Prédiction des Prix et Actifs du S&P 500 grâce à du Machine Learning simple

TD MID-TERM – 3A – DDEFi

Mardoché Clabessi | Ahmed Koffi | Elie-Emmanuel Motchoffo | Cédric Loussert | Sosthène Zongo

1 Table des matières

2 Introduction et objectifs

Les marchés financiers sont influencés par de nombreux facteurs économiques, techniques et comportementaux. Ce TD vise à prédire les prix et les rendements des actifs du S&P 500 à l'aide de techniques de machine learning simples.

Objectifs :

1. Collecter et préparer des données financières pertinentes.
2. Implémenter au moins 2 modèles prédictifs efficaces.
3. Évaluer les performances des modèles à l'aide de métriques appropriées.
4. Conclure sur les analyses retenues et proposer des ouvertures cohérentes.

3 Présentation des Variables

3.1 Variables explicatives

Les variables explicatives sont de 2 ordres : Variables macroéconomiques et Indicateurs techniques.

3.1.1 Variables macroéconomiques

Les variables macroéconomiques sont des indicateurs économiques qui mesurent la performance globale et la santé d'une économie, comme le PIB, l'inflation, le taux de chômage et les taux d'intérêt. Ces variables peuvent fournir des informations sur les conditions économiques générales susceptibles d'affecter la performance des actions individuelles ou du marché boursier dans son ensemble.

- **Indice de Volatilité (VIX) :** Indicateur de la peur sur les marchés, représentant la volatilité attendue sur 30 jours. Données issues de Yahoo Finance.
- **Taux d'intérêt (EFFR) :** Fixé par la Réserve fédérale, il influence les coûts d'emprunt et la croissance économique. Données issues de FRED.
- **Taux de Chômage (UNRATE) :** Indicateur de la santé économique, mesurant la part de chômeurs dans la population active. Données mensuelles issues de FRED.
- **Indice de Confiance des Consommateurs (UMCSENT) :** Mesure du sentiment des consommateurs envers l'économie. Données mensuelles issues de FRED.
- **Indice du Dollar Américain (USDX) :** Mesure de la valeur du dollar par rapport à un panier de devises. Données issues de Yahoo Finance.

3.1.1.1 Indice de volatilité (VIX)

L'indice de volatilité CBOE, également connu sous le nom de VIX, mesure la volatilité attendue de l'indice S&P 500 sur les 30 prochains jours. Il est souvent appelé l'« indice de la peur » ou le «

baromètre de la peur », car il est utilisé pour évaluer le niveau de peur ou de stress sur le marché boursier.

- Si le VIX est faible, cela indique que les investisseurs ne sont pas trop préoccupés par la volatilité à court terme du marché. Ils se sentent confiants et optimistes quant à l'avenir proche.
- À l'inverse, si le VIX est élevé, cela peut indiquer que les investisseurs sont plus incertains quant à l'avenir et sont plus susceptibles de vendre leurs actions, ce qui peut entraîner une baisse des prix des actions.

Nous avons téléchargé les données historiques **quotidiennes** depuis Yahoo Finance.

3.1.1.2 Taux d'intérêt (EFFR)

Le taux d'intérêt fait référence au pourcentage auquel l'emprunt ou le prêt d'argent est effectué. Les banques centrales, comme la Réserve fédérale des États-Unis, fixent les taux d'intérêt pour contrôler l'inflation et gérer la santé globale de l'économie.

- Lorsque les taux d'intérêt sont élevés, il devient plus coûteux pour les particuliers et les entreprises d'emprunter de l'argent, ce qui peut entraîner un ralentissement de la croissance économique.
- Lorsque les taux d'intérêt sont bas, il devient moins coûteux d'emprunter de l'argent, ce qui peut encourager la consommation et l'investissement, entraînant une croissance économique.

Le taux d'intérêt est une variable macroéconomique importante à inclure comme feature dans un modèle de prévision des prix des actions, car les variations des taux d'intérêt peuvent affecter la rentabilité des entreprises et la performance globale du marché boursier. Par exemple, si les taux d'intérêt sont bas, il peut être moins coûteux pour les entreprises d'emprunter de l'argent et d'investir dans la croissance, ce qui peut entraîner une hausse des prix des actions. De même, si les taux d'intérêt sont élevés, il peut devenir plus coûteux pour les entreprises d'emprunter de l'argent, ce qui peut entraîner une baisse des prix des actions.

Les données historiques **quotidiennes** ont été téléchargées depuis FRED.

