Master 2 Informatique et Science des Données

Université Paris Saclay

Rapport projet data

D. Jeannel

Vendredi 22 décembre 2023

Objectif : anticiper les évolutions hebdomadaires d'actions via des méthodes de Machine Learning

Problématique

L'objet du travail est, à partir de données boursières journalières, d'anticiper l'évolution hebdomadaire (5 jours prochains) des cours boursiers en utilisant des outils de machine learning.

A partir d'historiques de cotations boursières, on cherche à prévoir sur les 5 prochains jours l'évolution en pourcentage du cours d'une action : baisse (−2%), stabilité (entre −2% et +2%) ou hausse (+2%). Le problème revient à considérer une modélisation multi-classes.

Différentes approches peuvent être étudiées : modèles statistiques, méthodes basées sur les arbres, méthodes à noyaux, réseaux de neurones…

Le choix des techniques est laissé à votre appréciation et à votre intuition.

Les résultats des prévisions seront comparés par des mesures de performance détaillés dans les paragraphes suivants.

Données

Le fichier Excel « Historiques_cours_boursiers.xlsx » contient les historiques boursiers journaliers de 13 cours boursiers de 2000 à 2023 ou de 2004 à 2023 ou de 2010 à 2023.

Les onglets du fichier Excel portent les Ticker (index cours boursier sous Yahoo Finance) :

Ticker	Intitulé	Historique données
^IXIC	NASDAQ Composite	03/01/2000 au 20/12/2023
TTE	TotalEnergies	03/01/2000 au 20/12/2023
GE	General Electric	03/01/2000 au 20/12/2023
RMS.PA	Hermès International	03/01/2000 au 20/12/2023
TSLA	Tesla	29/06/2010 au 20/12/2023
JNJ	Johnson & Johnson	29/06/2010 au 20/12/2023
BRK-B	Berkshire Hathaway	03/01/2000 au 20/12/2023
UNH	UnitedHealth	03/01/2000 au 20/12/2023
NVDA	NVIDIA	03/01/2000 au 20/12/2023
MSFT	Microsoft	03/01/2000 au 20/12/2023
AAPL	Apple	03/01/2000 au 20/12/2023
GOOG	Alphabet	19/08/2004 au 20/12/2023
AMZN	Amazon	03/01/2000 au 20/12/2023

Dans chaque onglet Excel, pour chaque ticker, on dispose des informations journalières :

- Open : prix de l'action à l'ouverture ;
- High : prix le plus élevé du jour de cotation (H) ;
- Low : prix le plus bas sur la journée (L) ;
- Close : prix de clôture de l'action en fin de séance (C) ;
- Volume : nombre d'actions achetées et vendues d'une action au cours de la journée (V) ;
- Adjusted : mesure du profit de l'investissement par rapport au risque d'investissement sur une période donnée. Plus le risque d'un investissement est faible, plus le rendement (adjusted) est élevé.

Le prix de clôture, le prix le plus élevé, le prix le plus bas et le volume de transactions sont retenus pour établir des indicateurs.

Variable à expliquer (target) et Variables explicatives (features)

Variable à expliquer

Elle est définie de la manière suivante à partir du prix de clôture :

Tendance	Définition
Hausse	si 100. $C_{t+5}/C_t > 2\%$
Stable	si $-2\% < 100. C_{t+5}/C_t < 2\%$
Baisse	si 100. $C_{t+5}/C_t < -2\%$

Variables explicatives

Les informations journalières permettent de construire des indicateurs de trading basés sur les moyennes mobiles exponentielles des prix de clôture (C).

La définition de la moyenne exponentielle (EMA) est la suivante :

$$EMA_t(T) = \alpha \sum_{i=0}^{T} (1-\alpha)^i C_{t-i} \text{ avec } \begin{cases} \alpha = \dfrac{2}{T+1} \\ T \text{ la période considérée} \end{cases}$$

A partir de l'EMA, on peut construire les indicateurs de trading suivant :

1. *Indicateurs MACD*

Indicateur MACD	**Formulation**	**Valeurs des variables**
Ligne MACD	$MACD_t(T_1,T_2) = EMA_t(T_1) - EMA_t(T_2)$	T1 = 12, T2 = 26
Signal MACD	$Signal_t(T_3) = \alpha \sum_{i=0}^{T_3} (1-\alpha)^i MACD_{t-i}(T_1,T_2)$	T3 = 9
Histogramme	$H_t(T_1,T_2,T_3) = MACD_{t-i}(T_1,T_2) - Signal_t(T_3).$	

Dans la pratique, les praticiens utilisent l'une des deux règles suivantes pour anticiper les hausses ou les baisses de tendance à partir du MACD :

- Variation haussière si la ligne MACD est positive MACDt > 0 ou si l'histogramme est positif Ht > 0 ;
- Variation à la baisse si la ligne MACD est négative MACDt < 0 ou si l'histogramme est négatif Ht < 0.

