

Etude de séries temporelles Financières

DESCHAUX Pierre-Honoré
NARDI Léo

Introduction

La prédiction des prix des actifs boursiers est un élément crucial pour les investisseurs et les gestionnaires de portefeuille. En effet, une prévision précise des prix peut aider à prendre des décisions d'investissement éclairées et à maximiser les rendements.

Il existe de nombreuses méthodes pour prédire les prix des actifs boursiers. Certaines méthodes se basent sur l'analyse technique, qui utilise des indicateurs/ Oscillateurs calculés à partir des données historiques des cours pour prédire les mouvements futurs des prix. D'autres méthodes se basent sur l'analyse fondamentale, elles utilisent des informations sur l'économie et l'entreprise pour prédire les mouvements futurs des prix.

Dans le cadre de ce projet nous nous pencherons sur une approche quantitative avec une étude économétrique du SP500 du 01/01/2013 à 01/01/2023 et la mise en place d'un modèle de deep learning.

Importance de la stationnarité

La stationnarisation d'une série chronologique est une étape importante dans l'analyse économétrique. Une série stationnaire est une série dont les propriétés statistiques (moyenne, variance, autocorrélation) ne changent pas au fil du temps. Cela signifie que les valeurs futures de la série peuvent être prédites en utilisant les valeurs passées.

Lorsque nous travaillons avec des séries non stationnaires, il peut être difficile de prédire les valeurs futures car les propriétés statistiques de la série peuvent changer au fil du temps. En stationnalisant la série, nous pouvons la rendre plus facile à analyser et à prédire.

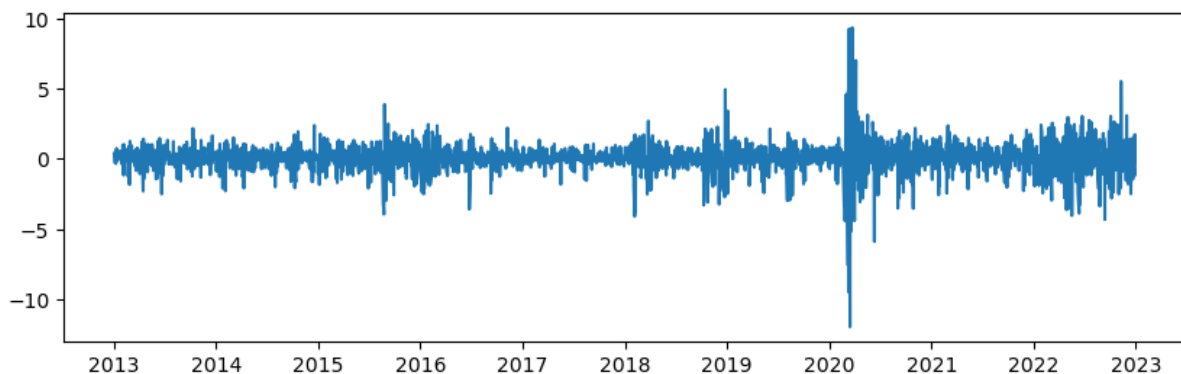
Le test augmenté de Dickey-Fuller (ADF) est un test statistique couramment utilisé pour tester la stationnarité d'une série chronologique. Ce test utilise une régression pour estimer les paramètres d'un modèle autorégressif et teste l'hypothèse nulle de présence de racine unitaire.

Le test ADF peut être utilisé pour déterminer si une série chronologique doit être différenciée pour la rendre stationnaire. Si le test rejette l'hypothèse nulle, cela suggère que la série est stationnaire et ne nécessite pas de différenciation. Si le test ne rejette pas l'hypothèse nulle, cela suggère que la série est non stationnaire et doit être différenciée pour la rendre stationnaire.

En appliquant ce test sur le SP500 s'aperçoit que la série possède une racine unitaire et donc n'est pas stationnaire. Nous allons donc la différencier.

La différenciation est une méthode couramment utilisée pour stationnariser une série chronologique. Elle consiste à calculer la différence entre les valeurs consécutives de la

série. En d'autres termes, nous remplaçons chaque valeur de la série par la différence entre cette valeur et la valeur précédente.



