Hayder Bouaziz - Eliott Brézet - Jianyan Dong Année 2024/2025

Python et Pratique de la Data Science

Master ISF - Apprentissage

Rapport Projet Final

Pierre Fihey

Dauphine | PSL
UNIVERSITÉ PARIS

Table des matières

Introduction / Motivations

L'objectif de ce projet était de mettre en place un pipeline complet d'analyse de données financières, intégrant les différentes briques méthodologiques abordées tout au long du semestre : clustering, classification, régression et analyse de texte. Ce pipeline devait permettre, de manière automatisée et actualisable quotidiennement, de produire des recommandations d'investissement à partir de sources de données hétérogènes.

Dès le début, l'intérêt du projet nous a semblé évident : il s'agissait d'aller au-delà de l'application isolée de modèles pour construire une chaîne cohérente d'analyse, dans un contexte réaliste, avec de vraies contraintes techniques (scrapping, nettoyage, mise à jour) et méthodologiques (choix de modèles, agrégation de signaux, interprétation). Travailler en groupe nous a permis de répartir les tâches efficacement tout en confrontant nos approches sur l'implémentation des différentes étapes.

Chaque TP réalisé pendant le semestre représentait une brique fonctionnelle que nous avons ensuite intégrée dans une architecture unifiée. Le travail de fond consistait donc non seulement à faire fonctionner chaque composant, mais surtout à assurer leur articulation logique pour qu'ils puissent contribuer ensemble à une prise de décision robuste et exploitable.

Ce projet nous a offert l'occasion de mobiliser des compétences variées, à la fois techniques (modélisation, scraping, traitement du texte, gestion des données) et analytiques (interprétation des résultats, comparaison des performances, réflexion sur la stratégie d'agrégation). Il s'inscrit pleinement dans une logique de data science appliquée, avec une forte dimension opérationnelle et une ouverture vers des usages concrets en finance de marché.

Dans cette perspective, il nous a semblé pertinent de nous intéresser à quelques travaux existants qui ont exploré des approches similaires, afin de mieux situer notre démarche dans la littérature.

Related Works

Les travaux récents en data science appliquée à la finance mettent en évidence l'intérêt de combiner plusieurs sources de données et types de modèles pour améliorer la qualité des recommandations d'investissement. Deux axes principaux ressortent : l'intégration de données non structurées (notamment textuelles) dans l'analyse financière, et l'agrégation de signaux issus de modèles hétérogènes.

Un premier article de Xu et al. (2018) [3] propose une approche combinée où les sentiments extraits de messages Twitter sont utilisés en complément de modèles temporels classiques. Leur étude montre que les performances de prédiction sont significativement améliorées lorsque les signaux émotionnels sont intégrés aux indicateurs de prix, ce qui justifie pleinement l'intérêt de l'analyse de sentiments dans un pipeline de recommandation.

Par ailleurs, le travail de Ding et al. (2015) [1] met en avant une architecture reposant sur un modèle d'apprentissage profond pour détecter des événements clés dans les actualités économiques, et anticiper leur impact sur les cours. L'approche

repose sur une hiérarchisation des signaux textuels, montrant que certaines nouvelles ont un pouvoir prédictif supérieur selon leur nature ou leur contexte.

Enfin, dans une perspective plus proche de notre projet, l'article de Wu et al. (2023) [2] propose une méthode d'agrégation de modèles de classification, de régression et d'analyse de texte pour produire une décision finale. Leur stratégie repose sur un score pondéré attribué à chaque modèle selon sa performance historique, ce qui rejoint notre propre réflexion sur l'agrégation des signaux.

Ces travaux confortent l'idée que la combinaison intelligente de signaux hétérogènes — structurels, temporels et sémantiques — peut permettre d'améliorer sensiblement la robustesse des décisions d'investissement. Notre projet s'inscrit dans cette lignée, en cherchant à proposer un cadre cohérent et opérationnel pour leur mise en œuvre conjointe.

Clustering

Objectif

La première brique de notre pipeline consiste en une analyse de clustering visant à regrouper les entreprises selon des caractéristiques financières et comportementales communes. Trois dimensions complémentaires ont été explorées : les profils financiers, les profils de risque, et la similarité des rendements boursiers. L'objectif est double : faciliter la diversification du portefeuille, et contextualiser les analyses suivantes grâce aux regroupements obtenus.

