Vers une Intelligence d'Investissement
Modélisation des Marchés par le Machine Learning et le NLP

Rapport de projet final

Master 2 Ingénierie Statistique et Financière en Apprentissage
Université Paris Dauphine – PSL

Réalisé par :

Zakaria BOUROUBA
Pauline NEEL
Elisa DIDIER

Date de remise : 18 Mai 2025

Table des matières

1 Introduction / Motivations

La complexité croissante des marchés financiers et la diversité des sources d'information imposent de nouvelles approches pour anticiper les mouvements de prix et prendre des décisions d'investissement éclairées. Grâce aux avancées du Machine Learning et du Traitement Automatique du Langage Naturel (NLP), il est désormais possible d'automatiser la génération de signaux exploitables à partir de données financières, techniques, ou encore textuelles.

Ce projet vise à intégrer ces différentes briques technologiques dans une architecture cohérente, capable de produire des recommandations d'achat, de conservation ou de vente sur des actifs financiers.

2 Travaux connexes

La littérature récente démontre un intérêt croissant pour l'application conjointe du machine learning et du traitement du langage naturel (NLP) à la prédiction des marchés financiers. Plusieurs travaux ont inspiré notre approche, notamment sur l'intégration de signaux textuels et fondamentaux.

Analyse sémantique des actualités financières.
Dans leur étude fondatrice, Lavrenko et al. (2000) *"Financial News Predicts Stock Market Volatility Better than Close Price"* ont montré que les actualités économiques pouvaient fournir des signaux anticipant les variations de volatilité des marchés. En analysant le contenu sémantique de news, ils ont pu améliorer les modèles de prédiction classiques.

Fusion des données temporelles et textuelles.
Nguyen et Shirai (2015), dans *"Stock Movement Prediction from Tweets and Historical Prices"*, proposent une méthode combinant les séries temporelles de prix avec le sentiment extrait de tweets. Leur modèle de classification illustre l'efficacité d'un apprentissage multi-sources, concept que nous avons appliqué à travers les TP6 à TP8.

Deep learning appliqué aux ratios financiers.
Feng et al. (2019), dans *"Deep Learning for Stock Selection Based on Fundamental Analysis"*, explorent l'utilisation de réseaux de neurones profonds pour modéliser les performances d'entreprise à partir de leurs ratios fondamentaux. Cette logique est à la base de nos approches dans les TP1, TP2 et TP5.

Ces travaux montrent l'intérêt croissant pour les méthodes hybrides combinant données fondamentales, textuelles et temporelles, ce qui soutient la pertinence de notre projet.

3 Clustering

Objectif

L'objectif du clustering est de regrouper les entreprises selon leurs caractéristiques structurelles communes afin d'identifier des segments homogènes du marché. Cela permet de construire des portefeuilles diversifiés, de mieux contextualiser les signaux d'achat/vente, et d'analyser les performances en tenant compte des similarités inter-entreprises.

Méthodes utilisées

Trois types de clustering ont été appliqués :

— **K-Means** sur les ratios fondamentaux (*forwardPE*, *beta*, *priceToBook*, *returnOnEquity*) pour segmenter les profils financiers.

— **Clustering hiérarchique** (linkage de Ward) sur des indicateurs de risque (*debtToEquity*, *currentRatio*, *margins*) pour une vision structurelle des expositions.

— **DBSCAN et Clustering hiérarchique** sur la matrice de corrélation des rendements journaliers pour détecter des groupes de titres fortement liés en performance de marché.

L'évaluation de la qualité des clusters s'est faite à l'aide du *Silhouette Score* et de représentations en t-SNE pour une visualisation 2D.

Résultats et interprétations

Les résultats du clustering ont été analysés à travers plusieurs axes :

— **Clustering fondamental (K-Means)** : Le coude d'inertie indiquait $k = 3$ comme nombre optimal de clusters. Chaque groupe présentait des profils distincts : le premier rassemblait des entreprises à forte rentabilité (*ROE* élevé), le deuxième des valeurs cycliques à beta élevé, et le troisième des sociétés sous-évaluées avec un faible *price-to-book*.

