Final Project: Proposal

Sophia Yang (xya134), Yingzi Yuan (yya309), Weiwei Zhang (wza124)

1. Research questions

List 3 questions that you intend to answer (1 point)

1. How to accurately predict future stock market price based on historical time-series data?
2. Which model has the best prediction accuracy? Compare the traditional machine learning models and deep learning models.
3. What other features can help increase the prediction accuracy? Explore extra features using the Yahoo Finance API (yfinance).

2. Dataset utilization

List all the datasets you intend to use (1 point)

- First, we use an existing dataset: S&P 500 stock data to develop and tune our models. The dataset can be found at https://www.kaggle.com/datasets/camnugent/sandp500/data. It has the stock market data for 5 years for S&P 500 companies.
- Then, we use the Yahoo Finance API (https://ranaroussi.github.io/yfinance/index.html) to collect more recent data, for example stock data in 2024 & 2025, and use this new dataset to further enhance our models.
- Finally, we also extract additional features using Yahoo API that might enhance our prediction accuracy, and form a stronger dataset and model as our final result.

3. Methodology

Give us a rough idea on how you plan to use the datasets to answer these questions. (2 points)

- Data Collection: Kaggle & Yahoo Finance API
- Data Exploration: We will conduct EDA for basic understanding of data, but EDA is not the focus of our project, since we have a very clear target: prediction.
- Data Cleaning: For Kaggle dataset, only minimum effort for validating the data, but for dataset collected from Yahoo API, it is possible to build data pipeline for cleaning and validation.
- Data Integration: Need to access external source using Yahoo API.
- Data Analysis:
    - What analysis do you intend to do?
        - Traditional machine learning models such as: ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), XGBoost / LightGBM / Random Forest, Support Vector Regression (SVR).
        - Deep learning models such as LSTM (Long Short–Term Memory), GRU (Gated Recurrent Unit), CNN for Time–Series, Transformer–based models (like Time–Series Transformer).
    - How to evaluate your analysis results? (e.g., evaluation metrics, confidence intervals, benchmark)
        - Root mean squared error for predicted values
- Data Product: prediction model and visualization

4. Expected impact

Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

Predicting price movements can help traders optimize buying and selling strategies.

5. Potential challenges

Identify any anticipated obstacles and how you plan to address them. (1 point)

Challenge: Choosing Between Traditional ML vs. Deep Learning

Traditional models (XGBoost, Random Forest) work well for small datasets, while LSTM / Transformer models require large datasets. If the dataset is too small, deep learning may overfit.

How to address them: identify overfitting, create larger dataset with API, study regularization methods, etc.