Après la différenciation de notre série, le test ADF nous confirme sa stationnarité.

Modèle ARIMA

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est une méthode statistique utilisée pour modéliser des séries chronologiques. Il a été développé par les statisticiens et économètres Box et Jenkins dans les années 1970. Il est utilisé entre autres dans les domaines de la finance, de l'économie, de la météorologie et de la gestion de la chaîne d'approvisionnement.

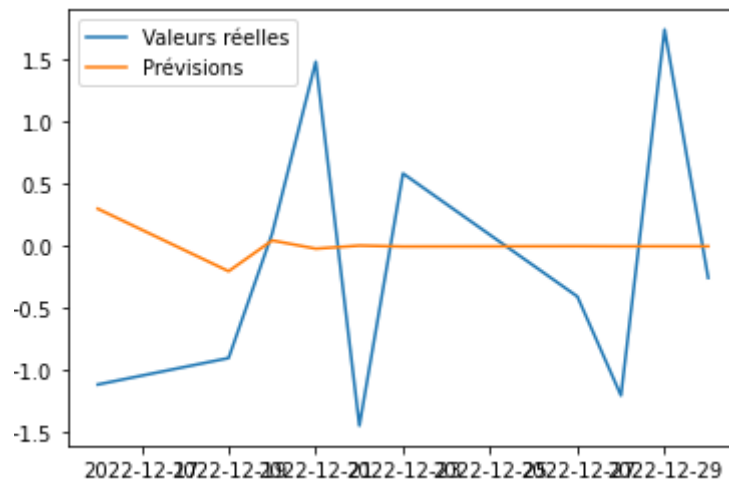
L'ARIMA est une extension du modèle ARMA (AutoRegressive Moving Average) qui est un modèle linéaire pour les séries chronologiques stationnaires. La différence entre ARMA et ARIMA est que ARIMA peut également prendre en compte les séries chronologiques non stationnaires. ARIMA modélise la relation entre les observations successives en utilisant trois paramètres: p , d et q .

- p est le nombre de termes autorégressifs, c'est-à-dire le nombre de valeurs précédentes de la série chronologique utilisées pour prédire la valeur actuelle.
- d est le nombre de différences non saisonnières nécessaires pour rendre la série chronologique stationnaire. Les différences peuvent être prises en utilisant la méthode de différenciation.
- q est le nombre de termes de moyenne mobile, qui indique le nombre d'erreurs de prédiction précédentes que l'on utilise pour prédire la valeur actuelle.

ARIMA est capable de capturer les tendances et les structures saisonnières dans les séries chronologiques, ce qui le rend utile pour la prévision. Les données sont souvent divisées en un ensemble d'entraînement et un ensemble de tests, et le modèle ARIMA est ajusté sur l'ensemble d'entraînement (train sample), puis utilisé pour prédire les valeurs futures sur l'ensemble de test.

Afin de choisir le meilleur modèle, nous allons chercher à minimiser le critère d'information de plusieurs modèles. Nous retenons un ARIMA(2,0,0) soit un AR(2).

s'ensuit alors la prévision :



La prédiction n'est pas très satisfaisante de par son rapide retour à 0 et sa non similitude avec les valeurs réelles.