— **Clustering de risque (hiérarchique)** : Trois clusters ont également été retenus. Les moyennes des ratios ont permis de distinguer des entreprises à forte liquidité (*currentRatio* élevé), des sociétés endettées (*debtToEquity* élevé), et un groupe intermédiaire avec des marges opérationnelles stables.

— **Clustering des rendements (corrélation + DBSCAN)** : En appliquant DBSCAN (avec $eps = 45$, $min\_samples = 3$), nous avons détecté des noyaux d'entreprises fortement corrélées en termes de performance boursière. Ces groupes coïncident souvent avec les secteurs industriels (tech, énergie, santé) et peuvent être exploités pour diversifier les stratégies ou adapter les modèles prédictifs à des clusters spécifiques.

— **Évaluation de la cohérence** : Les *Silhouette Scores* calculés pour chaque méthode ont confirmé la cohérence des clusters (notamment pour K-Means sur les données fondamentales), avec des scores souvent supérieurs à 0.4.

Ces résultats permettent une segmentation pertinente de l'univers d'investissement. Ils serviront à adapter les signaux de prédiction à des groupes d'entreprises homogènes, renforçant ainsi la robustesse et la pertinence des décisions générées par les modèles supervisés.

4 Classification des signaux d'achat / vente

Objectif : L'objectif de cette partie est de prédire la décision d'investissement à horizon 20 jours sous forme de classe : **achat (2)**, **conserver (1)** ou **vente (0)**. Ces signaux visent à orienter les décisions de trading de manière automatisée en s'appuyant sur des indicateurs techniques.

Méthodes utilisées : À partir des données de clôture, nous avons généré plusieurs indicateurs techniques (SMA, EMA, RSI, MACD, bandes de Bollinger, ROC, volatilité, etc.) à l'aide de la bibliothèque `ta`. Le label a été défini selon le rendement futur à 20 jours :

$$\text{label} = \begin{cases} 2 & \text{si rendement } > 5\% \text{ (achat)} \\ 1 & \text{si } -5\% \leq \text{rendement} \leq 5\% \text{ (conserver)} \\ 0 & \text{si rendement } < -5\% \text{ (vente)} \end{cases}$$

Nous avons testé plusieurs modèles de classification supervisée :

— XGBoost Classifier,

— Random Forest,

— SVM, KNN,

— Régression logistique.

Les modèles ont été entraînés avec validation croisée (`GridSearchCV`) et évalués sur un split 80/20.

Résultats : Les meilleurs résultats ont été obtenus avec le modèle **XGBoost**, avec une précision moyenne de **70%** et un F1-score supérieur à **0.72** pour la classe `Buy`. La répartition des performances par classe a mis en évidence une capacité plus fiable à prédire les extrêmes (achat et vente), les observations intermédiaires étant naturellement plus incertaines.

L'utilisation de **SHAP values** a permis de mieux comprendre les variables influentes : les indicateurs MACD, RSI et Bollinger étaient les plus décisifs pour les décisions d'achat, tandis que la volatilité et le ROC jouaient un rôle clé dans la détection des signaux de vente.

5 Prédiction du rendement à J+1

Objectif

Cette section vise à estimer de manière continue le rendement futur d'une action à un jour (J+1), à partir d'un historique glissant de ses rendements passés. Contrairement à la classification étudiée dans la section précédente, il s'agit ici d'un problème de **régression**, permettant de quantifier la variation attendue du prix plutôt que de la classifier.

Méthodes utilisées

Les données ont été transformées en séries temporelles à fenêtre glissante de 30 jours. Quatre modèles de régression ont été testés :

— **Régression linéaire**,

— **Random Forest Regressor**,

— **K-Nearest Neighbors Regressor**,

— **XGBoost Regressor**.

Chaque modèle a été évalué à l'aide d'une validation croisée temporelle (80/20) et optimisé via `GridSearchCV`. Les performances ont été mesurées à l'aide de deux métriques principales :

— **MSE** (Mean Squared Error),

— **RMSE** (Root Mean Squared Error).

Résultats

Le modèle **XGBoost Regressor** a obtenu les meilleures performances globales sur la majorité des séries, avec un **RMSE inférieur à 0.02** dans la majorité des cas.

Un soin particulier a été apporté à la standardisation des rendements, à la structuration temporelle sans fuite d'information, et à la visualisation des prédictions comparées aux rendements réels. Cette étape constitue un complément continu et chiffré à la prédiction discrète des signaux vue précédemment.

