

ML TEMPS RÉEL SUR FLUX DE DONNÉES

Présenté par :

- EL BOUGRINI Nassim: EMI, Génie industriel.
- HAFSI Siham : ENSA KHOURIBGA , Informatique et ingénierie des données
- CHAHIDI Khadija: ENSA Al Hoceima, Ingénierie des données
- RAHMOUN Hayat:FST Marrakech, Modélisation Calcul Scientifique pour L'ingénierie Mathématique
- AGOUMI Achraf: ENSA Al Hoceima, Ingénierie des données
- NACHOUR Ilham: ENSA Al Hoceima, Ingénierie des données

encadré par :

- Thierry BERTIN

Plan

Confluent Cloud

les données

Prétraitement des données

Visualisation

Ingénierie des caractéristiques

Modèle ARIMA

XGboost

Random Rorest

Modèle *ARIMA*

Définition

L'ARIMA, acronyme pour "AutoRegressive Integrated Moving Average", est un modèle statistique utilisé pour analyser et prévoir des séries temporelles. Une série temporelle est une séquence de données observées à intervalles réguliers dans le temps, comme les ventes mensuelles d'un produit, les températures quotidiennes, les cours boursiers, etc.

L'ARIMA est largement utilisé dans des domaines tels que l'économie, la finance, la météorologie et l'analyse de données.

L'idéologie derrière ARIMA s'articule autour de trois composants fondamentaux :
AutoRegressive (AR) : La méthode autorégressive utilise un cadre de régression pour prédire une variable en fonction de ses valeurs passées. Un modèle AR d'ordre p peut s'écrire :

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

y_t représente la variable à prédire au temps t . p est le nombre de décalages temporels utilisés. Les coefficients $\phi_1, \dots, \phi_{t-p}$ déterminent l'influence de chaque variable retardée sur la valeur actuelle de y_t . La valeur de y_t est déterminée par une régression linéaire multiple de p variables indépendantes $y_{t-1}, y_{t-2}, \dots, y_{t-p}$.

Moving Average (MA) : La méthode de la moyenne mobile (MA) vise à capturer la relation entre une observation et les erreurs résiduelles des prédictions précédentes. Dans un modèle MA d'ordre q , il peut être exprimé comme suit :

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$$

Le terme ε_t représente l'erreur ou le résidu au temps t , qui est la différence entre la valeur réelle de y_t et la valeur prévue basée sur les prédictions précédentes. La valeur y_t est calculée sur la base des q erreurs y_t commises par le modèle précédent. Ainsi, chaque terme successif regarde un peu plus loin dans le passé pour incorporer les erreurs commises par ce modèle dans le calcul actuel.

Integrated (I): est une technique pour rendre les données de séries chronologiques stationnaires, ce qui est une exigence cruciale pour modéliser des données de séries chronologiques. La stationnarité est importante dans l'analyse des séries chronologiques car elle garantit que les propriétés statistiques des données, telles que la moyenne et la variance, restent constantes dans le temps. Cette hypothèse est nécessaire car si le modèle évolue dans le temps, il devient difficile d'estimer les paramètres avec précision. Pour atteindre la stationnarité, une approche courante est la différenciation, qui consiste à calculer les différences entre des observations consécutives dans la série chronologique. La différenciation aide à stabiliser la moyenne de la série chronologique en éliminant les changements de son niveau, en supprimant efficacement les tendances et la saisonnalité. Dans le modèle ARIMA, le paramètre « d » représente le nombre de fois où l'opération de différenciation est appliquée à la série chronologique. Chaque opération de différenciation soustrait la valeur de l'observation courante y_t par la valeur décalée y_{t-d}



En combinant ces trois éléments, un modèle ARIMA est spécifié en utilisant trois paramètres principaux :

- **p** (ordre AR) : Le nombre de retards utilisés pour la régression autoregressive.
- **d** (ordre d'intégration) : Le nombre de différenciations nécessaires pour rendre la série temporelle stationnaire.
- **q** (ordre MA) : Le nombre de retards utilisés pour la moyenne mobile.