3.1.1.3 Taux de chômage civil (UNRATE):

Le taux de chômage civil mesure le pourcentage d'individus dans la population active civile qui sont au chômage mais activement à la recherche d'un emploi et prêts à travailler. Il est calculé par le Bureau of Labor Statistics (BLS) et est publié mensuellement.
Un taux de chômage faible indique généralement que l'économie est forte, que les entreprises embauchent et se développent. En revanche, un taux de chômage élevé peut indiquer que l'économie est faible, que les entreprises suppriment des emplois et ralentissent leur activité.

Les données historiques mensuelles ont été téléchargées depuis FRED. Ces données sont **mensuelles**.

3.1.1.4 Indice de confiance des consommateurs (UMCSENT):

L'indice de confiance des consommateurs mesure la confiance des consommateurs, reflétant leur sentiment par rapport à l'état actuel et futur de l'économie.
L'indice de confiance des consommateurs est une mesure utile à inclure, car la confiance des consommateurs est étroitement liée à leurs dépenses, qui stimulent la croissance économique et sont un moteur clé des profits des entreprises. Lorsque les consommateurs sont confiants et optimistes pour l'avenir, ils sont plus enclins à dépenser de l'argent pour des biens et services. À l'inverse, lorsque les consommateurs sont pessimistes et incertains, ils sont moins enclins à dépenser, ce qui peut entraîner une baisse des ventes et des profits pour les entreprises ainsi qu'une chute des prix des actions.

Les données historiques mensuelles ont été téléchargées depuis FRED. Ces données sont **mensuelles** tout comme **UNRATE.**

Indice du dollar américain (USDX):

L'indice du dollar américain (USDX) mesure la valeur du dollar américain par rapport à un panier d'autres devises majeures.
Un dollar plus fort peut rendre les exportations plus chères, ce qui nuit aux entreprises exportatrices et peut entraîner une baisse des prix des actions. À l'inverse, un dollar plus faible peut rendre les exportations moins chères, rendant les entreprises plus compétitives sur le marché mondial, ce qui peut entraîner une hausse des prix des actions.

Les **données historiques quotidiennes** ont été téléchargées depuis Yahoo Finance.

3.1.2 Indicateurs techniques

Les indicateurs techniques sont des calculs mathématiques basés sur le prix et/ou le volume d'un titre. Ils sont utilisés pour identifier les modèles et les tendances du prix d'un titre, et peuvent être utilisés pour faire des prédictions sur les mouvements futurs des prix.

- **Moyennes Mobiles (SMA, EMA) :** Indicateurs de tendance basés sur les prix historiques.
- **RSI (Relative Strength Index) :** Indicateur de momentum pour détecter les zones de surachat ou de survente.
- **MACD (Moving Average Convergence Divergence) :** Mesure de la relation entre deux moyennes mobiles pour identifier les tendances.

3.1.2.1 Prix d'ouverture 'Open'

Reflète le sentiment du marché dès l'ouverture. Les gaps d'ouverture sont souvent exploités par les traders pour des signaux de retournement ou de continuation.

3.1.2.2 High et Low

High et **Low** : Capturent les niveaux extrêmes intrajournaliers, cruciaux pour détecter des niveaux de support/résistance et identifier des *breakouts*.

3.1.2.3 Volume

Indique l'intensité des échanges. Un volume élevé sur les niveaux **High** ou **Low** peut signaler des renversements ou confirmer des cassures significatives.

3.1.2.4 Moyenne Mobile

La moyenne mobile sur une fenêtre donnée (14 jours dans notre cas) calculée à la date t permet de calculer la moyenne sur la fenêtre des 14 premiers jours consécutifs avant la date t des cours d'un instrument financier. Elle permet de mettre en exergue les tendances dans les séries temporelles

3.1.2.5 Convergence et divergence des moyennes mobiles (MACD) :

Le MACD (Moving Average Convergence Divergence) est un indicateur technique populaire utilisé pour identifier les tendances et la dynamique du prix d'un titre. Il est calculé en soustrayant la moyenne mobile exponentielle (EMA) sur 26 périodes du prix d'un titre de l'EMA sur 12 périodes du même prix. Cela élimine tous les signaux de trading à court terme, produisant ainsi une ligne de signal plus stable à moyen terme.

Le MACD est souvent utilisé en combinaison avec une ligne de signal, qui est l'EMA sur 9 périodes du MACD. Le croisement de la ligne MACD rapide avec la ligne de signal plus lente indique un changement de dynamique.

3.1.2.6 Indice de force relative (RSI) :

Le RSI est un indicateur technique qui aide les traders à comprendre si une action ou un autre instrument financier est suracheté ou survendu. Il compare l'amplitude des gains récents aux pertes récentes pour déterminer les conditions de surachat et de survendu d'un actif.