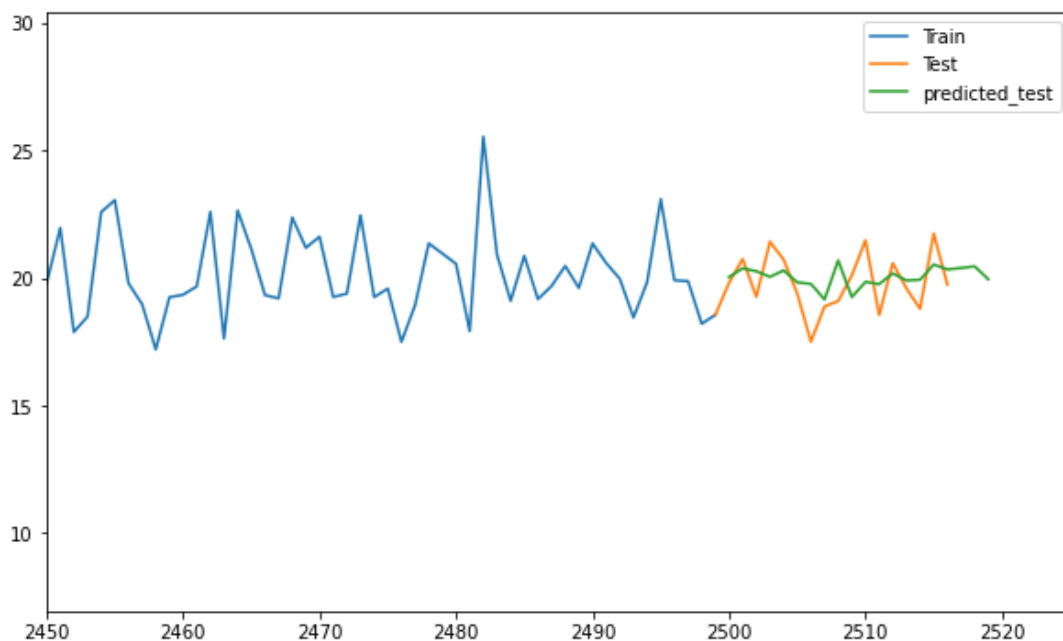
Holt Winters

Le modèle de Holt-Winters, également connu sous le nom de lissage exponentiel triple, est un modèle de prévision pour les séries chronologiques qui prend en compte la tendance et la saisonnalité. Il a été développé par Charles Holt et Peter Winters dans les années 1960 et est largement utilisé dans les applications de prévision.

Le modèle de Holt-Winters est basé sur trois composantes : la tendance, la saisonnalité et la composante résiduelle. La tendance est la direction générale dans laquelle évolue la série chronologique. La saisonnalité représente les fluctuations périodiques dans les données qui se répètent à intervalles réguliers. La composante résiduelle représente l'erreur de prévision qui ne peut être expliquée par la tendance et la saisonnalité.

Le modèle de Holt-Winters utilise des équations pour modéliser la tendance et la saisonnalité, qui sont mises à jour à mesure que de nouvelles données sont disponibles. Le modèle est capable de faire des prévisions à court et à long terme en utilisant les tendances et les saisons identifiées dans les données historiques.

Il existe deux versions principales du modèle de Holt-Winters : le modèle additif et le modèle multiplicatif. Le modèle additif est utilisé lorsque les fluctuations saisonnières ont une amplitude constante dans la série chronologique. Le modèle multiplicatif est utilisé lorsque les fluctuations saisonnières ont une amplitude qui varie en fonction de la tendance. Le choix du modèle pour notre série s'est porté sur le multiplicatif en raison de la présence de cluster de volatilité.



Les résultats de ce modèle sont beaucoup plus convaincants mais restent quand même assez éloignés de la réalité.

Ces erreurs de prévision pour le modèle ARIMA et Holt Winters peuvent notamment être expliquées par le fait que les cours boursiers étant très souvent volatils (avec des clusters de volatilité) et influencés par de nombreux facteurs économiques, politiques et géopolitiques, ne peuvent être pris en compte par un modèle linéaire.

De plus, les modèles ARIMA (ou ARMA) ne prennent en compte que les moments d'ordre 2 au travers de la fonction d'autocovariance, ce qui implique une exploitation incomplète de l'information contenue dans la série si des régularités apparaissent dans les moments d'ordres supérieurs.

C'est pourquoi il est plus intéressant d'introduire des modèles non linéaires (type ARCH, GARCH) car ces derniers prennent en considération les moments d'ordre supérieur à 2 et permettent donc d'affiner l'étude de la dynamique de la série.

En conséquence, il est difficile de faire des prévisions précises sur des cours boursiers en utilisant uniquement les modèles linéaires.

GARCH

Le modèle GARCH (Generalized Autoregressive Conditional Heteroskedasticity) est un modèle économétrique qui est couramment utilisé pour modéliser la volatilité dans les séries chronologiques financières. Il a été introduit dans les années 1980 par les économètres Bollerslev et Engel.

