plusieurs modèles de classification supervisée pour prédire les mouvements futurs d'un actif financier. Chaque modèle est optimisé à l'aide d'une recherche aléatoire sur grille (`RandomizedSearchCV`) qui permet d'optimiser rapidement, avec validation croisée temporelle (`TimeSeriesSplit`) pour éviter toute fuite d'information dans le temps. Les performances sont ensuite évaluées sur un jeu de test indépendant selon des métriques standards et personnalisées.

3.1 Modèles Entraînés

Les modèles suivants ont été entraînés, chacun selon un paradigme d'apprentissage différent :

- **XGBoost** : algorithme de boosting par gradient très performant basé sur des arbres de décision. Il corrige les erreurs des arbres précédents à chaque itération pour réduire l'erreur globale.
- **Random Forest** : ensemble d'arbres de décision entraînés sur des sous-échantillons aléatoires de données. La prédiction finale résulte du vote majoritaire, ce qui réduit le sur-apprentissage.
- **Gradient Boosting** : méthode d'optimisation séquentielle où chaque nouvel arbre est ajusté pour corriger les résidus du modèle précédent. Plus lent que XGBoost, il reste néanmoins efficace sur des structures simples.
- **Régression Logistique** : modèle linéaire probabiliste qui estime directement la probabilité qu'un exemple appartienne à une classe. Simple, interprétable et robuste lorsque les relations sont linéaires.
- **Réseau de Neurones (MLP)** : perceptron multicouche capable de modéliser des relations complexes non linéaires grâce à ses couches cachées. Il nécessite davantage de données et un réglage fin des hyperparamètres.
- **Ensemble Pondéré** (`VotingClassifier`) : combinaison pondérée des prédictions des modèles précédents. Chaque modèle contribue à la décision finale proportionnellement à ses performances observées.

3.2 Métriques d'Évaluation

Les modèles sont comparés à l'aide des métriques suivantes :

- **Accuracy** (exactitude)
- **AUC** (aire sous la courbe ROC)
- **F1 Score**
- **Précision**

— **Rappel**
— **AP** (Average Precision)
— **LogLoss** : perte logarithmique
— **Profit Score** : indicateur de performance financière personnalisé où chaque vrai positif génère +1% et chaque faux positif -0,5%.

3.3 Résultats Expérimentaux

Les performances obtenues sur l'ensemble de test sont résumées dans le tableau ci-dessous :

TABLE 1 – Performances des modèles optimisés sur le jeu de test

Modèle	**Accuracy**	**AUC**	**F1**	**Précision**	**Rappel**	**AP**	**LogLoss**	**Profit**
Régression Logistique	0.548	0.601	0.346	0.589	0.245	0.581	0.685	0.670
XGBoost	0.492	0.528	0.653	0.490	0.981	0.514	0.703	**1.975**
Ensemble Pondéré	0.464	0.491	0.565	0.468	0.712	0.525	0.709	1.290
Réseau de Neurones	0.462	0.488	0.561	0.466	0.705	0.522	1.049	1.265
Gradient Boosting	0.478	0.486	0.288	0.431	0.217	0.482	—	—
Random Forest	0.454	0.459	0.481	0.449	0.519	0.455	—	—

3.4 Analyse

Le modèle XGBoost offre la meilleure performance en termes de **F1 Score** (0,653) et de **Rappel** (0,981), ce qui en fait un candidat particulièrement adapté dans un contexte où les faux négatifs sont coûteux. De plus, il affiche le **Profit Score** le plus élevé (1,975), ce qui indique un bon compromis entre signal et bruit pour des décisions financières.

En revanche, bien que la **Régression Logistique** présente la meilleure **AUC** (0,601), elle souffre d'un rappel très faible (0,245), ce qui la rend peu fiable pour détecter les mouvements haussiers. Le **modèle d'ensemble** par vote pondéré atteint une performance intermédiaire (Profit Score de 1,290), confirmant qu'une combinaison pondérée peut atténuer la variance des modèles individuels.

3.5 Conclusion

Malgré une AUC globalement modeste pour tous les modèles, les résultats suggèrent qu'il est possible de générer une stratégie profitable en exploitant un compromis entre rappel et précision, notamment via XGBoost. Le Profit Score s'impose ici comme une métrique clé d'évaluation, adaptée au contexte applicatif visé.

4 Stratégie de trading

Dans le monde financier d'aujourd'hui, l'utilisation de l'**apprentissage automatique** pour le trading est en train de devenir un enjeu majeur. Nous allons vous présenter une **stratégie de trading simplifiée** qui expose les principes de base de l'utilisation des prédictions pour prendre des décisions d'investissement. Bien qu'elle ne soit pas aussi complexe que ce que les grandes institutions utilisent, elle couvre tout de même les éléments essentiels d'un système de trading basé sur les données.