6 Analyse de sentiment des actualités financières

Objectif

L'objectif est d'évaluer l'impact des nouvelles économiques sur les mouvements de prix à court terme, en extrayant un score de sentiment à partir des articles de presse associés aux entreprises étudiées.

Méthodes utilisées

Le processus s'est déroulé en trois étapes :

— **TP6 :** Récupération des articles via l'API `newsapi.org`, avec filtrage par mot-clé (nom de l'entreprise) et par date.

— **TP7 :** Fine-tuning de modèles de type **BERT** (*FinBERT* et *BERT-base*) sur des jeux de données labellisés en sentiments financiers. Chaque article est ensuite classé comme *positif*, *neutre* ou *négatif*.

— **TP8 :** Alignement temporel des sentiments avec les séries de prix pour mesurer leur influence sur les variations à J+1, J+3 et J+5.

Résultats et interprétations

L'analyse a mis en évidence plusieurs tendances intéressantes :

— Les signaux de sentiment ont montré une **corrélation positive** avec les rendements futurs à J+1 dans environ **65% des cas**, surtout autour des périodes de publications de résultats.

— Le modèle **FinBERT fine-tuné** a obtenu une **accuracy moyenne de 79%** sur le jeu de test annoté, dépassant les performances de *BERT-base*.

— L'impact des actualités « négatives » sur la baisse des cours a été plus marqué que celui des nouvelles « positives » sur la hausse.

— L'intégration de ces signaux dans la stratégie globale a permis d'éviter plusieurs positions perdantes, en neutralisant les signaux d'achat lorsque le sentiment médiatique était défavorable.

Cette étape a enrichi la stratégie d'investissement par une composante qualitative issue du NLP, en ajoutant un filtre contextuel utile à la robustesse de la prise de décision.

7 Stratégie d'agrégation des signaux

Objectif

L'objectif est de centraliser les différents signaux issus des modules précédents (clustering, classification, régression, sentiment) dans une architecture unique de prise de décision, permettant de produire une recommandation d'investissement unifiée : achat, vente ou conservation.

Méthodes utilisées

Un script principal `main.py` orchestre l'exécution de tous les modules. Chaque signal produit une sortie normalisée, traduite en score entre 0 (vente) et 1 (achat). Ces scores sont ensuite combinés selon une règle de pondération fixe ou dynamique.

— **Score de classification :** Moyenne pondérée des probabilités issues de TP3.

— **Score de régression :** Transformation du rendement prédit en score via une fonction sigmoïde.

— **Score de sentiment :** +1 pour positif, 0 pour neutre, –1 pour négatif, repondéré entre 0 et 1.

— **Score de cluster :** Adaptation du poids selon la volatilité moyenne du cluster.

Le score final est obtenu par la formule :

$$\text{Score final} = \alpha \cdot S_{\text{classif}} + \beta \cdot S_{\text{regr}} + \gamma \cdot S_{\text{sentiment}} + \delta \cdot S_{\text{cluster}}$$

avec $\alpha + \beta + \gamma + \delta = 1$, ajustés empiriquement.

Une décision est prise selon le seuil :

$$\text{Décision} = \begin{cases} \text{Achat} & \text{si score final} > 0.65 \\ \text{Vente} & \text{si score final} < 0.35 \\ \text{Conserver} & \text{sinon} \end{cases}$$

Résultats et interprétations

L'intégration des signaux a permis d'améliorer la robustesse des décisions par rapport aux approches isolées. Des backtests simples ont montré que :

- Le taux de faux positifs a diminué de 15% en moyenne grâce au filtre de sentiment.
- L'aggrégation a permis d'éviter les achats sur titres volatils au sein de clusters instables.
- Une approche adaptative des poids, basée sur la performance historique par entreprise, offre un potentiel d'amélioration futur.

Cette étape clôt le pipeline en offrant une stratégie cohérente, interprétable et modulaire, facilement réutilisable ou extensible pour d'autres actifs ou périodes.

8 Synthèse des performances

Cette section résume les résultats clés obtenus tout au long du projet, en mettant en lumière les performances des modèles dans chaque bloc fonctionnel. Elle permet d'identifier les approches les plus efficaces pour une intégration future dans un système d'investissement automatisé.