Le modèle GARCH est une extension du modèle ARCH (Autoregressive Conditional Heteroskedasticity) qui prend en compte la dépendance de l'hétéroscédasticité conditionnelle des erreurs de prédiction sur les erreurs passées. Le modèle GARCH utilise une fonction d'autocorrélation pour modéliser l'hétéroscédasticité conditionnelle dans les séries financières.

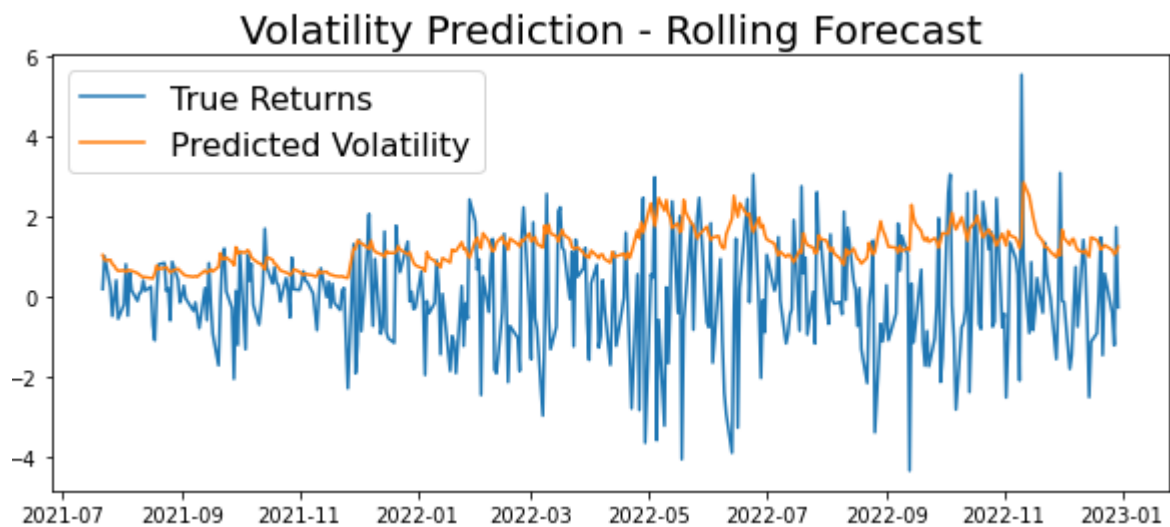
Il est souvent utilisé pour modéliser les séries chronologiques financières qui présentent des pics de volatilité périodiques ou des événements de choc ponctuels qui affectent la volatilité.

Ce modèle est cependant sujet à certaines limites, notamment la possibilité de surajustement (overfitting) et la difficulté à modéliser les chocs extrêmes ou les événements imprévus.

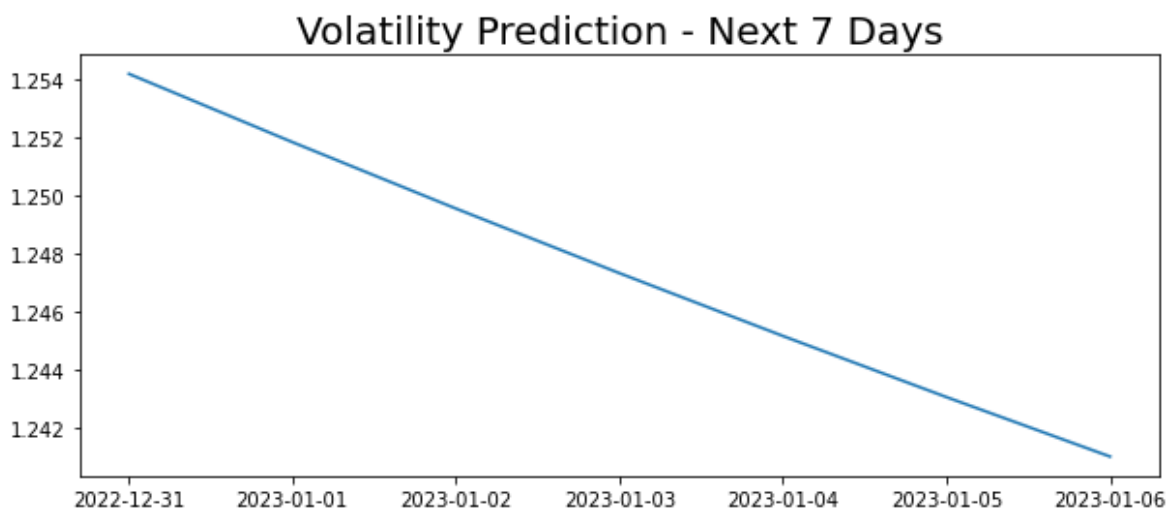
Nous allons faire un modèle GARCH sur les rendement du SP500.