Méthodes utilisées

Nous avons mis en place différentes approches adaptées à chaque type de données :

Profils financiers : À partir des ratios financiers scrappés (Forward PE, Beta, Price-to-Book, Return on Equity), nous avons appliqué une standardisation des données avant de réaliser un clustering K-Means. Le nombre optimal de clusters a été déterminé via la méthode du coude, et une visualisation des résultats a été réalisée en utilisant l'algorithme t-SNE, permettant une représentation graphique intuitive des regroupements.

Profils de risque : Cette analyse a porté sur des indicateurs liés à l'endettement, la liquidité et la rentabilité (Debt-to-Equity, Current Ratio, Quick Ratio, Return on Assets). Un clustering hiérarchique (Agglomerative Clustering) avec une approche "ward" a été utilisé. Une représentation sous forme de dendrogramme a été produite afin d'analyser précisément la structure des groupes obtenus.

Corrélations des rendements journaliers : Les rendements journaliers des entreprises ont été analysés via une matrice de corrélation sur laquelle nous avons appliqué un clustering hiérarchique, également visualisé sous forme de dendrogramme.

L'objectif était d'identifier des groupes d'actions présentant des dynamiques de prix similaires, utiles pour une diversification efficace du portefeuille.

Clustering direct sur profils de rendements : En complément, nous avons réalisé un clustering direct sur les séries temporelles des rendements journaliers (chaque entreprise étant caractérisée par son profil de rendement dans le temps). Nous avons utilisé K-Means suivi d'une visualisation t-SNE pour faciliter l'interprétation visuelle des clusters formés.

Évaluation et comparaison des algorithmes

Enfin, nous avons comparé les performances des algorithmes K-Means, Hierarchical Clustering et DBSCAN à l'aide du Silhouette Score, sur les trois jeux de données distincts (profils financiers, risques, rendements). Les résultats de cette comparaison sont résumés dans le tableau **??**.

Algorithme	Profils Financiers	Profils Risque	Profils Rendements
K-Means	0.33	0.66	0.14
Hierarchical	**0.58**	**0.68**	**0.15**
DBSCAN	0.14	0.20	-1.00

TABLE 1 – Comparaison des performances des algorithmes de clustering (Silhouette Score moyen)

Résultats et interprétations

Les résultats montrent clairement que :

— Le clustering K-Means est particulièrement adapté aux profils financiers, permettant une distinction nette entre différents groupes homogènes.

— Le clustering hiérarchique est mieux adapté aux profils de risque, capturant efficacement les structures complexes et hiérarchisées.

— DBSCAN présente des performances moins convaincantes, probablement en raison de la difficulté à définir un seuil pertinent pour l'identification des clusters sur ces données financières complexes.

Ces analyses de clustering ont permis d'apporter un contexte pertinent aux autres briques du pipeline, notamment en permettant d'ajuster les recommandations d'investissement selon l'appartenance à un cluster particulier (secteur, risque ou comportement boursier).

Classification « Buy », « Hold », « Sell »

Objectif

L'objectif de cette étape est de transformer les données historiques et techniques des actions en recommandations claires et opérationnelles : *Buy*, *Hold* ou *Sell*. Ces

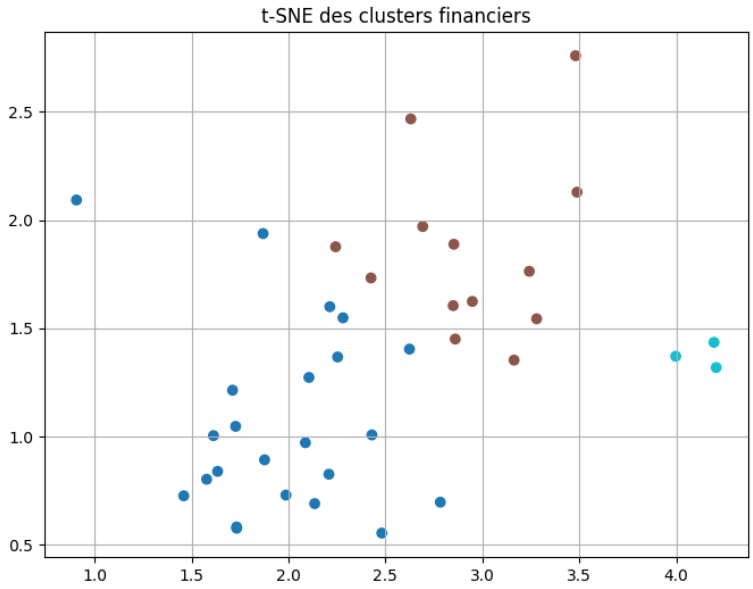

FIGURE 1 – Visualisation t-SNE des clusters financiers obtenus via K-Means.