Une valeur supérieure à 70 sur le RSI est considérée comme surachetée, ce qui signifie que le prix de l'action est élevé et pourrait bientôt baisser. En revanche, une valeur inférieure à 30 sur le RSI est considérée comme survendue, ce qui signifie que le prix de l'action est bas et pourrait augmenter bientôt.

3.2 Variable cible

Dans cette étude, nous examinons l'historique des prix du **S&P 500**, un indice boursier qui suit l'évolution des actions de 500 des plus grandes entreprises cotées en bourse aux États-Unis. Notre objectif est de prédire le prix futur de cet indice en privilégiant principalement des modèles de **Machine Learning de régression**, plutôt que les approches classiques de **modélisation des séries temporelles.**

En effet, les méthodes dites traditionnelles supposent que la valeur future de l'indice peut être prédite uniquement à partir de ses valeurs passées, sans tenir compte de variables exogènes.
Tandis que l'approche machine learning permet offre une flexibilité quant au choix des variables.

Pour ce faire, on étudie un **rendement logarithmique** (à partir des prix journaliers). Cette transformation réduit l'impact des extrêmes et rend les données plus stationnaires.

4 Préparation des données

Les données ont été collectées à l'aide d'API vers des sources fiables : Yahoo Finance et FRED.

Notre série temporelle principale (S&P 500) comporte de nombreux jours manquants en raison des week-ends et des jours fériés aux États-Unis. Nous y remédierons en traitant les données de la série temporelle comme un signal continu, en supposant que les week-ends et les jours fériés ne prennent pas de temps.

4.1 Gestion des données monthly (UNRATE, UMCSENT)

En ce qui concerne les données mensuelles, nous devons les convertir en données journalières afin que chaque date pour laquelle le S&P 500 dispose de données soit également associée à des valeurs macroéconomiques. Nous avons utilisé l'approche **Forward Fill** en trois étapes :

- **Temps 1** : on crée un dataframe indicé sur dates en continue dont la valeur un 1er du mois correspond à celle du UNRATE/ UMCSENT correspondant.
- **Temps 2** : On utilise l'approche "Forward Fill" pour renseigner les valeurs manquantes. L'idée est de parcourir le dataframe dans l'odre chronologique, puis à une valeur manquante rencontrée, on lui assigne de la dernière valeur disponible.
- **Temps 3** : on fait une concaténation avec la série de SP500 en supprimant les valeurs manquantes.

4.2 Gestion de valeurs manquantes et abérrantes

Si le S&P 500 est coté à une date donnée, mais qu'une de nos autres variables ne l'est pas, nous pouvons imputer la valeur manquante en utilisant la dernière donnée disponible dans l'ordre chronologique.

En ce qui concerne les valeurs aberrantes mises en évidence par les boxplots, celles-ci sont principalement attribuables à la période de la pandémie de COVID-19. Cette période a entraîné une chute importante du cours du S&P 500, ce qui explique la présence de nombreuses valeurs nettement inférieures aux seuils inférieurs des boxplots.

À l'inverse, les valeurs aberrantes majoritairement positives observées pour l'indice VIX trouvent leur origine dans le comportement des investisseurs en période de stress. Pendant ces phases, les investisseurs achètent massivement des options pour se protéger, ce qui augmente leur prix et, par conséquent, la volatilité du marché. Cette relation est cohérente avec la corrélation négative entre le VIX et le S&P 500.

Enfin, les valeurs exceptionnellement élevées de l'indicateur UNRATE (taux de chômage) s'expliquent par l'arrêt brutal des activités économiques pendant la pandémie. Face à cette crise, de nombreuses entreprises ont dû procéder à des restructurations importantes pour survivre, ce qui a considérablement augmenté le chômage.

5 Modélisation et interprétation

5.1 Modèles traditionnelles vs modèles Machine learning

Critères	**Approche traditionnelle**	**Approche Machine Learning**
Points positifs	1. Simple à implémenter et à interpréter	Prédisent des relations complexent entre variables
	Requiert peu de données exogènes	Intègre des données externes variées pour améliorer la précision
	Basé sur des principes statistiques solides	Adapté aux problèmes non linéaires et dynamiques
Points Négatifs	Pas très efficace si relation non linéaires	Nécessite une préparation approfondie des données (feature engeneering)
	Ignore la plupart du temps les variables exogènes	Peut être plus difficile à interpréter
	Moins adapté aux évolutions rapides du marché	Demande d'avantage de ressource pour le calcul

Nous avons opté pour la modélisation pour une approche machine learning compte tenu de la flexibilité quant au choix des variables explicatives que celle-ci offrait même si plus complexe à interpréter.