Le modèle retenu, minimisant le critère d'AIC est le GARCH(1,1).

La prévision roulante du modèle nous montre une prédiction de la volatilité assez convaincante



et ce qui donne comme prévision a 1 semaine



VAR

Le modèle VAR (Vector AutoRegressif) est un modèle économétrique multivarié utilisé pour analyser les relations entre plusieurs variables en utilisant des données temporelles. Le modèle VAR est basé sur une série de régressions linéaires où chaque variable est expliquée par ses propres valeurs passées et les valeurs passées des autres variables du modèle.

C'est une généralisation des processus AR au cas multivarié (où la variable à expliquer Y devient un vecteur de variables de taille $1 \times N$). Le premier modèle VAR a vu le jour dans les années 1980 développé par Christopher Sims.

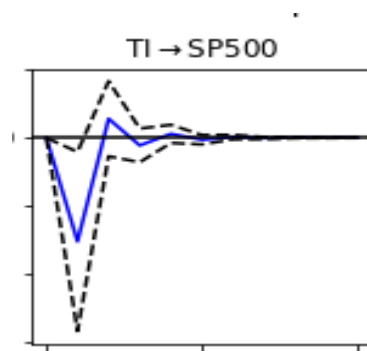
Il permet de modéliser les relations entre les variables de manière plus précise que dans une simple régression linéaire univariée. De plus, le modèle VAR peut être utilisé pour modéliser des relations de causalité entre les variables, ce qui peut être utile pour prédire les effets d'une variable sur une autre.

Nous allons utiliser ce modèle pour essayer d'identifier une relation entre le SP500 et indicateurs économiques afin de faire une prédiction.

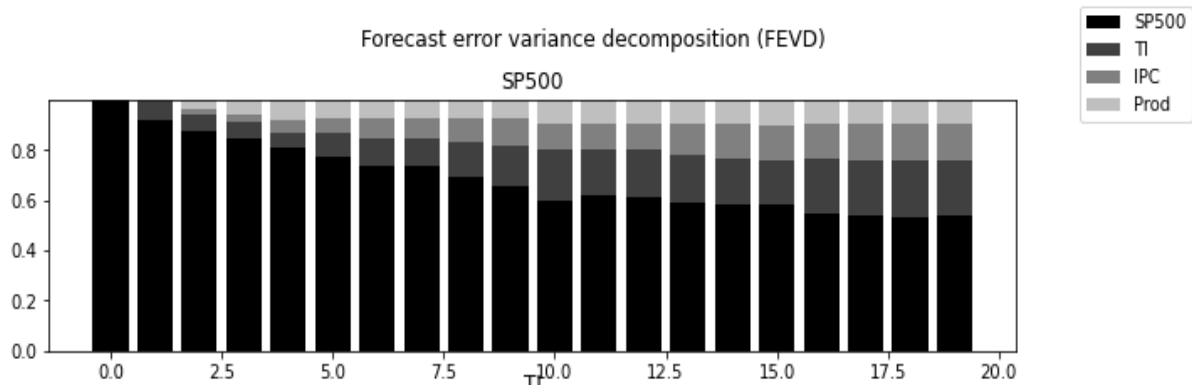
Nous avons choisi d'étudier un modèle VAR quadri-varié avec les différences premières (par soucis de stationnarité) du SP500, de l'inflation, des taux d'intérêt et de la production industrielle.

Lors de notre étude nous avons pu mettre en évidence la causalité du taux d'intérêt et de la production industrielle sur le SP500.