Le processus de création d'un modèle ARIMA implique généralement les étapes suivantes :

Vérifier la stationnarité des séries chronologiques : soit en utilisant diagramme de série chronologique ou bien Graphique ACF et graphique PACF

Différenciation : Si la série n'est pas stationnaire, appliquer des différenciations jusqu'à ce que la série devienne stationnaire.

Identification des ordres : Utiliser des outils comme les graphiques ACF (fonction d'autocorrélation) et PACF (fonction d'autocorrélation partielle) pour identifier les ordres p , d et q optimaux.

Construction du modèle : Basé sur les ordres identifiés, construire le modèle ARIMA.

Ajustement et évaluation : Ajuster le modèle aux données historiques et évaluer ses performances en utilisant des métriques telles que l'erreur quadratique moyenne (RMSE) ou l'erreur absolue moyenne (MAE).

Prévisions : Utiliser le modèle ajusté pour faire des prévisions sur les valeurs futures de la série temporelle.

Comment choisir p, q, d ?

1. Ordre d (Différenciation) :

- Si la série temporelle montre une tendance évidente, effectuez une différenciation ($d=1$ ou plus) jusqu'à ce que la série semble stationnaire.
- Si la série semble déjà stationnaire, $d=0$ suffit.

2. Ordre p (pour AR) :

- Observez les pics dans le PACF. Si le pic au retard k est significatif, envisagez d'utiliser $p=k$.
- Si les pics dans le PACF s'éteignent rapidement, cela peut indiquer que peu de termes AR sont nécessaires.

3. Ordre q (pour MA) :

- Observez les pics dans l'ACF. Si le pic au retard k est significatif, envisagez d'utiliser $q=k$.
- Si les pics dans l'ACF s'éteignent rapidement, cela peut indiquer que peu de termes MA sont nécessaires.

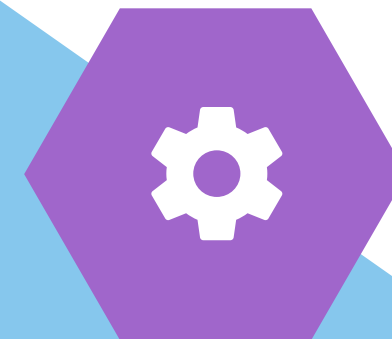
XGBoost



C'est quoi?



Pourquoi?

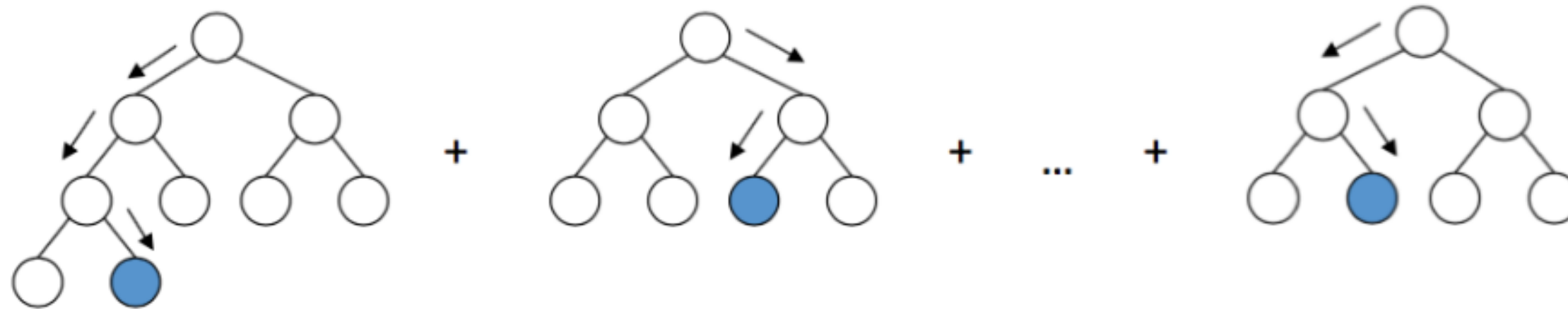


Comment?

C'est quoi?

