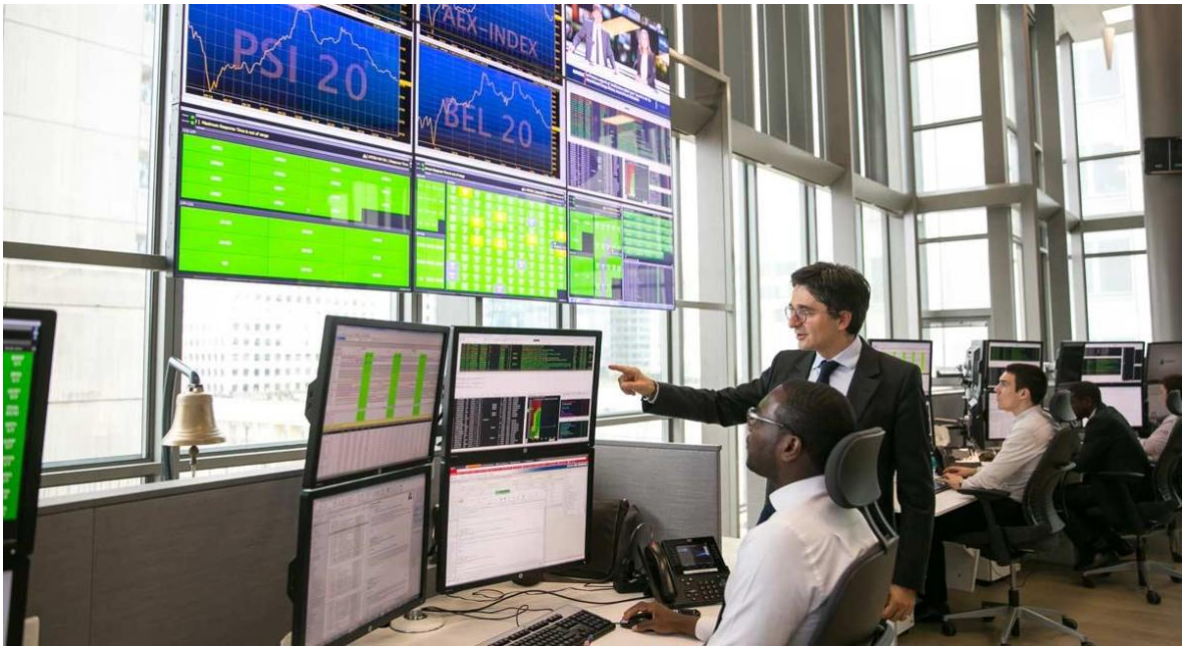
ATTILA Guillaume
ESILV IF2

Projet Machine Learning :
Prédiction des tendances de marchés boursiers



Sommaire

I. Présentation du Sujet et Contextualisation p.3

II. Description du Problème p.3

III. Solutions et Argumentaire pour le Choix des Algorithmes p.3

- Analyse de la Corrélation de Pearson
- Modèles de Régression Temporelle
- Apprentissage Automatique

IV. Protocole expérimental p.5

- Collecte des Données
- Nettoyage des Données
- Études des Caractéristiques
- Entraînement des Modèles
- Évaluation des Modèles
- Enregistrement et exploitation du Modèle

V. Résultats et limites p.11

VI. Améliorations et ouvertures p.12

VII. Conclusion p.13

VIII. Auto-évaluation p.13

IX. Sources p.13

I. Présentation du Sujet et Contextualisation

Le projet s'inscrit dans le domaine de la finance, avec un accent particulier sur le marché boursier. Le but est d'exploiter les techniques d'apprentissage automatique pour anticiper les fluctuations des prix des actions, une tâche à la fois complexe et cruciale dans le domaine financier. Cette capacité de prédiction est essentielle pour les investisseurs et les institutions financières, car elle peut influencer les décisions d'investissement et de trading, ainsi que la gestion des risques. Le contexte actuel du marché, caractérisé par sa volatilité et l'influence de facteurs économiques, politiques et sociaux, rend la tâche de prédiction encore plus pertinente et difficile.

II. Description du Problème

Le problème central est de prédire si le prix d'une action va augmenter ou diminuer, en se basant sur des données historiques. Les fluctuations des prix des actions sont influencées par une multitude de facteurs, y compris les performances financières de l'entreprise, les conditions du marché global, les événements politiques, et même le sentiment du marché extrait des actualités et des médias sociaux. Les modèles doivent donc capturer et analyser ces diverses données pour faire des prédictions précises.

- **Type de Problème** : Classification binaire. L'objectif est de prédire deux classes possibles - augmentation (1) ou diminution (0) du prix de l'action, grâce aux données historiques des prix des actions (Cela comprend les prix d'ouverture, de fermeture, les plus hauts, les plus bas et le volume des actions sur une période donnée).
- **Fonctionnalités Intrinsèques** : Données directement extraites du marché boursier (prix, volume).
- **Entrée** : Un dataset de données financières avec prix d'ouverture, fermeture, le volume et l'amplitude. Pour la suite de ce projet, nous prendrons un dataset des prix de l'action Apple entre le 01/01/2020 et le 31/12/2022.
- **Sortie Attendue** : Une variable binaire indiquant la hausse (1) ou la baisse (0) du prix de l'action.
- **Critères d'Évaluation** : Précision, rappel, score F1, et analyse des erreurs pour juger de la performance des modèles.