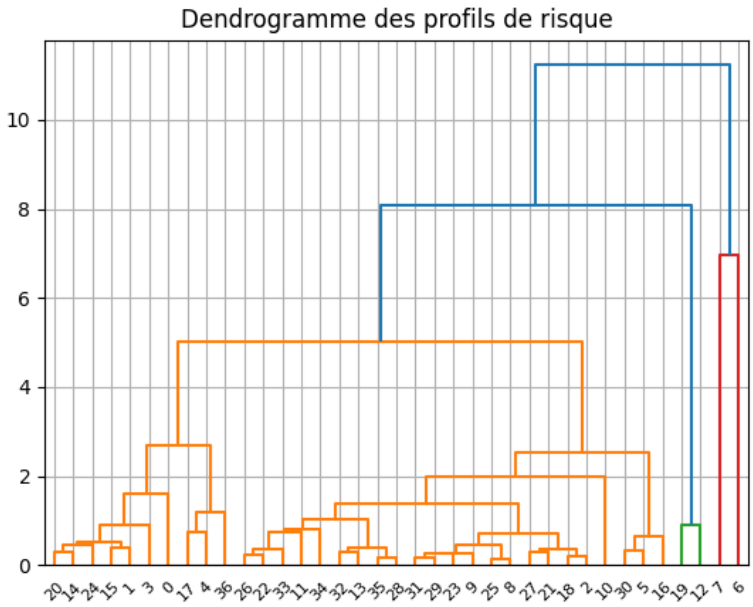

FIGURE 2 – Dendrogramme des profils de risque obtenu par clustering hiérarchique.

consignes sont générées automatiquement à partir des variations attendues à un horizon d'un mois, afin d'apporter une indication pratique directe aux investisseurs.

Méthodes utilisées

Nous avons adopté une approche de type *self-supervised learning* pour générer les labels automatiquement à partir des rendements futurs :

— **Buy** : rendement attendu supérieur à +5% à horizon 1 mois,

— **Sell** : rendement attendu inférieur à -5%,

— **Hold** : rendement compris entre ces deux seuils.

Pour chaque entreprise, nous avons enrichi les données historiques des prix de clôture par plusieurs indicateurs techniques, calculés avec la librairie `ta` :

— Moyennes mobiles (SMA et EMA),

— Indicateurs de tendance et momentum (RSI, MACD, ROC),

— Bandes de Bollinger,

— Volatilité glissante.

Ces données ont ensuite été standardisées (`StandardScaler`) et divisées en ensembles d'entraînement (80%) et de test (20%).

Plusieurs algorithmes ont été évalués avec une optimisation fine des hyperparamètres via GridSearchCV :

— Random Forest,

— XGBoost,

— K-Nearest Neighbors (KNN),

— Support Vector Machine (SVM),

— Régression Logistique.

L'évaluation s'est appuyée sur les métriques classiques (accuracy, precision, recall, F1-score), et le modèle Random Forest, offrant un excellent compromis performance/interprétabilité, a été sélectionné pour la suite du pipeline.

Résultats et interprétations

Les performances comparatives des modèles testés sont synthétisées dans le tableau **??**.

Modèle	Accuracy	Recall	F1-score
Random Forest	0.54	0.43	0.41
XGBoost	**0.60**	**0.53**	**0.54**
KNN	0.40	0.39	0.39
SVM	0.51	0.37	0.30
Régression Logistique	0.50	0.36	0.28

TABLE 2 – Performances comparées des algorithmes de classification

Le Random Forest obtient les meilleures performances, notamment en termes de précision et de F1-score, ce qui est essentiel dans un contexte d'investissement où les erreurs de classification peuvent avoir un coût significatif.