XGBoost :

L'approche de base de XGBoost consiste à combiner les prédictions de plusieurs modèles faibles, généralement des arbres de décision, pour former un modèle global plus puissant. Chaque arbre est construit de manière séquentielle en essayant de corriger les erreurs du modèle précédent.





pourquoi XGBoost ?

Haute Performance

XGBoost est conçu pour être très rapide et efficace, ce qui le rend adapté au traitement de grands ensembles de données et de problèmes complexes

Précision

XGBoost combine les prédictions de nombreux modèles faibles pour former un modèle global plus précis. Cela lui permet de capturer des motifs complexes et de réaliser de meilleures prédictions

Régularisation

XGBoost intègre des techniques de régularisation telles que la réduction de la profondeur des arbres et la pénalisation des termes de complexité (L1 et L2). Cela permet de réduire l'overfitting et d'améliorer la généralisation du modèle



pourquoi XGBoost ?

Sélection de Fonctionnalités

XGBoost fournit des scores d'importance des fonctionnalités, ce qui vous permet de sélectionner les fonctionnalités les plus pertinentes pour votre modèle et de comprendre quelles fonctionnalités contribuent le plus aux prédictions.

Gestion des Données Manquantes

XGBoost est capable de gérer automatiquement les données manquantes lors de l'entraînement et de la prédiction

Gestion des Données Temporelles

XGBoost est capable de gérer les données temporelles en incorporant des caractéristiques retardées et en permettant des prédictions sur des fenêtres temporelles futures

Comment?



1

XGBoost commence par générer un arbre décisionnel simple.

2

L'algorithme calcule ensuite la perte entre les prédictions de l'arbre et les valeurs cibles.

3

XGBoost génère ensuite un nouvel arbre décisionnel qui est optimisé pour réduire la perte

4

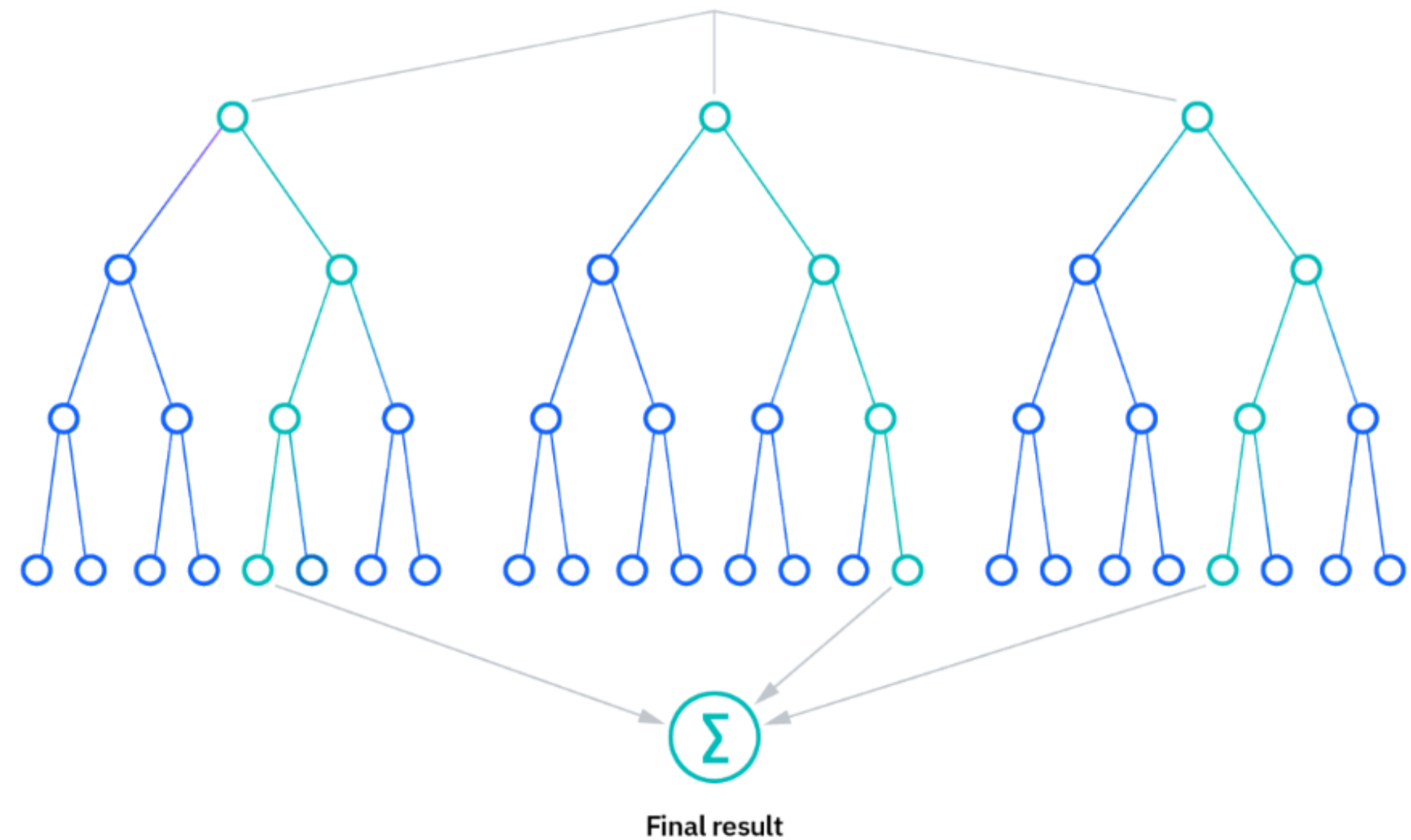
L'algorithme itère les étapes 2 et 3 jusqu'à ce qu'il atteigne un seuil de précision souhaité.
.