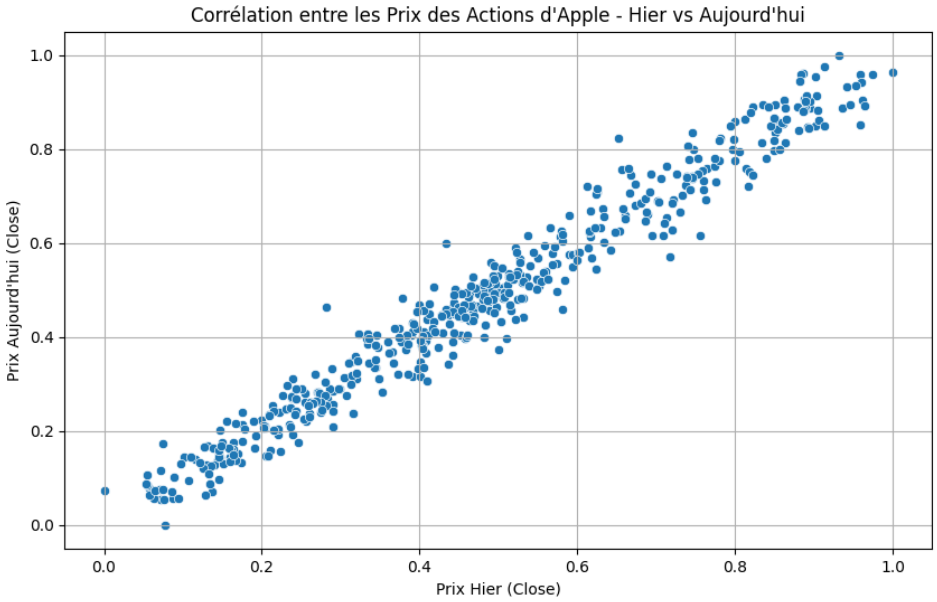
III. Solutions et Argumentaire pour le Choix des Algorithmes

La première étape serait de mettre en lumière le fait que les données futures dépendent en partie des données du passé. Pour cela nous réalisons une analyse statique afin d'étudier la corrélation entre les données.

Analyse de la Corrélation de Pearson

```
Corrélation de Pearson: 0.995823042834103, P-value: 0.0
```

La corrélation de Pearson est très élevée (près de 1), ce qui indique une forte relation linéaire positive entre le prix d'une action un jour donné et son prix le jour précédent.
La p-value de 0.0 suggère que cette corrélation est statistiquement significative.
On peut donc affirmer que le prix d'aujourd'hui est liées aux prix des jours précédents.



Pour essayer de prédire les cours il y a deux types d'approche, la première est :

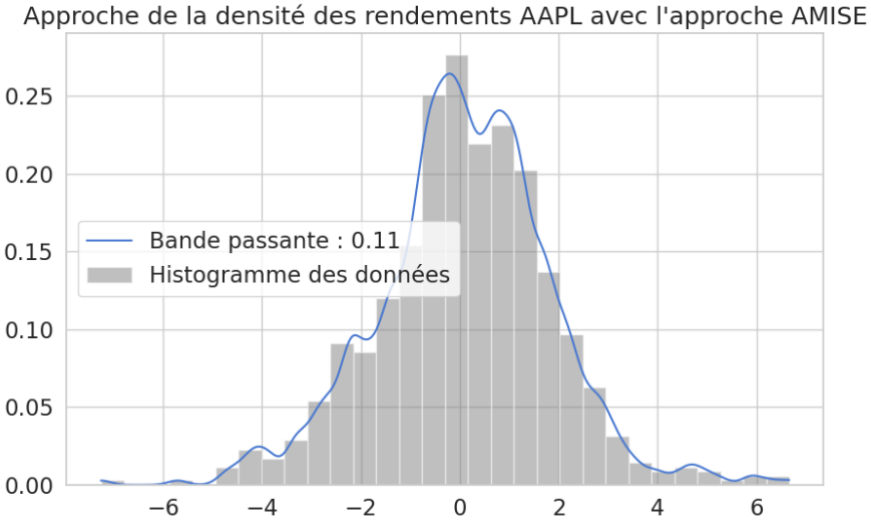
Modèles de Régression Temporelle : Utilisez des modèles de régression comme ARIMA (AutoRegressive Integrated Moving Average), AMISE (Asymptotic Mean Integrated Squared Error) ou encore LOOCV pour utiliser les valeurs passées pour prédire les valeurs futures.

L'estimation de densité par AMISE est une méthode utilisée en statistiques pour estimer la fonction de densité de probabilité d'un ensemble de données. Le principe fondamental de l'AMISE est d'optimiser le choix de la bande passante lors de l'utilisation d'estimateurs de densité à noyau, ce qui permet de comprendre comment les différentes valeurs de la variable sont distribuées et d'anticiper la probabilité d'occurrence de futures valeurs.

L'estimation de densité est particulièrement utile pour la modélisation de données où l'on ne suppose pas a priori une distribution spécifique, permettant ainsi une plus grande flexibilité et adaptabilité du modèle aux caractéristiques réelles des données observées.