5.2 Modèles Machine learning implémentés

5.2.1 Motivation de Modélisation et Présentation des métriques

Les modèles que nous avons décidé d'implémenter sont : **LinearRegression** et **RandomForestRegressor**. Le premier étant assez efficace pour capturer les dépendances linéaires et le second pour les dépendances non linéaires.

Notre point de départ est un univers de 16 features. Pour rappel chacune de ces features est affectée d'un shift (décalage dans le temps). Autrement dit on souhaite utiliser les informations passées pour prédire le futur. Ainsi de toutes nos features, la plus récentes est le prix d'ouverture du SP500. Le prix d'ouverture fait donc partie des variables que l'on utilise pour prédire le prix au close. Toutefois si on souhaite prédire à un horizon plus éloignée, on peut faire une récursivité afin de projeter dans le temps les variables explicatives pour faire notre prédiction (Un point qui sera détaillé en ouverture).

Pour chacun des modèles, nous allons faire une recherche de features pertinentes afin de réduire l'univers des features ce qui aura pour effet d'améliorer la performance des modèles et accélérer le temps d'entrainement.

Notre travail peut servir à diverses fins, selon le but de l'utilisateur final, et peut être appréhendé soit comme un problème de classification, soit comme un problème de régression. Dans le cadre du **problème de régression**, les métriques **RMSE** et **MAE** sont utilisées pour évaluer la précision des prédictions en termes de valeurs continues, en mesurant respectivement l'erreur quadratique moyenne et l'erreur absolue moyenne. Le **R2**, quant à lui, permet d'apprécier la corrélation entre les prédictions et les valeurs réelles, offrant un avantage unique par rapport aux autres métriques : il ne dépend pas de l'ordre de grandeur de la variable prédite.

Par exemple, un **MAE** de 10 peut être excellent pour des données de l'ordre du millier, mais moins pertinent pour des données de l'ordre de la dizaine, tandis que **R2** reste une mesure cohérente quel que soit l'échelle des données. Dans le cadre d'un **problème de classification binaire**, nous mesurons la capacité du modèle à prédire correctement la direction des mouvements du marché, en utilisant la métrique **accuracy** pour évaluer la proportion de prédictions correctes concernant la tendance du S&P 500, qu'il s'agisse de prévisions de hausse ou de baisse.

On utilise la validation croisée temporelle avec `TimeSeriesSplit`, ce qui est adapté aux séries temporelles. La validation croisée est effectuée avec 10 splits et, pour chaque split, le modèle de régression linéaire est ajusté et évalué sur les ensembles d'entraînement et de test.

5.2.2 LinearRegression : Entrainement et analyse des résultats

Pour extraire les variables explicatives les plus importantes dans le cadre du modèle de régression linéaire, nous avons standardisé les features afin de rendre comparables les coefficients du modèle après entraînement. Cette approche a permis de sélectionner huit variables essentielles pour le modèle final :

- Le prix à l'ouverture du jour de la prédiction.
- Le prix bas ("low") et le prix haut ("high") de la veille.
- Le volume de transactions de la veille.
- L'indice VIX deux jours avant.
- Le prix à l'ouverture de la veille.
- Le prix de clôture deux jours avant.
- Le prix de clôture de la veille.

La suppression des variables jugées superflues a conduit à une amélioration notable des performances du modèle, avec une réduction de 4 % de la MAE (Mean Absolute Error) et de 3 % du RMSE (Root Mean Squared Error).

Cependant, il est intéressant de noter l'absence des variables macroéconomiques parmi les facteurs sélectionnés. Cela peut s'expliquer par leur influence à long terme sur le marché (plusieurs mois), impactant davantage les tendances globales que les fluctuations à court terme du S&P 500.

Le bilan du modèle est globalement satisfaisant. Avec une MAE de 28 et un RMSE de 36, les erreurs de prédiction restent relativement faibles par rapport à l'ordre de grandeur du S&P 500. De plus, l'accuracy de 100 % indique que le modèle prédit parfaitement la direction du marché une fois le prix à l'ouverture connu. Enfin, un coefficient R^2 de 0,94 démontre que le modèle explique 94 % de la variance des données, attestant d'une excellente capacité prédictive.