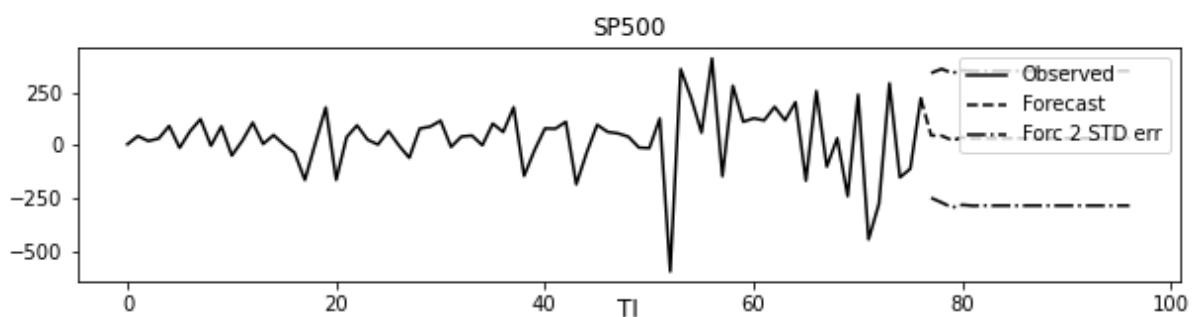
De plus il est intéressant de noter qu'un choc de Taux d'intérêt a un impact négatif sur le SP500



Nous observons que la variance des erreurs de prévisions des rendements du SP500 est expliquée à hauteur de 20% par le choc du taux d'intérêt à partir de l'horizon 10. En d'autres termes : un choc à l'instant présent sur le taux d'intérêt fait varier la prévision de la série de la croissance du PIB de 20%. L'inflation et la production industrielle expliquent aussi environ 10% chacun la variance des erreurs de prévision.



Les résultats de la prévision du SP500 avec un retard de 1 nous montre qu'il n'est pas possible de prédire de manière pertinente le SP500 avec ce modèle au vu du fort retour vers 0 et de l'intervalle de confiance plutôt étendu.



Le choix d'autres variables et de retards plus étendus en fonction du nombre d'observations peuvent être une piste d'amélioration du modèle.

MACHINE LEARNING

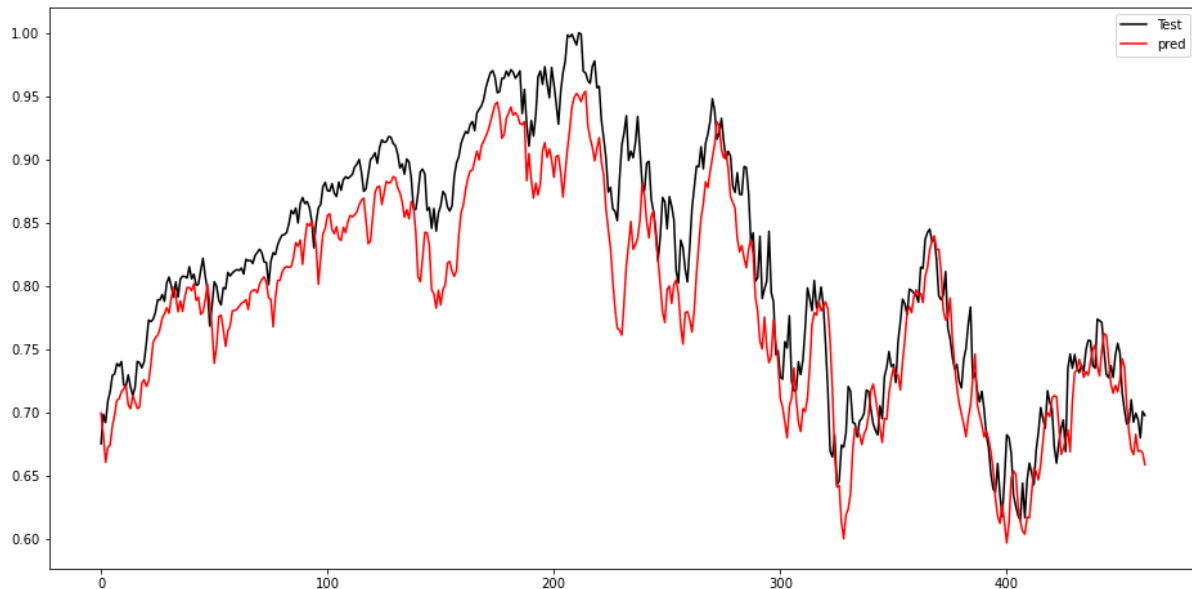
L'utilisation de l'apprentissage automatique et de l'apprentissage profond pour la prévision de séries temporelles financières a suscité un intérêt croissant ces dernières années. Ces techniques permettent de modéliser des relations complexes et non linéaires entre les variables et peuvent capturer des tendances et des motifs dans les données qui peuvent être difficiles à détecter avec des méthodes traditionnelles. Les approches de machine learning, terme introduit par Arthur Samuel en 1959, telles que les machines à vecteurs de support (SVM), les forêts aléatoires (RF) et les arbres de décision à boosting de gradient (GBTD) ont été utilisées pour prédire les séries temporelles financières. Ces méthodes peuvent être entraînées sur des données historiques pour apprendre à prédire les tendances futures, on peut citer Arthur Samuel et Tom M. Mitchell.