L'optimisation de la largeur de bande pour minimiser l'erreur conduit à un équilibre entre le biais et la variance de l'estimateur de densité, trouvant ainsi le compromis optimal entre le lissage excessif (sous-estimation de la variance) et le surajustement (surestimation de la variance) des données. Cela permet de produire une estimation précise et robuste de la distribution des données, qui peut ensuite être utilisée pour évaluer les probabilités et les risques dans la prise de décision ou pour la prédiction de tendances dans des domaines tels que la finance.

Voici une approche de la densité des rendements de Apple par la méthode AMISE avec différentes bandes passantes :



Apprentissage Automatique : Des modèles peuvent apprendre des dépendances temporelles complexes, démontrant ainsi l'influence des données passées.

Dans le cadre de ce projet on se penchera sur cette deuxième approche, et on sélectionnera deux algorithmes d'apprentissage automatique : le Random Forest Classifier et le réseau de neurones Long Short-Term Memory (LSTM). Voici un aperçu des raisons justifiant le choix de ces algorithmes :

Le **Random Forest Classifier** se distingue par sa robustesse et sa précision, étant particulièrement reconnu pour sa capacité à gérer efficacement un grand nombre de caractéristiques tout en minimisant le risque de surajustement, un problème courant avec les arbres de décision individuels. Ce modèle excelle dans la gestion des données non linéaires, un atout majeur dans l'analyse des séries temporelles financières, où la non-linéarité est souvent la norme. Un autre avantage significatif du Random Forest est sa capacité à fournir des informations précieuses sur l'importance des différentes caractéristiques. Cette caractéristique peut être cruciale pour comprendre quels facteurs influencent le plus les prédictions des prix des actions.

Les réseaux **Long Short-Term Memory (LSTM)** sont spécialement conçus pour reconnaître et exploiter les dépendances à long terme dans les données de séries temporelles. Cette capacité est particulièrement précieuse dans le domaine de la prédiction des prix des actions, où la compréhension des tendances passées peut être cruciale pour anticiper les mouvements futurs. En tant que variante des réseaux de neurones récurrents, les LSTM sont aussi bien adaptés au traitement des séries temporelles financières. Leur flexibilité et adaptabilité permettent de traiter divers types d'entrées, ce qui est bénéfique dans le cadre de la diversité des données financières. De plus, les LSTM sont capables de capturer des relations complexes et non linéaires dans les données, une capacité essentielle dans un environnement financier souvent caractérisé par sa volatilité et son imprévisibilité.

IV. Protocole expérimental

1. Collecte des Données

Sélection des Sources de Données

Le choix de la source de données est crucial car il détermine la qualité et la fiabilité des informations sur lesquelles l'algorithme sera construit. Yahoo Finance a été identifié comme la principale source pour les raisons suivantes :

- Accessibilité : Yahoo Finance offre un accès facile et gratuit à des données historiques et en temps réel pour une large gamme de titres financiers.
- Complétude : Il fournit une vue d'ensemble complète du marché avec non seulement les prix des actions mais aussi des données fondamentales, telles que les dividendes et les ratios financiers.
- Fiabilité : En tant que plateforme reconnue, Yahoo Finance est considérée comme une source fiable de données financières.
- Facilité d'Intégration : Les données peuvent être extraites facilement via des API ou des bibliothèques Python dédiées comme yfinance.

Téléchargement des Données Historiques des Prix des Actions

Les données historiques sont essentielles pour analyser les tendances passées et construire des modèles prédictifs. La nature des données récupérées inclura le prix d'ouverture et de clôture, prix hauts et bas, le volume, le prix ajustés de clôture (pour tenir compte des ajustements pour les actions divisées et les dividendes pour une analyse plus précise).

```
Aperçu des données collectées :
                Open      High       Low     Close  Adj Close    Volume  Target
Date
2023-01-03  0.060800  0.044422  0.000000  0.000700   0.000694  0.675849       1
2023-01-04  0.012530  0.012631  0.012797  0.018760   0.018611  0.499315       0
2023-01-05  0.015948  0.000000  0.008297  0.000000   0.000000  0.436765       1
2023-01-06  0.000000  0.035765  0.010125  0.064399   0.063887  0.488887       1
2023-01-09  0.063506  0.080046  0.080439  0.071819   0.071248  0.358705       1
```

2. Nettoyage des Données

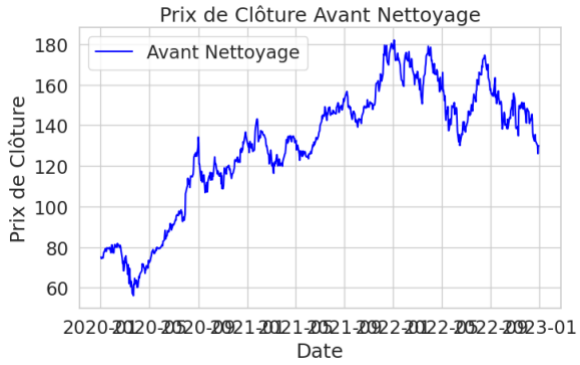
Gestion des Valeurs Manquantes ou Incorrectes