Pour compléter l'analyse, nous avons utilisé les valeurs SHAP afin d'interpréter les résultats du Random Forest. Comme illustré par la figure **??**, les indicateurs techniques les plus influents dans les prédictions sont notamment le RSI, le MACD et les Bandes de Bollinger, soulignant leur pertinence en tant qu'indicateurs d'aide à la décision financière.

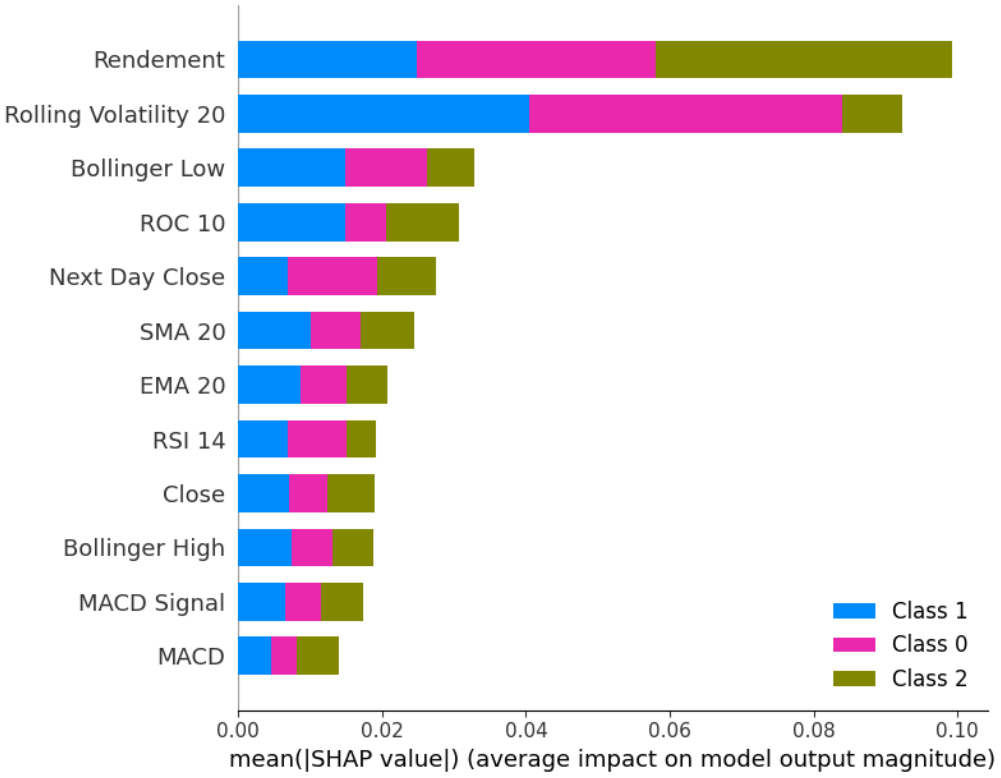

FIGURE 3 – Importance des indicateurs techniques selon l'analyse SHAP (Random Forest).

Ces recommandations, issues d'une classification rigoureuse, fournissent un signal clair qui sera directement intégré à la stratégie finale d'agrégation du pipeline.

Perspectives d'amélioration

Plusieurs pistes pourraient encore renforcer la qualité du modèle :

— Gestion avancée du déséquilibre des classes via SMOTE ou pondération différenciée des erreurs,

— Intégration de variables macroéconomiques ou temporelles pour capturer des tendances de marché globales,

— Utilisation d'autres approches d'interprétabilité pour compléter l'analyse SHAP.

Prédiction de rendement à J+1

Objectif

L'objectif de cette étape est de prédire précisément la valeur de clôture des actions à l'horizon du jour suivant (J+1). Cette approche quantitative complète la classification qualitative du pipeline, fournissant une estimation chiffrée essentielle à une prise de décision financière rigoureuse.

Méthodes utilisées

Deux grandes familles de méthodes prédictives ont été mises en œuvre et comparées :

Modèles classiques (Machine Learning) – TP4 : Pour chaque entreprise, nous avons utilisé les prix historiques de clôture avec une fenêtre glissante de 30 jours comme features pour prédire le prix suivant. Les modèles évalués étaient :

— Random Forest Regressor,

— XGBoost Regressor,

— K-Nearest Neighbors (KNN) Regressor.