2. *Indicateurs RSI*

Le RSI (Relative Strength Index) est un indicateur qui mesure la puissance d'un mouvement boursier en période de hausse ou de baisse. Il est déduit à partir du prix de clôture. Sa formule est la suivante:

$$RSI_t(N) = \frac{\sum_{i=0}^{N-1}(C_{t-i} - C_{t-i-1})\mathbb{I}\{C_{t-1} > C_{t-i-1}\}}{\sum_{i=0}^{N-1}|C_{t-i} - C_{t-i-1}|} \cdot 100$$

Selon les valeurs de N, plusieurs règles empiriques sont proposées :

- si N = 21, les spécialistes considèrent que la tendance du marché est à la hausse si RSI > 50 et à la baisse si RSI < 50 ;
- si N = 14, le marché est jugé haussier si RSI > 70 et baissier si RSI < 30.

Les deux règles sont à évaluer en pratique.

3. *Indicateurs ATR*

L'ATR (Average True Range) est une mesure de la volatilité du marché. Il est fonction du prix de clôture, du prix le plus haut et du prix le plus bas. On considère sur plusieurs instants le plus grand des écarts entre le prix le plus haut moins le prix de clôture, la valeur absolue du prix le plus haut moins le prix de clôture, et la valeur absolue entre le prix le plus bas moins le prix de clôture. L'ensemble des écarts obtenus est ensuite moyenné.

Son écriture mathématique est assez simple. Pour une date i, on définit le True Range TR_i, puis l'ATR qui est une moyenne des écarts sur une période N :

$$TR_i = \max\left(H_i - L_i, |H_i - C_i|, |L_i - C_i|\right)$$

$$ATR(N)) = \frac{\sum_{i=1}^{N} TR_i}{N}$$

L'indicateur dépend du nombre de périodes N. L'augmentation de N conduit à un meilleur lissage de la volatilité du cours boursier. Les professionnels utilisent des valeurs de N variant de 7 à 14.

Une forte valeur de l'ATR indique une volatilité accrue sur le marché. Un renversement de prix avec une augmentation de l'ATR peut indiquer un renversement de cours.
Une faible valeur de l'ATR est signe d'une volatilité faible. Une période prolongée de faibles valeurs d'ATR peut indiquer une zone de consolidation et la possibilité d'un mouvement de continuation.

4. *Indicateurs Volume de transactions*

Le volume de transactions donne une information sur la liquidité et l'intérêt d'une action sur le marché. Plus les volumes de transaction à l'achat ou à la vente sont élevés, plus l'intérêt pour l'action est important, et inversement.
L'indicateur le plus utilisé en pratique est le CMF (Chaikin Money Flow). Il est fonction du prix le plus haut, du prix le plus bas, du prix de clôture et du volume de transactions :

$$CMF_t(N) = \frac{\sum_{i=0}^{N} \frac{(C_{t-i} - L_{t-i}) - (H_{t-i} - C_{t-i})}{H_{t-i} - L_{t-i}} \cdot V_{t-i}}{\sum_{i=0}^{N} V_{t-i}}$$

Les professionnels recommandent de prendre N = 21 ou 28. L'interprétation est la suivante :

- Lorsque le CMF est supérieur à 0.25 (resp. inférieur à −0.25), le marché est à la hausse : le prix de clôture et le volume de transactions augmentent ;
- lorsque le CMF est inférieur à −0, 25, il s'agit d'un signal baissier du marché ;
- Si le CMF est reste en dessous de zéro et que le prix de clôture monte, alors il y a un risque de retournement de marché.