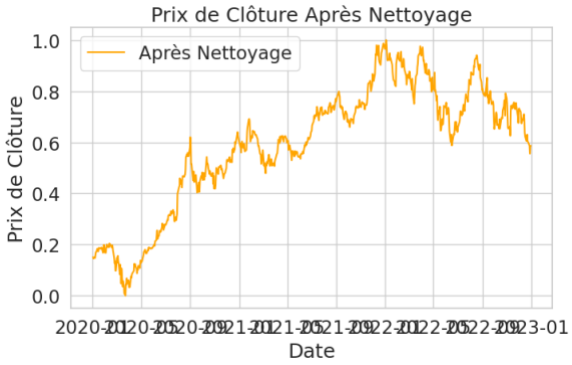
Avant de procéder à toute forme d'analyse, il est primordial de s'assurer que les données sont complètes et exactes. On va d'abord **Identifier les Valeurs Manquantes** (Détecter les données manquantes dans chaque série temporelle qui pourraient affecter l'analyse) et mettre en place une **Stratégies de Traitement** (Remplacement des données manquantes par la valeur précédente).

Normalisation ou Standardisation des Données

Les différentes échelles des données financières nécessitent une normalisation pour permettre aux modèles de traiter les caractéristiques sur un pied d'égalité.

Ainsi nous **Normalisons les données** (Réduire les données à une échelle commune sans distorsion des différences dans les gammes de valeurs) et nous les **Standardisons** (Centrer les données en soustrayant la moyenne et en divisant par l'écart-type pour obtenir une distribution standard).





3. Études des Caractéristiques

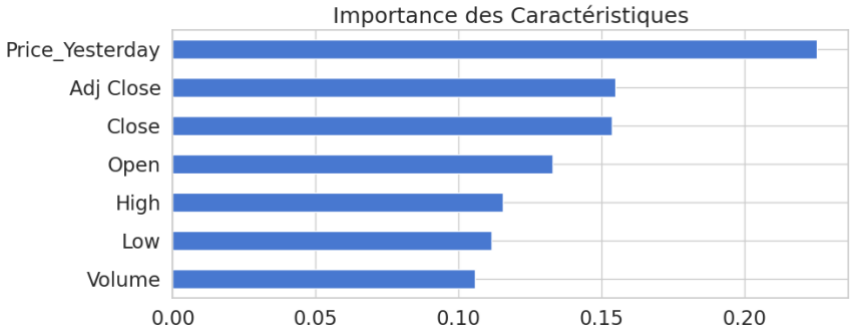
Identification des Caractéristiques Pertinentes

L'analyse de corrélation permet d'évaluer la relation entre chaque caractéristique et la variable cible pour identifier les variables potentiellement prédictives.

```
Target       1.000000
Close        0.045386
Adj Close    0.043406
Low         -0.007142
High        -0.014188
Open        -0.057403
Volume      -0.097048
```

L’importance des Caractéristiques par l’utilisation des techniques statistiques et des algorithmes de machine learning (comme les arbres de décision) permet d’estimer l'importance des caractéristiques.

```
Importance des caractéristiques:
Price_Yesterday     0.225306
Adj Close           0.155055
Close               0.153749
Open                0.132946
High                0.115691
Low                 0.111521
Volume              0.105732
```



Application de Techniques de Réduction de Dimensionnalité

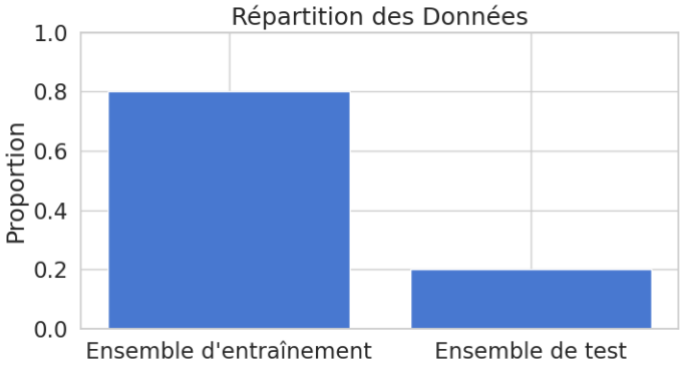
Le **PCA (Analyse en Composantes Principales)** transforme les caractéristiques en un ensemble réduit de variables non corrélées qui capturent la majeure partie de la variance dans les données.

```
Précision du modèle : 0.47577092511013214
```

4. Entraînement des Modèles

Division des Données

Les données sont divisées en **ensembles d'entraînement, de validation et de test**, pour permettre l'entraînement des modèles, le réglage des hyperparamètres et l'évaluation finale des performances.



Entraînement des Modèles sur l'Ensemble d'Entraînement

Le processus d'apprentissage sert à entrainer les modèles en utilisant les données d'entraînement et des mécanismes pour prévenir le surajustement, comme le dropout pour les LSTM. Il y a un **suivi des performances** pour s'assurer que les modèles apprennent effectivement des patterns prédictifs.