L'analyse des résidus révèle une dispersion globalement aléatoire autour de la ligne centrale, indiquant que le modèle capture bien les relations linéaires. La variance des résidus semble constante, confirmant une homoscédasticité, bien que quelques valeurs extrêmes (supérieures à 100 ou inférieures à -100) soient présentes. Globalement, les résidus valident la pertinence du modèle, tout en mettant en lumière des observations exceptionnelles méritant une attention particulière.

5.2.3 RandomForest Regressor : Entrainement et analyse des résultats

Pour la construction de ce modèle, un algorithme de sélection des variables explicatives (feature selection) a été suivi d'un algorithme d'optimisation des hyperparamètres. Les variables retenues sont :

- Le prix à l'ouverture.
- Les prix haut ("high") et bas ("low") de la veille.
- Les prix d'ouverture et de clôture de la veille.
- L'indice VIX.
- La moyenne mobile.
- Le MACD (Moving Average Convergence Divergence).
- Le prix de clôture deux jours avant.

Il convient de noter, une fois de plus, l'absence de variables macroéconomiques parmi les facteurs sélectionnés.

Les performances du modèle Random Forest Regressor sont décevantes. Les métriques montrent des résultats médiocres, avec une MAE (Mean Absolute Error) de 235 et un RMSE (Root Mean Squared Error) de 270, ce qui représente des erreurs importantes par rapport à l'ordre de grandeur du S&P 500 (~5000-6000). Le coefficient R2R^2R2, avec une valeur de -0,94, indique que le modèle échoue à expliquer la variance des données, se révélant même moins performant qu'une simple moyenne. De plus, l'accuracy directionnelle, à 51 %, est proche du hasard.

Les prédictions quasi constantes ne capturent pas les variations significatives du S&P 500, comme en témoigne le graphe comparant les valeurs réelles et prédites. Par ailleurs, la courbe d'apprentissage révèle un surajustement extrême : bien que le RMSE en entraînement soit faible, les erreurs sur les données de test sont considérables, indiquant une incapacité du modèle à généraliser correctement.

Ces résultats suggèrent des limites importantes dans l'approche utilisée, nécessitant une révision des choix de variables, des hyperparamètres et potentiellement de l'algorithme employé.

5.2.4 Cas d'utilisation de notre modèle

Un exemple concret d'application de notre modèle se situe dans le domaine du trading haute fréquence (High Frequency Trading - HFT). Dans ce secteur, les intervenants prennent des positions qu'ils conservent sur des périodes très courtes, souvent inférieures à une journée, parfois réduites à quelques heures ou même quelques minutes.

Notre algorithme offre deux avantages clés pour ce type de stratégie :

1. **Prédiction directionnelle fiable** : Il est capable de prédire avec précision le mouvement de hausse ou de baisse à la clôture du marché, dès que le prix d'ouverture est connu au début de la journée.
2. **Définition de seuils de take-profit** : Le modèle peut également aider à établir des seuils de prise de profit (take-profit), c'est-à-dire des niveaux de gain à partir desquels la position se clôture automatiquement, optimisant ainsi les sorties de position.

Grâce à une accuracy de 100 %, cette approche garantirait un bilan sans aucune perte sur les trades réalisés. Une telle performance pourrait transformer les opérations en HFT, en maximisant la rentabilité tout en minimisant les risques.

6 Conclusion

Ce projet visait à prédire les prix de clôture des actions en utilisant des modèles de machine learning conçus pour analyser les séries temporelles financières. Après une analyse approfondie des données historiques, un des modèles déployés a démontré des performances remarquables, avec une accuracy de 100 %. Ce résultat exceptionnel est corroboré par un coefficient de détermination R^2 de 94 %, confirmant que les prédictions suivent étroitement les tendances des valeurs réelles. Cependant, le modèle ne capture pas parfaitement les amplitudes des variations, ce qui explique la présence d'une erreur absolue moyenne (MAE) non nulle.

Malgré ces résultats prometteurs, notamment pour des applications comme le trading automatisé, certaines limites subsistent. Le modèle reste sensible aux événements imprévus, tels que les crises économiques ou les annonces inattendues, et sa fiabilité repose fortement sur la qualité et l'historique des données utilisées. Ces contraintes soulignent la nécessité de compléter cette approche par des modèles capables d'intégrer des variables exogènes et de s'adapter à des environnements dynamiques.

Ce projet constitue une base solide pour des développements futurs, notamment l'exploration de modèles plus avancés comme les réseaux neuronaux récurrents (RNN), les Transformers ou les modèles hybrides combinant des approches statistiques et machine learning. Une validation rigoureuse dans des conditions de marché réelles et la prise en compte de facteurs macroéconomiques pourraient également améliorer les performances globales et renforcer la robustesse des prédictions.