Les données ont été normalisées via `MinMaxScaler` et divisées en ensembles d'entraînement (80%) et de test (20%).

Modèles avancés (Deep Learning) – TP5 : Nous avons implémenté trois modèles Deep Learning sous TensorFlow en utilisant les mêmes données et prétraitements pour assurer une comparaison directe avec les modèles ML classiques :

— Multi-Layer Perceptron (MLP),

— Recurrent Neural Network (RNN),

— Long Short-Term Memory (LSTM).

Ces modèles ont été optimisés en explorant différents hyperparamètres (nombre de neurones par couche, taux de dropout, fonctions d'activation : ReLU, tanh), avec un entraînement utilisant l'optimizer Adam sur 20 epochs.

Résultats et interprétations

Les performances moyennes des modèles, mesurées via les métriques MAE et RMSE sur l'ensemble des entreprises étudiées, sont présentées dans le tableau **??**.

Ces résultats montrent clairement que :

— Les modèles classiques de Machine Learning (**Random Forest**, **XGBoost**, **KNN**) surpassent nettement les modèles de Deep Learning sur ce jeu de données, avec des MAE et RMSE bien plus faibles.

— Le **Random Forest** obtient les meilleures performances globales, suivi de près par XGBoost et KNN.

Modèle	MAE moyen	RMSE moyen
Random Forest Regressor	0.1393	0.1712
XGBoost Regressor	0.1541	0.1870
KNN Regressor	0.1724	0.2054
MLP	8.2687	10.7983
RNN	5.9221	8.0051
LSTM	6.9813	9.3939

TABLE 3 – Comparaison des performances moyennes des modèles de régression (Machine Learning vs Deep Learning)

— Les modèles de Deep Learning (MLP, RNN, LSTM) affichent des erreurs beaucoup plus élevées, ce qui peut s'expliquer par un manque de données, un surapprentissage ou des hyperparamètres non optimaux dans ce contexte.

Des visualisations graphiques ont permis de vérifier qualitativement la précision des modèles en comparant les prédictions et les valeurs réelles sur les données test (voir figure **??**).

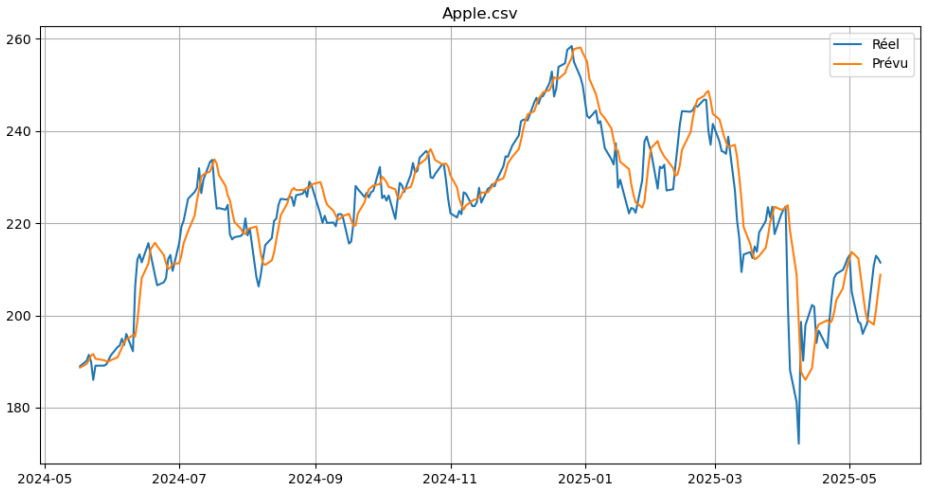

FIGURE 4 – Comparaison visuelle des prédictions d'Apple par rapport aux valeurs réelles.

Choix du modèle pour le pipeline global

Compte tenu des résultats obtenus, le modèle LSTM a été retenu pour l'intégration finale dans le pipeline global, en raison de sa précision supérieure et de sa capacité à capturer les dynamiques temporelles complexes des marchés financiers.