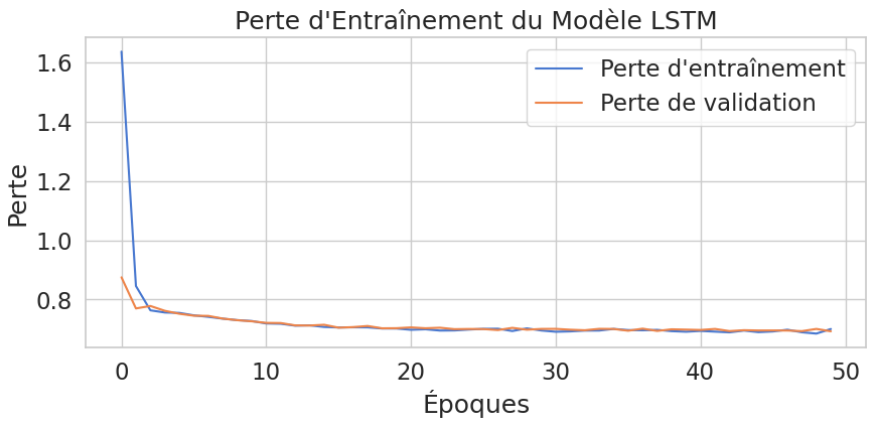
- **Modèle de forêt aléatoire (Random Forest)**

En utilisant une bibliothèque de machine learning telle que scikit-learn en Python, le modèle est construit avec la classe **RandomForestClassifier**. Une fois le modèle instancié, il est entraîné sur un ensemble de données d'entraînement (**X_train**, **y_train**), où **X_train** contient les caractéristiques et **y_train** les étiquettes correspondantes. L'entraînement du modèle (**rf_model.fit(X_train, y_train)**) lui permet d'apprendre à associer les caractéristiques aux étiquettes.
Après l'entraînement, le modèle est utilisé pour faire des prédictions sur un ensemble de test (**X_test**) via **rf_model.predict(X_test)**, produisant un ensemble de prédictions **y_pred**. Ces prédictions sont ensuite comparées aux vraies étiquettes (**y_test**) pour évaluer la performance du modèle.
La précision, calculée par **accuracy_score(y_test, y_pred)**, mesure la proportion de prédictions correctes par rapport au total. De plus, un rapport de classification (**classification_report(y_test, y_pred)**) est généré pour fournir un aperçu détaillé de la performance du modèle, incluant des métriques telles que la précision, le rappel et le score F1 pour chaque classe.

```
Random Forest - Précision: 0.5364238410596026
Random Forest - Rapport de Classification:
              precision    recall  f1-score   support

           0       0.57      0.53      0.55        81
           1       0.50      0.54      0.52        70

    accuracy                           0.54       151
   macro avg       0.54      0.54      0.54       151
weighted avg       0.54      0.54      0.54       151
```

- **Modèle de réseau de neurones LSTM (Long Short-Term Memory)**

Le modèle commence par une architecture séquentielle (via **Sequential()**) dans Keras, permettant l'empilement linéaire des couches. La première couche LSTM reçoit des séquences en entrée. Cette couche est suivie par une couche **Dropout** (technique utilisée pour prévenir le surajustement (overfitting) dans le réseau de neurones.) avec un taux de 20%, utilisée pour réduire le surajustement en désactivant aléatoirement des neurones pendant l'entraînement. Une configuration similaire est répétée avec une autre couche LSTM et Dropout. La dernière couche, une couche **Dense** avec une activation **Linéaire**, est typique pour les problèmes de classification binaire. Le modèle est ensuite compilé avec l'optimiseur Adam et utilise la perte de **binary_crossentropy**, adaptée aux tâches de classification binaire. La précision est choisie comme métrique d'évaluation.
L'entraînement est effectué sur 50 époques avec des lots de 32 observations, en utilisant à la fois les données d'entraînement (**X_train_lstm**, **y_train**) et de validation (**X_test_lstm**, **y_test**).



Ajustement des Hyperparamètres

Les techniques comme la recherche sur grille ou la recherche aléatoire sont appliquées pour explorer l'espace des hyperparamètres. La recherche sur grille est une méthode exhaustive qui essaie systématiquement toutes les combinaisons possibles d'hyperparamètres dans un espace de recherche prédéfini. Cette méthode est très systématique mais peut être très coûteuse en termes de temps et de ressources computationnelles, en particulier si l'espace des hyperparamètres est grand ou si le modèle est complexe et long à entraîner.

5. Évaluation des Modèles

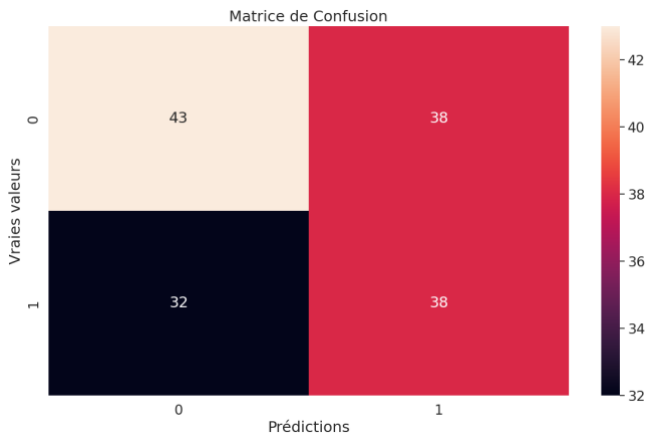
Évaluation sur l'Ensemble de Validation

Après avoir entraîné les modèles, il est crucial de les évaluer en utilisant un ensemble de données distinct de celui utilisé pour l'entraînement. Cela permet de s'assurer que les modèles peuvent généraliser et effectuer des prédictions précises sur de nouvelles données.