5. Autres variables explicatives.

A côté des indicateurs de tradings décrits précédemment, il est possible de créer d'autres indicateurs à partir du prix de l'action à l'ouverture (Open), du prix le plus élevé du jour de cotation (High), du prix le plus bas sur la journée (Low), du prix de clôture de l'action en fin de séance (Close), du nombre d'actions achetées et vendues d'une action au cours de la journée (V) et de la mesure du profit de l'investissement par rapport au risque d'investissement sur une période donnée.

Cela peut être des indicateurs retardés, des écarts…. La création d'autres indicateurs est laissée au soin du datascientist·e.

Evaluation des méthodes

Les prévisions des méthodes sur les échantillons de test seront évaluées sur la base de métriques basées sur les observations :

Les métriques retenues sont :

— Accuracy. Elle quantifie la *proportion de modalités correctement prédites* par rapport au nombre total de modalités de chaque observation. Elle est définie de la manière suivante :

$$Accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{\left| Y_i \wedge \hat{Y}_i \right|}{\left| Y_i \vee \hat{Y}_i \right|}$$

— Hamming. C'est la *métrique d'évaluation la plus courante* dans la littérature. Elle se définit comme le *ratio du nombre d'erreurs de prévisions des modalités par le nombre total de modalités.*

$$Hamming = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{\left| \mathbb{I}\left\{ Y_i \neq \hat{Y}_i \right\} \right|}{|L|}$$

et sur des métriques basées sur les modalités de la target (hausse, stable, baisse) :

En ce qui concerne les métriques basées sur les étiquettes, il existe deux façons de procéder :

— L'approche *macro-moyenne* (macro-averaging). Pour chaque étiquette, les quantités (TP, FN, FP) et les métriques $(Precision(Pr) = \frac{TP}{TP+FP}$, $Recall(Re) = \frac{TP}{TP+FN}$, $F-measure = 2.\frac{Pr.Re}{Pr+Re})$ sont déterminées. La moyenne de la métrique choisie est déduite comme mesure finale. Autrement dit, les métriques sont calculées individuellement pour chaque label et la moyenne est obtenue en les divisant sur le nombre d'étiquettes :

$$\text{Macro-average} = \frac{1}{|L|}\sum_{i=1}^{|L|} M_i \text{ où } M = Pr, Re, F-measure$$

— et l'approche de *micro-moyenne* (micro-averaging). Sur toutes les étiquettes, les quantités (TP, FP, FN) sont déterminées, puis la métrique choisie $(Pr,\ Re,\ F-measure)$ est calculée sur les sommes obtenues :

$$\text{Micro-average} = M(\sum_{i=1}^{|L|} TP_i, \sum_{i=1}^{|L|} FN_i, \sum_{i=1}^{|L|} FP_i) \text{ où } M = Pr, Re, F-measure$$

Travail demandé

- Sélectionner 6 indices boursiers parmi les 13 contenues dans le fichier Excel ;
- Créer des variables explicatives (features) ;
- Proposer différentes modélisations de la variable à expliquer (target). Un minimum de 8 méthodes est demandé ;
- Construire des modélisations sur des échantillons d'apprentissage ;
- Tester les modélisations choisies sur les échantillons de test ;
- Evaluer la performance des prévisions sur les échantillons tests en utilisant les métriques globales et locales ;
- Faire des recommandations, des interprétations sur les performances des méthodes et sur la problématique.

Consignes pour l'évaluation :

- Le travail à rendre est un rapport au format pdf. Un seul fichier sera accepté et considéré dans l'évaluation. Le fichier pdf devra contenir le rapport et le code de calcul. La date limite est fixée au 31 janvier 2024 18 heures.

- Le plan du rapport à adopter est le suivant :
    1. Introduction
    2. Données
    3. Méthodologie
    4. Résultats
    5. Conclusions et perspectives
    6. Annexe : code de calcul

- Pour respecter l'anonymat dans la correction des rapports, le fichier pdf devra comporter uniquement le numéro d'identifiant de l'étudiant·e, sans son nom, sans son prénom (ex : 123456.pdf). Idem dans le document pdf, il ne devra faire référence en aucun cas au nom et au prénom de l'étudiant·e.

- La qualité du texte (style, fautes d'orthographes), l'interprétation des résultats, les remarques, les recommandations… seront prise en compte dans la notation.

- Un graphique ou une série de graphiques mal adaptés ou n’apportant aucune valeur ne sera pas prise en compte dans l’évaluation.

- Le non-respect de ces consignes sera pris en compte dans la notation finale.