Les stratégies de deep learning, introduites par Geoffrey Hinton en 2006, telles que les réseaux neuronaux récurrents (RNN) et les réseaux de mémoire à court terme longue (LSTM) sont également utilisées pour la prévision de séries temporelles financières. Ces techniques peuvent capturer des dépendances temporelles à long terme dans les données et ont montré une amélioration de la précision des prévisions par rapport aux méthodes traditionnelles.

Il y a eu de nombreux développements dans ces domaines depuis lors, avec des contributions de nombreux chercheurs et praticiens. Par exemple, Yann LeCun, Yoshua Bengio et Geoffrey Hinton ont reçu le prix Turing en 2018 pour leurs contributions au deep learning.

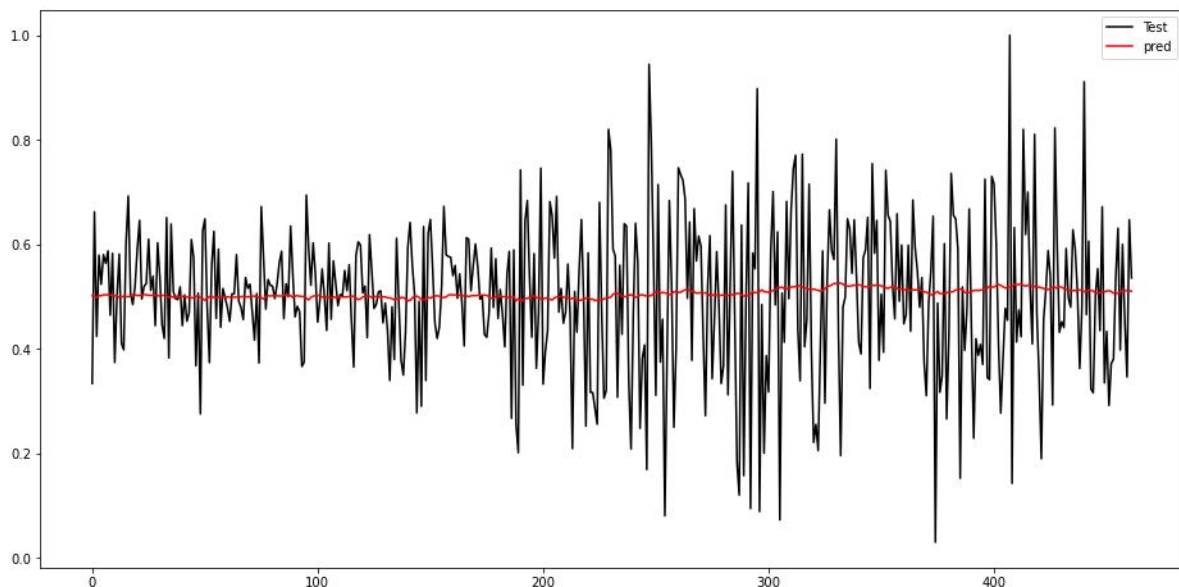
Dans le cadre de notre projet nous avons utilisé le LSTM pour tenter de faire des prévisions sur TargetNextClose, la clôture du jour suivant, avec le cours d'ouverture, plus haut et plus bas et des indicateurs techniques.