- **Évaluation des Performances :** La précision de la classification mesure la proportion des prédictions correctes parmi toutes les prédictions effectuées. Le rappel évalue la capacité du modèle à identifier correctement toutes les instances positives réelles. Le score F1 fournit un équilibre entre la précision et le rappel, utile lorsque les classes sont déséquilibrées. L'AUC (Area Under the Curve) mesure la performance globale du modèle en évaluant l'aire sous la courbe ROC, qui compare le taux de vrais positifs au taux de faux positifs à différents seuils.

```
Rapport de Classification pour Random Forest :
              precision    recall  f1-score   support

           0       0.57      0.53      0.55        81
           1       0.50      0.54      0.52        70

    accuracy                           0.54       151
   macro avg       0.54      0.54      0.54       151
weighted avg       0.54      0.54      0.54       151

Rapport de Classification pour LSTM :
              precision    recall  f1-score   support

           0       0.60      0.43      0.50        81
           1       0.51      0.67      0.58        70

    accuracy                           0.56       151
   macro avg       0.55      0.55      0.56       151
weighted avg       0.56      0.54      0.56       151
```

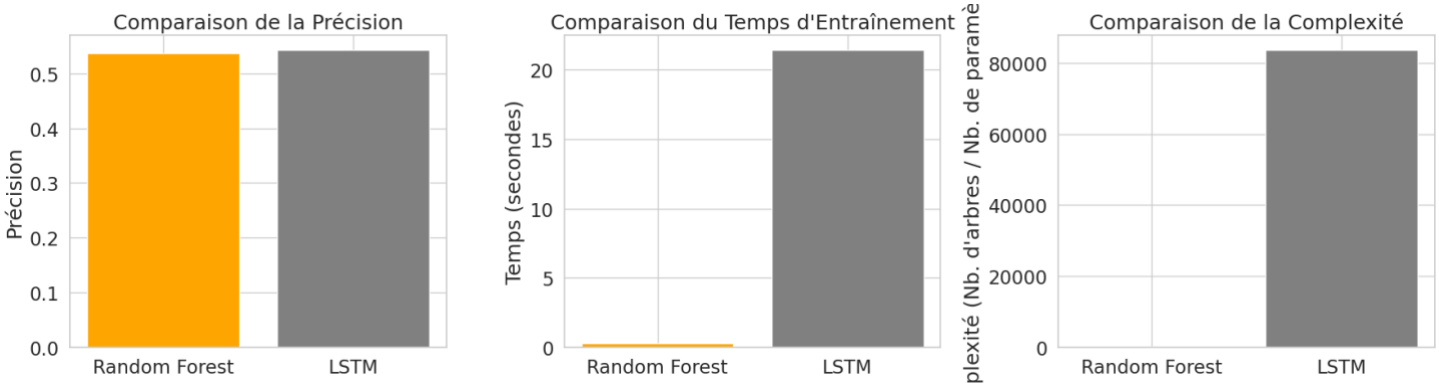
- **Interprétation des Résultats :** L'analyse de la matrice de confusion pour le modèle Random Forest donne une vue d'ensemble des performances du modèle, en montrant quantité notable de vrais positifs et de vrais négatifs cependant il faut noter qu'il y a également beaucoup de faux positifs et de faux négatifs.



Comparaison des Modèles

Cette phase implique la mise en concurrence des différents modèles pour sélectionner celui qui offre la meilleure performance. Les modèles sont évalués en fonction de leurs métriques de performance et de leur pertinence par rapport au contexte spécifique du problème de prédiction des tendances du marché boursier.

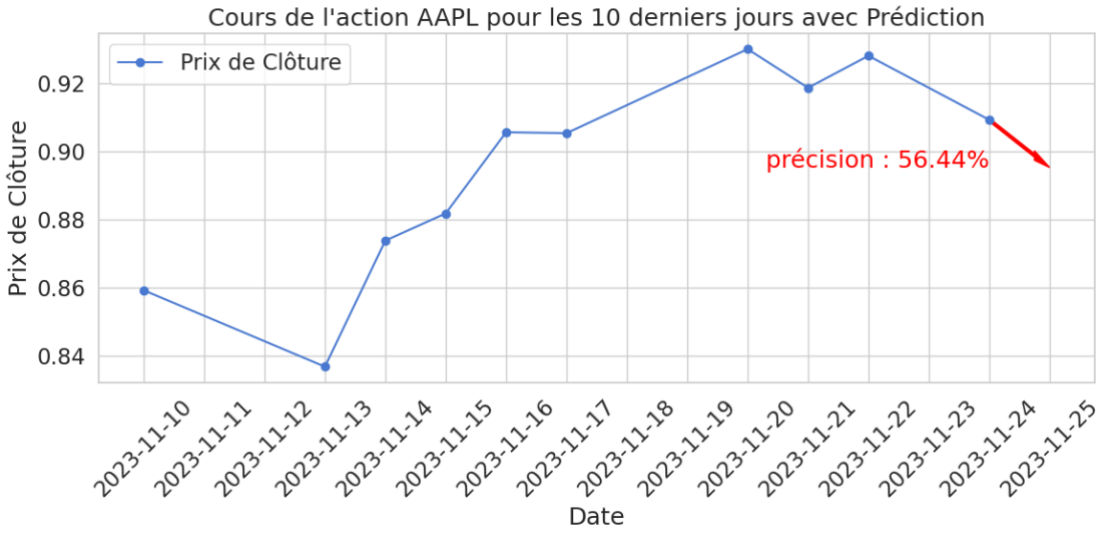
- La sélection du meilleur modèle passe par la définition des critères clairs basés sur la performance, la complexité, et le temps de calcul. Il faut aussi considérer les implications commerciales des différentes performances des modèles, telles que les coûts des faux positifs par rapport aux faux négatifs.
- Pour la prise de décision, on préférera les modèles qui offrent un bon équilibre entre la performance et la possibilité d'expliquer les prédictions mais aussi la meilleure robustesse (performance stable sur différentes périodes et conditions de marché).