Notre stratégie se déroule en plusieurs étapes clés, chacune jouant un rôle dans la manière dont nous générons les signaux de trading et gérons nos investissements.

4.1 Choisir le Meilleur Modèle de Prédiction

Au cœur de tout cela se trouve un **modèle de classification binaire**. Imaginez-le comme un outil qui prédit la probabilité qu'un prix d'actif augmente dans le futur. Pour choisir le meilleur modèle, nous nous concentrons sur ce que l'on appelle l'**Aire Sous la Courbe (AUC)**. Nous privilégions l'AUC car elle nous offre une mesure très fiable de la capacité de notre modèle à distinguer entre les "prix qui montent" et les "prix qui ne montent pas", quel que soit le seuil spécifique que nous fixons. Une AUC élevée signifie que notre modèle est efficace, ce qui est crucial pour des prédictions fiables.

4.2 Trouver le Juste Milieu : Optimiser Notre Seuil de Confiance

Une fois que nous avons notre modèle, l'étape suivante, primordiale, consiste à déterminer le **seuil de confiance optimal** (P_{th}) pour l'achat. Ce seuil représente en fait la probabilité minimale que notre modèle doit prédire avant que nous n'envisagions même d'acheter. Nous trouvons ce "juste milieu" en analysant les données historiques. Notre objectif est de trouver le P_{th} qui nous donne les meilleurs rendements moyens sur les trades qui ont été profitables. Ainsi, nous n'agissons que sur les signaux les plus prometteurs, en nous concentrant sur la maximisation de nos gains potentiels.

4.3 La Stratégie de Trading Elle-Même : Simplifiée pour la Clarté

Notre stratégie de trading, que nous avons intégrée dans une classe nommée `SimplifiedTradingStrategy`, utilise les données de prix et les probabilités issues de notre modèle. Elle dispose de règles claires pour l'entrée et la sortie des positions, ainsi que d'une gestion de l'argent de base.

Voici ses paramètres principaux :

- **Seuil de confiance (P_{th})** : C'est le déclencheur d'achat, défini lors de notre étape d'optimisation.
- **Stop-Loss (SL)** : Un pourcentage de perte prédéfini qui clôture automatiquement un trade pour éviter des pertes plus importantes.
- **Take-Profit (TP)** : Un pourcentage de gain prédéfini qui clôture un trade pour sécuriser les bénéfices.
- **Coûts de Transaction** : Un petit pourcentage fixe appliqué à chaque achat et vente, tout comme les frais de trading réels.

Et comment elle fonctionne, jour après jour :

1. **Gestion des Positions Ouvertes** : Chaque jour, si nous avons un trade ouvert, la stratégie vérifie si les conditions de sortie sont remplies – comme atteindre notre stop-loss, atteindre notre take-profit, ou si la prédiction du modèle pour une hausse diminue significativement (nous utilisons 0.45 comme un simple signal de "vente"). Si l'une de ces conditions se produit, nous clôturons la position et enregistrons tous les détails.
2. **Entrée dans de Nouveaux Trades** : Si nous ne sommes pas déjà en position et que la probabilité prédite par notre modèle dépasse notre P_{th}, la stratégie envisage d'acheter. Pour simplifier, nous investirons un pourcentage fixe de 10% de nos liquidités disponibles pour ce trade.

4.4 Comment Savoir Si Ça Marche ? Vérifier la Performance

Pour voir comment notre stratégie se comporte, nous comparons ses rendements à ceux d'une simple approche **Acheter et Conserver (Buy & Hold)** sur la même période de test. Nous examinons quelques chiffres clés :

- **Rendement Total (%)** : C'est notre profit ou perte global(e) pendant la période d'analyse.
- **Rendement Annualisé (%)** : Cela nous indique le profit ou la perte moyen(ne) par an, facilitant la comparaison des performances sur différentes durées.
- **Ratio de Sharpe** : Il s'agit d'une mesure de rendement ajusté au risque. Elle nous montre combien de rendement nous obtenons pour chaque unité de risque que nous prenons. Un Ratio de Sharpe plus élevé est toujours préférable !

Ces métriques nous donnent une bonne vue d'ensemble de la rentabilité et de la stabilité de notre stratégie.