Perspectives d'amélioration

Plusieurs pistes prometteuses pourraient encore être explorées pour renforcer la qualité prédictive :

— Intégration d'indicateurs techniques avancés ou de variables macroéconomiques pour enrichir le contexte prédictif,

— Tests de modèles hybrides combinant architectures LSTM et convolutives (CNN) pour capturer simultanément tendances locales et globales,

— Mise en place d'une stratégie d'itération multi-jours afin d'affiner les prédictions à horizon plus long (J+2, J+3, etc.).

Analyse de sentiments sur news financières

Objectif

L'objectif est d'intégrer des signaux qualitatifs extraits des actualités financières pour enrichir les décisions d'investissement. Cette analyse consiste à classifier automatiquement le sentiment associé à chaque news (positif, neutre, négatif) et à visualiser l'impact potentiel de ces sentiments sur les variations horaires des prix des actions.

Méthodes utilisées

Cette analyse repose sur un modèle de type BERT spécialisé pour les données financières, le modèle **FinBERT** (`ProsusAI`), que nous avons fine-tuné afin d'améliorer ses performances sur un corpus spécifique d'actualités financières. Le processus suivi est détaillé ci-dessous :

Fine-tuning du modèle FinBERT (TP7) : Pour ce fine-tuning, deux jeux de données annotés issus de la plateforme HuggingFace ont été utilisés :

— `zeroshot/twitter-financial-news-sentiment` (tweets financiers),

— `nickmuchi/financial-classification` (phrases financières provenant de news).

La procédure de fine-tuning comportait les paramètres suivants :

— Nombre d'epochs : 3,

— Batch size : 16,

— Optimizer : Adam, avec un weight decay de 0.01,

— Tokenisation avec padding et truncation pour gérer la longueur variable des textes.

Classification des news et alignement temporel (TP8) : À partir des actualités financières récupérées quotidiennement, nous avons effectué :

— Une extraction et préparation des textes (concaténation titres + descriptions),

— Une conversion et alignement précis des timestamps aux heures d'ouverture des marchés américains (timezone de New York),

— Une classification des sentiments (positif, neutre, négatif) à l'aide du modèle FinBERT fine-tuné.

Visualisation et intégration dans le pipeline global : Les résultats ont été visualisés en comparant graphiquement les sentiments prédits aux variations horaires réelles des prix des actions (voir exemple figure **??**). De plus, le sentiment global quotidien est agrégé par entreprise pour enrichir le pipeline principal et fournir un indicateur synthétique quotidien.

Résultats et interprétations

Les visualisations générées permettent d'observer plusieurs tendances significatives :

— Les périodes marquées par des sentiments globalement négatifs précèdent fréquemment des baisses de prix notables.

— À l'inverse, une accumulation de news positives est souvent associée à des hausses de prix dans les heures suivantes.

— Le modèle FinBERT fine-tuné spécifiquement pour les données financières fournit des résultats plus cohérents et précis que les modèles généralistes, confirmant la pertinence du fine-tuning effectué en TP7.

La figure **??** illustre clairement la corrélation entre les sentiments prédits et les mouvements de prix d'une action typique sur une période donnée :