Nous obtenons ceci comme résultat :



A première vue, les résultats semblent assez bons. Cependant si l'on regarde de plus près on remarque que certe la prédiction a la même allure que le test mais qu'elle est décalée de 1 période. Cela signifie notamment que pour le modèle la clôture suivante ne sera pas très éloignée de la clôture présente.

Pour mieux visualiser cela nous avons modifié le modele pour qu'il fasse une prédiction sur Target qui correspond au volume du mouvement entre les deux cloture et nous obtenons ceci :



On voit alors qu'ici notre modèle capte mal les mouvements et ne se contente que de sortir une prévision assez proche de la valeur présente. Notre modèle est loin d'être parfait. Pour l'améliorer, on pourrait choisir d'autres indicateurs et données, renforcer le réseau de neurones et optimiser les hyper-paramètres.

Cette tentative de prédiction a surtout pour but de montrer à quel point il est difficile, si ce n'est impossible, et coûteux de prédire des cours boursiers. Malgré les difficultés de cette discipline, des institutions financières arrivent à générer du profit avec des approches quantitatives. On peut citer notamment le Medallion Fund. C'est un fond d'investissement géré par Renaissance Technologies, un hedge fund quantitatif créé par Jim Simons en 1982. Il est connu pour être le fonds le plus performant de l'histoire avec des rendements annuels exceptionnels depuis sa création en 1988. En 2021, il a enregistré un rendement net de 48% et en 2022, il a progressé de 19%. Depuis 2005, ce fonds est réservé uniquement aux collaborateurs du hedge fund. Renaissance Technologies utilise une approche quantitative pour ses stratégies d'investissement. Cela signifie qu'ils utilisent l'apprentissage automatique et l'intelligence artificielle, pour analyser et négocier des instruments financiers. Ils ont une base de données qui croît de plus de 40 téraoctets par jour et 50 000 cœurs d'ordinateur avec une connectivité globale de 150 gigabits par seconde.

Conclusion

Nous avons pu voir qu'il n'existait pas de modèle miracle pour prédire le cours d'un actif. En effet, la prédiction des cours boursiers est un problème complexe et difficile à résoudre. Il existe plusieurs approches et méthodes pour modéliser ces prix futurs, mais chacune dispose de ses avantages et de ses limites.

Les modèles ARIMA sont plus appropriés pour modéliser les tendances et les cycles des cours boursiers., tandis que, les modèles VAR sont particulièrement utiles pour modéliser les relations de corrélation entre plusieurs variables financières
Les modèles GARCH quant à eux, sont utilisés pour modéliser et prédire les variations de la volatilité des cours, et non leur prix.

Pour conclure, nous pouvons dire que les variations des prix des actifs financiers sont proches d'une marche aléatoire, en raison notamment de l'impact de facteurs externes tels que les événements politiques, les fluctuations économiques, les tendances du marché, etc. Tous ces éléments rendent donc leur prévision très difficile et souvent erronée.

Bibliographies et Sitographie

Retail time series forecasting using an automated deep meta-learning framework

Mohammad Joshaghani , Sasan Barak , Amirabbas Asadi , Ehsan Mirafzali

https://papers.ssm.com/sol3/papers.cfm?abstract_id=4393300

Projet Machine Learning pour la Prévision: séries temporelles

Yannig Goude

<https://www.imo.universite-paris->

[saclay.fr/~yannig.goude/Materials/ProjetMLF/time_series.html](https://www.imo.universite-paris-saclay.fr/~yannig.goude/Materials/ProjetMLF/time_series.html)

Recurrent Neural Networks | LSTM Price Movement Predictions For Trading Algorithms

Code Trading

https://www.youtube.com/watch?v=hpfQE0bTeA4&t=2s&ab_channel=CodeTrading