4.5 Visualiser la Performance

Pour rendre les choses encore plus claires, nous avons inclus un graphique. Il montre l'évolution de la valeur du portefeuille de notre stratégie au fil du

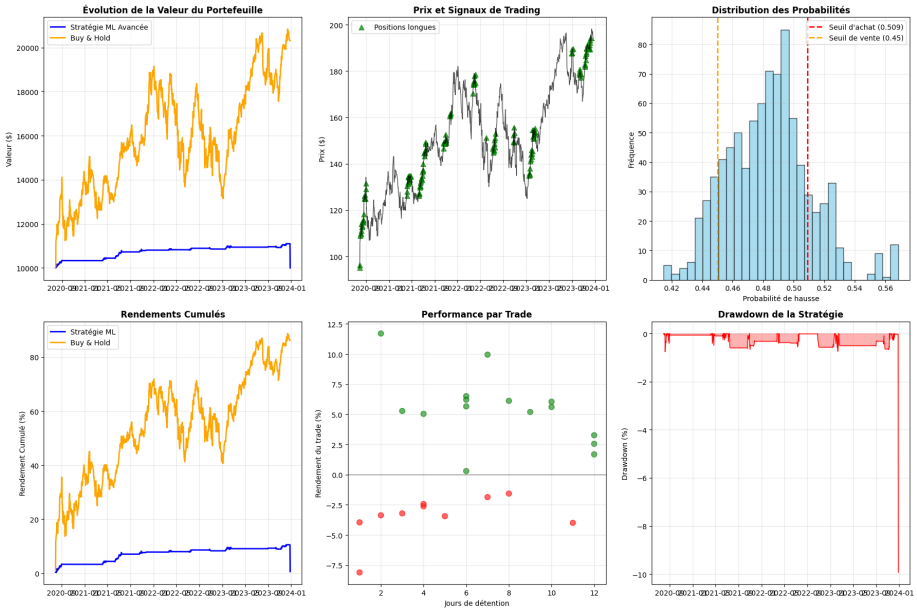

FIGURE 3 – Graphiques pour la visualisation des résultats

temps par rapport à une simple stratégie Buy and Hold. Cette comparaison visuelle nous aide à voir rapidement quand notre stratégie surperforme (ou sous-performe) et à comprendre sa trajectoire globale.

Cette stratégie de trading simplifiée offre un moyen direct de comprendre comment l'apprentissage automatique peut être utilisé en finance. Elle souligne l'importance d'avoir un bon modèle de prédiction, un seuil de décision optimisé et des règles disciplinées pour gérer vos trades. Bien qu'elle ne couvre pas toutes les complexités du trading réel, elle fournit une base solide pour des explorations plus avancées. On se rend bien compte que les résultats ne sont pas bons. La stratégie sous-performe par rapport à la stratégie Buy and Hold. Ces résultats prouvent la difficulté de battre le marché avec seulement des données financières. Avec l'utilisation du FNSPID, on peut s'attendre à des résultats probablement supérieurs à ceux présent ici.

5 Conclusion

Ce travail a permis de concevoir, implémenter et évaluer une stratégie de prédiction et de trading algorithmique fondée sur des techniques d'apprentissage automatique appliquées à l'action Apple Inc. (AAPL). À partir d'un ensemble de données enrichi combinant indicateurs techniques issus des données de marché et signaux de sentiment dérivés de Google Trends, un pipeline rigoureux de prétraitement a été mis en place afin de respecter la structure temporelle des données et de limiter les risques de surapprentissage.

L'expérimentation a porté sur plusieurs modèles de classification supervisée, avec une attention particulière portée à leur optimisation et à leur évaluation via des métriques standards, ainsi qu'une métrique spécifique — le Profit Score — mieux adaptée au cadre financier. Les résultats obtenus mettent en évidence la supériorité du modèle XGBoost en termes de capacité prédictive et de rentabilité simulée, même si la regression logistique montre une valeur d'AUC plus importante. Néanmoins, malgré certaines performances encourageantes, la stratégie élaborée reste en deçà de la performance d'une approche Buy and Hold sur la période de test.

Ces résultats soulignent la complexité intrinsèque de la prédiction des marchés financiers et la difficulté de construire une stratégie systématique réellement profitable à partir de données exclusivement quantitatives et techniques. Ils mettent également en lumière l'intérêt de compléter ces approches par des sources d'information plus riches, notamment des données textuelles issues de la presse spécialisée, intégrées via des modèles de traitement du langage naturel récents, comme avec le FNSPID.

Ce travail constitue ainsi une base solide pour des investigations futures, notamment en intégrant des architectures de type transformers ou en explorant des techniques d'apprentissage par renforcement sur des environnements de marché simulés.