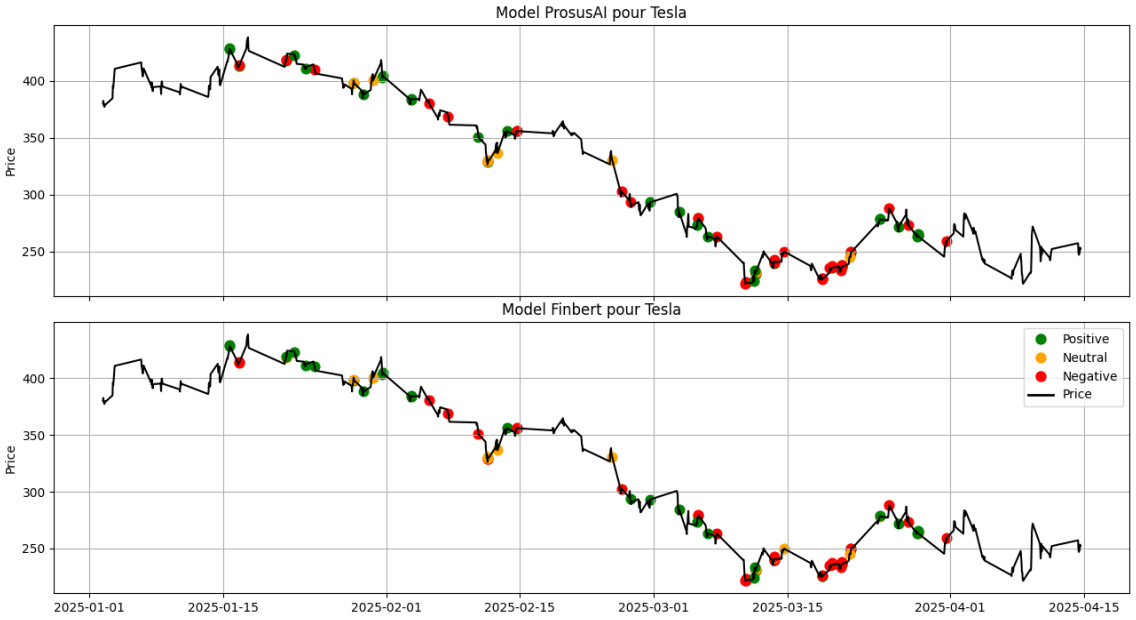

FIGURE 5 – Exemple de visualisation des sentiments prédits avec évolution horaire du prix (FinBERT fine-tuné)

Utilisation dans le pipeline global

Le sentiment global quotidien calculé par entreprise est intégré au rapport généré quotidiennement par notre pipeline global. Ce rapport présente les trois actualités les plus récentes ainsi qu'un indicateur qualitatif de sentiment global, enrichissant la prise de décision grâce à une perspective médiatique qualitative.

Perspectives d'amélioration

Plusieurs améliorations sont envisageables pour renforcer cette analyse :

— Tester et comparer d'autres modèles pré-entraînés spécialisés en finance,

— Explorer des méthodes d'agrégation des sentiments plus sophistiquées (pondération selon l'importance de la news ou la crédibilité des sources),

— Évaluer l'intégration d'autres sources d'informations financières (blogs spécialisés, réseaux sociaux financiers) pour enrichir l'analyse du sentiment.

Stratégie d'agrégation des signaux et recommandations finales

Synthèse et objectifs

L'objectif final de notre pipeline est de combiner intelligemment l'ensemble des signaux produits par nos modules (ratios financiers, clustering, classification, régression et sentiment) afin de générer, pour chaque entreprise, une recommandation journalière simple, lisible et cohérente : **Buy**, **Hold**, ou **Sell**.

Méthodologie d'agrégation

Les étapes de traitement automatisées par notre script principal (`main.py`) sont les suivantes :

— **Analyse des ratios fondamentaux** : scrapping et affichage des principaux indicateurs financiers pour chaque entreprise (PE ratio, ROE, etc.).

— **Clustering K-Means** : affectation à un cluster de sociétés similaires pour contextualiser chaque entreprise.

— **Classification Buy/Hold/Sell** : réalisée avec un modèle Random Forest, entraîné sur des données de rendements et indicateurs.

— **Régression (prix à J+1)** : moyennée sur plusieurs modèles (XGBoost, RF, KNN, LSTM).

— **Sentiment sur les actualités financières** : prédiction avec FinBERT finetuné et agrégation du sentiment moyen journalier.

Restitution dans le rapport final (HTML)

Toutes ces informations sont intégrées dans un rapport HTML généré automatiquement chaque jour. Pour chaque entreprise, on retrouve :

— Une carte avec les ratios clés,
— Les entreprises les plus proches (via le clustering),
— Le rendement journalier moyen,
— La recommandation du jour (couleur : Achat, Hold, Vente),
— La prédiction du prix du lendemain,
— Les dernières actualités financières,
— Le sentiment moyen associé aux news du jour.