**L'algorithme de
forêt aléatoire**

RandomForest

Définition:

Le **Random Forest** est un algorithme d'apprentissage automatique supervisé utilisé pour résoudre des problèmes de classification et de régression. Il appartient à la famille des méthodes d'ensemble, qui combinent les prédictions de plusieurs modèles individuels pour améliorer la précision et la stabilité.



Comment ça fonctionne ?

1.Création d'arbres de décision : Le Random Forest crée un ensemble d'arbres de décision, chacun étant formé sur un sous-ensemble aléatoire de données et de caractéristiques. Cela aide à réduire le surajustement (overfitting) et à améliorer la généralisation.

Comment ça fonctionne ?

2.Bootstrap agrégé (Bagging) : Le Random Forest utilise une technique appelée "bagging" pour former chaque arbre. Il sélectionne aléatoirement des échantillons avec remplacement à partir de l'ensemble de données pour chaque arbre, ce qui crée une diversité dans les ensembles d'entraînement.

Comment ça fonctionne ?

3.Sous-ensemble de caractéristiques : Pour chaque arbre, un sous-ensemble aléatoire de caractéristiques est sélectionné à partir de l'ensemble complet. Cela encourage la diversité entre les arbres et réduit la corrélation entre les prédictions.

Comment ça fonctionne ?

4.Vote majoritaire : Lors de la prédiction, les résultats de tous les arbres sont agrégés. En classification, le résultat est déterminé par un vote majoritaire. En régression, les prédictions sont en moyenne pour obtenir une valeur finale.

Pourquoi RandomForest ?





Performance améliorée : Les prix des actions sont influencés par de nombreuses variables complexes. En utilisant Random Forest, vous pouvez tirer parti de la puissance de plusieurs arbres de décision pour capturer ces relations complexes et obtenir de meilleures prédictions.



Gestion des données : Le Random Forest peut gérer les valeurs aberrantes et les données manquantes plus efficacement que les arbres de décision simples, ce qui est important compte tenu de la nature imprévisible des données financières.

Pourquoi RandomForest ?

 **Réduction du surajustement** :Étant donné que les prix des actions peuvent être volatils et influencés par des facteurs divers, Random Forest peut aider à réduire le risque de surajustement en combinant les prédictions de plusieurs arbres.

 **Analyse des caractéristiques** :Le Random Forest peut également fournir des informations sur l'importance relative des différentes caractéristiques (variables) dans la prédiction des prix des actions, ce qui pourrait aider à mieux comprendre les tendances du marché.

Merci !