6. Enregistrement et exploitation du Modèle

La sérialisation est le fait d'utiliser la bibliothèques **joblib** pour sauvegarder le modèle formé dans un fichier qui peut être chargé et utilisé ultérieurement.
Lors de l'**automatisation du processus de prédiction,** on charge le model entrainé et on exécute des prédictions à des intervalles réguliers. On peut également faire de la **surveillance en temps réel** afin de s'assurer que le modèle ne se dégrade pas en termes de performance ou qu'il n'y a pas de problèmes opérationnels.

La fonction finale est donc composée de toutes ces parties. Elle prend en argument le symbole de l'action à étudier et en sortie on a un graphique des 10 dernières périodes avec la prédiction et le pourcentage de précision de la prédiction. Voici à quoi cela s'apparente :



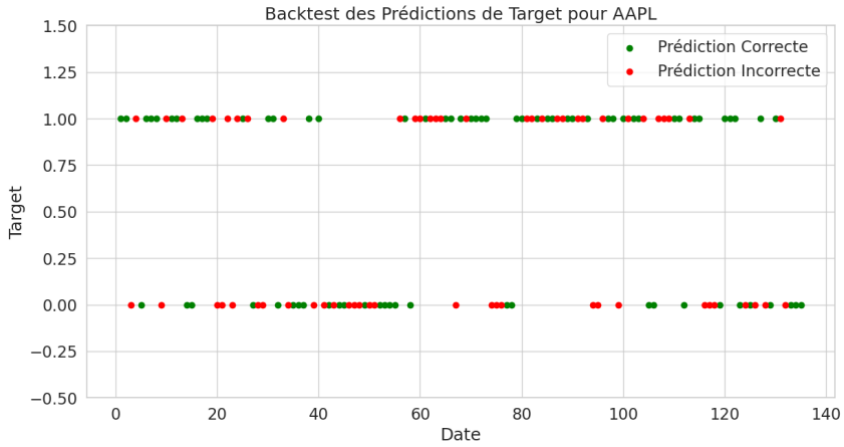
V. Résultats et Limites

Lors de l'exécution de l'algorithme le 26/12/2023, nous avons obtenu le graphe précèdent. Une **Prédiction de Baisse**, ce qui indique que le prix de l'action est plus susceptible de baisser plutôt que d'augmenter. Cela peut influencer les décisions de trading, telles que la vente d'actions, l'attente d'une opportunité d'achat à un prix inférieur, ou la mise en place de stratégies de couverture. Cependant, l'**exactitude des modèles** est à discuter. En effet, pour le modèle :

- **Random Forest (53.64% de précision)**, la précision est légèrement mieux que celle d'une décision aléatoire (50%), ce qui suggère que le modèle Random Forest n'est pas très efficace pour prédire la direction du marché dans cet ensemble de données spécifique.
- **LSTM (56,44% de précision)**, la précision est meilleure que celle du Random Forest, mais reste modeste. Ce niveau de précision indique que, bien que légèrement mieux que le hasard, le modèle LSTM a également du mal à capturer les tendances du marché de manière fiable.

Un backtest sur les 200 derniers jours de cette stratégie donne :

```
Précision: 55.56%
Prédictions correctes : 55.15%
Prédictions incorrectes : 44.12%
```



Bien que le modèle affiche une performance supérieure à une prédiction aléatoire, il y a encore une marge significative pour l'amélioration de la précision des prédictions.

D'autres parts, un **test de Causalité de Granger** donne ces résultats :