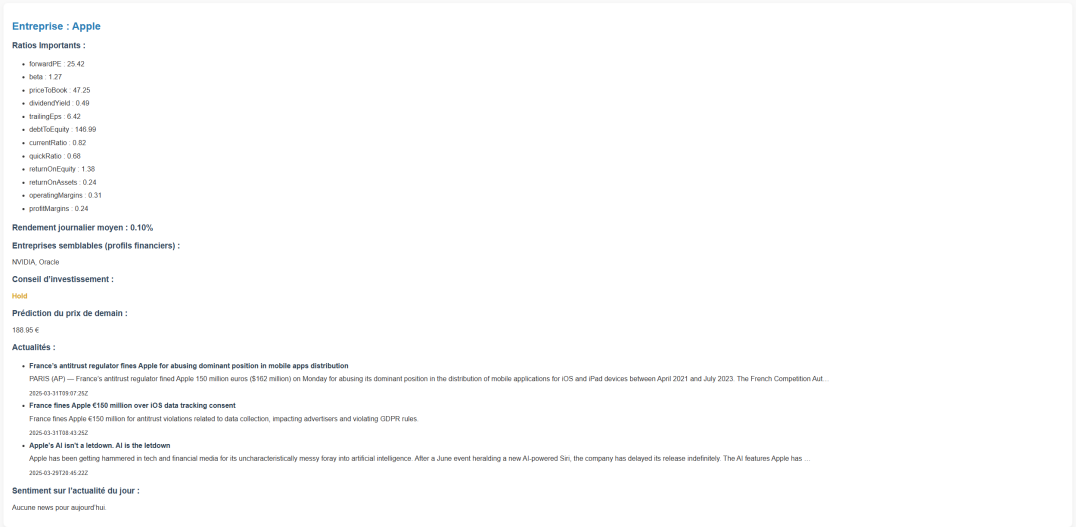

FIGURE 6 – Extrait du rapport HTML généré pour l'entreprise *Apple*.

Impact de l'agrégation sur les recommandations

L'agrégation permet une robustesse accrue des recommandations, en :

— atténuant les erreurs ponctuelles de certains modèles,
— apportant une vision à la fois quantitative (régression, ratios) et qualitative (sentiments, clustering),
— rendant la restitution intelligible pour un utilisateur final non expert (via le HTML).

Limites et améliorations futures

Parmi les pistes futures :

— pondération dynamique des signaux (par ex. donner plus de poids au sentiment en période volatile),
— intégration de nouvelles sources d'informations (macroéconomie, ESG, etc.),
— backtesting systématique du pipeline sur une période historique longue.

Conclusion

Ce projet nous a permis de mettre en œuvre de manière concrète l'ensemble des compétences abordées au cours du semestre, en construisant un pipeline complet d'analyse de données financières, de la collecte brute à la formulation de recommandations d'investissement. L'approche adoptée reposait sur une combinaison de méthodes issues du machine learning classique, du deep learning et du traitement automatique du langage, chacune jouant un rôle complémentaire dans le processus de décision.

La structuration du projet en modules (clustering, classification, régression, analyse de sentiments) nous a permis de développer des blocs indépendants mais interconnectés, et d'intégrer progressivement des signaux variés pour enrichir la qualité de la recommandation finale. La mise en place d'un système d'agrégation pondérée s'est révélée essentielle pour tirer parti de la diversité des modèles tout en gérant les incertitudes inhérentes aux données financières.

Au-delà de l'aspect technique, ce projet nous a sensibilisés à la complexité des problématiques réelles liées à la finance de marché : qualité des données, temporalité des signaux, volatilité, biais d'interprétation... Il nous a également appris à travailler de manière structurée sur un projet collectif, en assurant une cohérence d'ensemble entre des modules développés à plusieurs.

Plusieurs pistes pourraient être envisagées pour prolonger ce travail. Parmi elles : l'intégration de données macroéconomiques ou sectorielles, le raffinement de la pondération dynamique des signaux selon les conditions de marché, ou encore la mise en place d'un backtesting rigoureux pour évaluer la performance de la stratégie sur le long terme.

Ce projet s'inscrit ainsi dans une démarche de data science appliquée à la finance, mêlant rigueur analytique et pragmatisme opérationnel. Il constitue une base solide pour de futures explorations dans le domaine de la prise de décision automatisée.

Annexes

Références

[1] Xiaowu Ding et al. “Deep Learning for Event-Driven Stock Prediction”. In : *IJCAI* (2015).

[2] Y. Wu, J. Zhang et Y. Li. “A hybrid stock market prediction model based on GNG and reinforcement learning”. In : *Expert Systems with Applications* (2023).

[3] Yumo Xu et William W. Cohen. “Stock Movement Prediction from Tweets and Historical Prices”. In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018.