Project 4 Initial Proposal

Team Members: Adithya Ramanathan (ar74353), Parth Patki (pap2389)

1. Introduction and Problem Statement

The stock market is a dynamic, non-linear system influenced by numerous factors. Our project investigates the feasibility and limitations of predicting future stock closing prices using historical data. Specifically, we aim to compare the performance of various machine learning models—both classical and deep learning—on their ability to forecast the closing price of individual stocks, predict within-industry stock behavior, and assess cross-industry generalization.

We are particularly interested in understanding whether models trained on one stock can generalize to others within the same industry or across different sectors, and which models are best suited to each prediction type.

2. Data Sources

We plan to use the Yahoo! Finance Stock Dataset, accessed via the yfinance Python library. The dataset includes daily open, high, low, close, volume, dividends, and stock splits for major publicly traded companies. We will focus on a time period from 2000 to 2020, excluding the COVID-19 pandemic period to avoid anomalous patterns.

The industries selected were:
- Technology: AAPL, NVDA, GOOG
- Healthcare: ABBV, LLY, MRK
- Construction: TT, CRH, URI
- Food: ADM, GIS, KHC

3. Methods and Technologies

We plan to apply several models to predict stock closing prices, with emphasis on both traditional and neural network-based techniques:

- Baseline Models: Linear Regression, Random Forest, XGBoost Regressor
- Deep Learning Models:
  - Recurrent Neural Network (Simple RNN)
  - Long Short-Term Memory (LSTM v1, v2, v3)
  - Multilayer Perceptron (MLP)
  - Transformer (PyTorch implementation)

Technologies to Consider:
- Python (NumPy, Pandas, Scikit-learn)
- TensorFlow / Keras, PyTorch
- Google Colab for experimentation and training
- Matplotlib/Seaborn for visualizations

Techniques to Consider:
- Feature Engineering (Moving Averages, Lag Features)
- Data Scaling (Standard and Min-Max Scalers)
- Train/Test Split (80/20 with no shuffle for time series)
- Evaluation Metrics: MAE, MSE

4. Products to be Delivered

- A technical report detailing our models, training process, evaluation, and insights (delivered as a final report).
- Colab notebooks containing all preprocessing, training code, and visualizations (linked in the final report).