Clustering :

- K-Means : $k = 3$ optimal, Silhouette ≈ 0.45.
- DBSCAN : détection de groupes sectoriels pertinents.
- Hiérarchique : bonne séparation selon ratios de risque.

Classification (Buy / Hold / Sell) :

- Meilleur modèle : **XGBoost**, Accuracy $\approx 70\%$, F1 (Buy) > 0.72.
- Variables clés : RSI, MACD, bandes de Bollinger.

Régression (Rendement à J+1) :

- XGBoost Regressor : RMSE < 0.02 dans la majorité des cas.
- Régression linéaire : résultats plus faibles, sensible au bruit.

Analyse de sentiment :

- FinBERT fine-tuné performant sur actualités financières.
- Sentiment fortement corrélé aux annonces de résultats.

Stratégie globale :

— Pondération dynamique des signaux.

— Performance améliorée lorsque les clusters fondamentaux sont utilisés pour adapter les modèles.

Conclusion : XGBoost s'est avéré le modèle le plus performant de manière générale. L'hybridation des signaux permet d'accroître la robustesse et la précision du système de recommandation.

9 Limites du projet

Bien que notre pipeline ait démontré une certaine robustesse, plusieurs limites subsistent :

— **Dépendance aux données externes** : les performances dépendent fortement de la qualité des données financières et des actualités récupérées via les APIs, sujettes à des interruptions ou à du bruit informationnel.

— **Surapprentissage** : certains modèles (notamment XGBoost) peuvent surajuster les données historiques, réduisant leur capacité de généralisation en conditions réelles.

— **Absence de backtesting complet** : le projet ne comprend pas encore une évaluation rigoureuse via un portefeuille réel intégrant les coûts de transaction, les frais et la liquidité.

— **Limite temporelle** : la fenêtre d'analyse (2019-2024) pourrait ne pas couvrir certains régimes de marché extrêmes (crises, bulles, etc.).

Ces éléments ouvrent des pistes d'amélioration pour des versions futures plus robustes, réalistes et adaptées à des environnements de trading concrets.

10 Conclusion

Ce projet a permis de mettre en œuvre une stratégie complète de data science appliquée aux marchés financiers, en intégrant des signaux techniques, fondamentaux et textuels dans un pipeline cohérent et modulaire.

À travers les différents modules :

— Le **clustering** a structuré notre univers d'investissement,

— La **classification** a guidé les décisions discrètes d'achat/vente,

— La **régression** a permis d'estimer les rendements attendus à court terme,

— L'**analyse de sentiment** a apporté une composante qualitative contextuelle,

— Et l'**agrégation** a synthétisé ces signaux pour produire une stratégie robuste.

Les résultats obtenus confirment l'intérêt de combiner différentes sources d'information et modèles pour améliorer la prise de décision financière. Le modèle XGBoost a montré une forte performance tant en classification qu'en régression, tandis que le sentiment, lorsqu'il est bien aligné temporellement, s'est révélé particulièrement prédictif autour des événements clés.

Perspectives d'amélioration :

— Intégration de facteurs macroéconomiques et d'indicateurs de marché global,

— Prise en compte des coûts de transaction dans le calcul de performance,

— Construction d'un véritable portefeuille et évaluation via backtesting réel.

Ce projet démontre que l'intelligence artificielle, bien appliquée, peut renforcer l'analyse financière traditionnelle et ouvrir la voie à des stratégies plus systématiques et adaptatives.

Bibliographie

[1] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). *Mining the Web for Synonyms : PMI-IR versus LSA on TOEFL*. In Proceedings of the 12th European Conference on Machine Learning.

[2] Nguyen, T. T., & Shirai, K. (2015). *Topic modeling based sentiment analysis on social media for stock market prediction*. ACL-IJCNLP.

[3] Feng, F., He, X., Zhang, H., Chua, T. S., & Wang, M. (2019). *Deep learning for stock selection based on fundamental analysis*. In IEEE Transactions on Knowledge and Data Engineering.

[4] Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. In Advances in Neural Information Processing Systems.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv :1810.04805.

[6] Araci, D. (2019). *FinBERT : Financial Sentiment Analysis with Pre-trained Language Models*. arXiv preprint arXiv :1908.10063.