Rendu – Séance 1

Mise en place du Projet de Machine Learning Financier

November 19, 2025

1 Introduction

Cette première séance avait pour objectif de poser les fondations techniques du projet de Machine Learning appliqué à la finance. Nous avons réalisé trois grandes étapes :

- la création d'une arborescence professionnelle pour structurer le code,
- l'installation des librairies essentielles pour le traitement de données, le NLP, le machine learning et la création d'une WebApp,
- l'écriture d'un script Python permettant de récupérer automatiquement les données boursières nécessaires au projet via l'API publique de Yahoo Finance.

Ce document présente chacune de ces étapes en détaillant les motivations techniques et méthodologiques.

2 Création du squelette du projet

Nous avons créé la structure suivante :

```
FINANCE/
 .venv/
 data/
     raw/
 notebooks/
 deliverables/
 src/
     data/
     models/
     nlp/
     utils/
 README.md
```

Chaque dossier a un rôle bien précis :

- **.venv/** : environnement virtuel isolant toutes les librairies du projet.
- **data/raw/** : stockage des données brutes téléchargées (fichiers CSV).
- **notebooks/** : analyses exploratoires, visualisations, prototypes rapides.
- **deliverables/** : documents finaux, rapports, résultats exportés.
- **src/data/** : scripts de collecte, nettoyage et préparation des données.
- **src/models/** : modèles de machine learning (forecasting, classification, etc.).
- **src/nlp/** : analyse de sentiment via FinBERT ou d'autres modèles NLP.
- **src/utils/** : fonctions transversales (gestion des fichiers, métriques, etc.).

Cette organisation est inspirée des bonnes pratiques en data science et MLOps : chaque bloc du pipeline est isolé et réutilisable.

3 Installation des librairies essentielles

Nous avons ensuite activé l'environnement virtuel puis installé toutes les librairies utiles au projet. Voici leur rôle :

- **pandas** : manipulation de données tabulaires et séries temporelles.
- **numpy** : opérations numériques vectorisées, indispensable pour le ML.
- **yfinance** : accès direct à l'API de Yahoo Finance pour télécharger les données de marché.
- **matplotlib & seaborn** : visualisation de données financières.
- **scikit-learn** : modèles ML classiques (régression, classification, PCA, etc.).
- **torch** : réseaux de neurones, LSTM, CNN, Transformers légers.
- **transformers** : chargement de modèles NLP avancés comme FinBERT.
- **streamlit** : création d'une WebApp pour visualiser les prédictions du modèle.

Ces librairies couvrent l'ensemble du pipeline : de la collecte des données jusqu'au déploiement d'une interface finale.

4 Script Python pour la collecte des données

Nous avons écrit un script permettant de télécharger automatiquement les données boursières des quatre plus grandes entreprises liées à l'écosystème de l'intelligence artificielle :

- **MSFT** (Microsoft) : partenaire principal d'OpenAI, fournisseur majeur d'infrastructure via Azure.
- **NVDA** (Nvidia) : leader mondial des GPU utilisés pour entraîner les modèles IA.
- **AMD** : alternative stratégique en GPU, récemment partenaire d'OpenAI.
- **GOOG** (Alphabet) : créateur du modèle Gemini, équivalent public de DeepMind/OpenAI.

Ce choix représente les acteurs majeurs du "cercle économique" de l'IA : hardware (NVDA, AMD), cloud & services (MSFT), modèles IA (Alphabet).

Code du script (src/data/load$_data.py$)

```
import yfinance as yf
import pandas as pd
import os

def download_stocks(tickers, start_date="2020-01-01", folder="data/raw"):
    os.makedirs(folder, exist_ok=True)
    for t in tickers:
        print(f"Downloading {t} ...")
        df = yf.download(t, start=start_date)
        df.to_csv(f"{folder}/{t}.csv")
        print(f"Saved: {folder}/{t}.csv")

if __name__ == "__main__":
    tickers = ["MSFT", "NVDA", "AMD", "GOOG"]
    download_stocks(tickers)
```

Ce script télécharge automatiquement les données depuis le 1er janvier 2020, année de l'émergence des premiers modèles GPT modernes, jusqu'à aujourd'hui.

Les fichiers obtenus sont stockés dans `data/raw/`.

5 Conclusion

Cette première séance a posé des bases solides pour le projet. Nous avons :

- créé une architecture claire et professionnelle,
- installé un environnement logiciel complet,
- automatisé la collecte des données financières au coeur du projet.

Nous sommes maintenant prêts à entamer la prochaine étape : l'exploration des données et la création des premières features pour les modèles de prévision et d'analyse.