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=2.7407  , p=0.0982  , df_denom=751, df_num=1
ssr based chi2 test:   chi2=2.7516  , p=0.0972  , df=1
likelihood ratio test: chi2=2.7466  , p=0.0975  , df=1
parameter F test:         F=2.7407  , p=0.0982  , df_denom=751, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.2692  , p=0.7641  , df_denom=749, df_num=2
ssr based chi2 test:   chi2=0.5412  , p=0.7629  , df=2
likelihood ratio test: chi2=0.5410  , p=0.7630  , df=2
parameter F test:         F=1.6361  , p=0.1954  , df_denom=749, df_num=2Granger
```

Les p-values sont bien au-dessus du seuil de 0.05, cela indique qu'avec un ou deux retards, il n'y a pas de preuve significative de causalité de Granger. C'est-à-dire que même en considérant deux périodes précédentes, les prix des jours précédents ne semble pas être un

bon prédicteur pour le prix du jour suivant. Ainsi, l'utilisation des prix seuls ne sont pas suffisant seul comme indicateur.

VI. Améliorations et ouvertures

Améliorations possibles

Étant donné le niveau d'exactitude relativement faibles de ces modèles, s'appuyer uniquement sur les prédictions pour prendre des décisions de trading peut être risqué. On pourrait utiliser ces modèles comme un des nombreux outils d'analyse dans une stratégie de trading plus large, qui prendrait également en compte d'autres formes d'analyse (technique, fondamentale, sentimentale, etc.).

Ainsi les pistes d'amélioration possible serait le téléchargement d'indicateurs économiques pertinents comme les **indicateurs macroéconomiques** (PIB, taux d'intérêt et chômage). De plus, nous pourrions prendre en compte de nouvelles caractéristiques comme des **indicateurs Techniques** (moyenne mobile, RSI (Relative Strength Index), ou le MACD (Moving Average Convergence Divergence). Dans le but de créer des caractéristiques qui capturent les tendances sur différentes périodes (hebdomadaire, mensuelle). Explorer d'autres caractéristiques, ajuster les hyperparamètres ou augmenter la quantité de données pourrait améliorer la précision des modèles.

Cependant, les marchés reflètent le sentiment global des investisseurs. Prendre en compte les **données fondamentales spécifiques au secteur** (Nouvelles réglementations, changements de tarifs, etc) ou encore le sentiment des news (Analyse des nouvelles et des réseaux sociaux) semble être pertinent pour façonner un nouvel indicateur.

Analyse du sentiment de marché

Ainsi, nous tenterons d'ajouter à notre stratégie, l'influence d'un sentiment derrière une nouvelles pour la prise de décisions d'investissement. Une des solutions qui rend cela possible sont les LLM (Large Language Model). Il s'agit d'un type d'algorithme de deep learning qui est capable de comprendre et de générer du texte dans un large éventail de langues et de contextes, offrant des capacités impressionnantes de conversation, de rédaction, et d'analyse textuelle.

Pour ce module nous allons, dans un premier temps récupérer sur Yahoo finance les dernières nouvelles de l'action Apple puis nous allons voir le sentiment global de la journée. Beaucoup de librairie propose leurs modèles de LLM pré-entrainé (Langchain avec chatGPT de OpenAI, Transformers de Hugging Face pour BERT ou XLNet, etc). Ici nous utiliserons une approche assez simpliste, avec la fonction **SentimentIntensityAnalyzer** de le librairie **nltk**, pour déterminer un score de sentiment sur les titres des articles qui parlent d'Apple.

```
                                                 News Sentiment
0  I Stopped Buying Apple Products and Here's Wha...    Baisse
1  The Apple Store closes its doors for Black Friday    Neutre
2  Apple wants to break its alliance with Goldman...    Neutre
3  Swiss Central Bank Slashed AMC Stake, Sells Ap...    Baisse
4  We won't be able to uninvent it': Warren Buffe...    Hausse
5   20 Most Disabled-Friendly Countries in the World    Neutre

    le sentiment global du marché aujourd'hui est : Baisse
```

Pour améliorer notre approche, il serait judicieux de considérer l'article dans son intégralité et pas seulement le titre, tout en évaluant l'influence potentielle d'une information sur le cours de l'action. Par exemple, une nouvelle telle que la « démission de Tim Cook » pourrait avoir un impact négatif plus significatif que la fermeture des Apple Stores pour le Black Friday. Il serait également pertinent de créer une base de données historiques des nouvelles, en tenant compte des événements passés qui ont eu ou pourraient avoir un effet à long terme sur la performance de l'entreprise.

VII. Conclusion

En résumé, même si les modèles ont effectué des prédictions, leur efficacité limitée indique qu'ils ne devraient pas être utilisés isolément, mais plutôt en complément d'autres outils, comme des indicateurs techniques. Il est également crucial d'examiner une stratégie qui intègre les divers programmes précédemment développés. L'objectif de ces outils est d'aider les traders dans leurs prises de décision, et non de prendre ces décisions à leur place.

VIII. Auto-évaluation

Au cours de ce projet, j'ai développé une compréhension approfondie dans les modèles de machine learning tels que le Random Forest et LSTM mais aussi en LLM, et j'ai affiné mes compétences en prétraitement et analyse statistique des données (avec notamment la méthode AMISE). Ce travail a amélioré ma capacité à résoudre des problèmes complexes, renforçant ma pensée critique et ma compréhension du domaine financier. J'ai également acquis de précieuses compétences en visualisation de données.

IX. Sources

[1] Practical bandwidth selection in deconvolution kernel density estimation - A. Delaigle & I. Gijbels
https://www.sciencedirect.com/science/article/abs/pii/S0167947302003298

[2] Density estimation for statistics and data analysis - B.W. Silverman

[3] Conservation machine learning: a case study of random forests - Moshe Sipper & Jason H. Moore
https://www.nature.com/articles/s41598-021-83247-4

[4] Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks - Ralf C. Staudemeyer & Eric Rothstein Morris
https://arxiv.org/pdf/1909.09586.pdf

[5] NLTK – Documentation (Sample usage for sentiment)
https://www.nltk.org/howto/sentiment.html